ECOLOGICALLY VALID LONG-TERM MOOD MONITORING OF INDIVIDUALS WITH BIPOLAR DISORDER USING SPEECH

Zahi N. Karam[1]*, Emily Mower Provost*[1]*, Satinder Singh*[1]*,*
Jennifer Montgomery[2]*, Christopher Archer*[2]*, Gloria Harrington*[2]*, Melvin G. Mcinnis*[2]

Departments of: Computer Science and Engineering[1] and Psychiatry[2], University of Michigan

ABSTRACT

Speech patterns are modulated by the emotional and neurophysiological state of the speaker. There exists a growing body of work that computationally examines this modulation in patients suffering from depression, autism, and post-traumatic stress disorder. However, the majority of the work in this area focuses on the analysis of structured speech collected in controlled environments. Here we expand on the existing literature by examining bipolar disorder (BP). BP is characterized by mood transitions, varying from a healthy euthymic state to states characterized by mania or depression. The speech patterns associated with these mood states provide a unique opportunity to study the modulations characteristic of mood variation. We describe methodology to collect unstructured speech continuously and unobtrusively via the recording of day-to-day cellular phone conversations. Our pilot investigation suggests that manic and depressive mood states can be recognized from this speech data, providing new insight into the feasibility of unobtrusive, unstructured, and continuous speech-based wellness monitoring for individuals with BP.

***Index Terms*—** Speech Analysis, Bipolar Disorder, mood modeling

1. INTRODUCTION

Bipolar disorder (BP) is a common and severe psychiatric illness characterized by pathological swings of mania and depression and is associated with devastating personal, social, and vocational consequences (suicide occurs in up to 20% of cases [1]). Bipolar disorder is among the leading causes of disability worldwide [2]. The cost in the United States alone was estimated at $45 billion annually [3]. These economic and human costs, along with the rapidly increasing price of health care provide the impetus for a major paradigm shift in health care service delivery, namely to monitor and prioritize care with a focus on prevention. In this paper, we present our pilot investigation into methods to unobtrusively collect and analyze speech data for longitudinal wellness monitoring to meet this ever-growing need.

Speech patterns have been effectively used in clinical assessment for both medical and psychiatric disorders [1,4]. Clinicians are trained to record their observations of speech and language, which become a critical component of the diagnostic process. Recently, there have been research efforts exploring computational speech analysis as a way to assess and monitor the mental state of individuals suffering from a variety of psychological illnesses, specifically major depression (MD) [5–8], autism [9–12], and post-traumatic stress disorder (PTSD) [13–16].

Stress and anxiety have been studied extensively and elements of speech have been correlated with subjectively reported stress in PTSD [13–16]. Research efforts have demonstrated the efficacy of speech-based assessments for autism focusing on diagnosis [12], in addition to predicting the course and severity of the illness [10, 17, 18]. Variations in speech patterns have also been used for computational detection and severity assessment in major depressive disorder [5–8, 19–21]. However, most work in this area focuses on the assessment of participants over short periods of time, at most several weeks, [19] rendering it challenging to measure the natural fluctuations that accompany the illness trajectories. Additionally, the speech input is often highly-structured and collected in controlled environments [13, 14, 16], precluding an understanding of how acoustic patterns characteristic of natural speech variation correlate with mood symptomology.

This paper focuses on the estimation of mood state for individuals with BP. This disorder is characterized by fluctuating mood state, including periods of depression (lowered mood state), mania (elevated mood state), and euthymia (neither mania nor depression). The dynamic nature of the symptoms and temporal course of bipolar disorder are well suited to a comparative study of the acoustic patterns associated with mood and illness states. Furthermore, unlike previous work that examined individuals over relatively short periods of time, the population in our study is continuously monitored over the course of six months to a year using our cell phone-based recording software that unobtrusively records all outgoing speech.

The work presented in this paper represents a pilot analysis of our initial collection targeting six individuals with BP. The ground truth labels of our data are established through weekly structured interactions between the participant and a trained clinician. We demonstrate that we can detect the presence of mania and depression in these calls. We further test the hypothesis that speech collected in an unstructured setting (outside the clinician interaction) can be used to assess the underlying mood state. We provide evidence that mood-related variations recorded both from structured and unstructured cell phone conversation data are reflective of underlying mood symptomology and that the acoustic variations indicative of mood patterns across these conversation types differ. Furthermore, we highlight the features of the speech that are most correlated with the clinical assessment of manic and depressive mood states.

The novelty of our approach resides both in the longitudinal, ecological, and continuous collection of unstructured speech in diverse environments and in the acoustic analysis of the BP participant population, which exhibits mood-states at two extremes of the mood-state spectrum, depression and mania. Our results suggest that this style of data collection can be effectively used, highlighting the potential for autonomous ecologically valid monitoring for mental

This work was supported by the Department of Health and Human Services of the National Institutes of Health under award number R34MH100404 and the Heinz C. Prechter Bipolar Research Fund at the University of Michigan Depression Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

health assessment.

2. UM PRECHTER ACOUSTIC DATABASE (UM-PAD)

Description: The University of Michigan Prechter Acoustic Database (UM-PAD) consists of longitudinally collected speech from individuals diagnosed with bipolar disorder participating on the Prechter BP Longitudinal Study [22], a multi-year study that takes a multidimensional, biological, clinical, and environmental, approach to the study of BP.

Enrollment: UM-PAD contains speech data collected from six participants, four women and two men (average age 41 ± 11.2) diagnosed with bipolar disorder type I and with a history of rapid cycling, characterized by 4 or more episodes per year of mania, hypomania, or depression. Participants are recruited from the Prechter Longitudinal study and enrolled for 6 months to a year.

Protocol: Each participant is provided with a "smart phone" and an unlimited call/data plan for personal use and is encouraged to use the phone as their primary mode of contact. The phone is pre-loaded with an application that records only the participant's *outgoing* speech (i.e. no incoming speech is captured or recorded), at $8KHz$, whenever they make or receive a phone call. All collected speech is encrypted and transferred securely for analysis. The application, data transfer, and handling follow strict security and encryption guidelines approved by the internal review board (IRB HUM00052163) to ensure that the integrity and privacy of the collected data is not compromised.

Weekly Mood-State Labels: Ground truth measures of a participant's mood-state are obtained using weekly phone-based interactions with a clinician associated with this project. The clinician administers a twenty-minute recorded assessment that measures the mood-state of the participant over the past week. The assessments include the 17 item Hamilton Rating Scale for Depression (HAMD) [23] as well as the Young Mania Rating Scale (YMRS) [24] to assess the level of depression and mania, respectively. In the current stage of our collection, no participant has exhibited symptom severity associated with a manic episode. As a result, our objective is to detect hypomania (elevated mood state not reaching the severity of mania).

We categorize the mood assessments using thresholds set by the clinical team. The final labels are as follows: **Hypomanic:** YMRS$\geq$ 10 and HAMD $<$ 10. **Depressed:** HAMD$\geq$ 10 and YMRS $<$ 10. **Euthymic:** YMRS$<$ 10 and HAMD$<$ 10. **Mixed:** YMRS$\geq$ 10 and HAMD$\geq$ 10. However, mixed mood-state is not included in this paper due to its rarity in the collected data.

The weekly clinical assessments ("evaluation call") provide a measure both of the participant's mood-state over the past week and of the clinician's perception of the participant's current (during evaluation call) mood-state. We hypothesize that the labels obtained during an evaluation call will be most strongly associated with the participant's mood during that evaluation call and thus with the mood-related modulations of the speech recorded during the call. We further hypothesize that the set of calls disjoint from the evaluation calls, calls recorded outside of a clinical interaction, will possess a more subtle expression of mood symptomology, may involve masking of symptomology, and will correlate less strongly with the clinically assessed labels. It is important to note that the only data with labels are the evaluation calls.

Statistics on Recorded Calls: A total of 221.2 hours was recorded from $3,588$ phone calls. On average participants made 4.9 ± 4.3 calls per day, with an average duration of 222.0 ± 480.7 seconds and a median of 67.4 seconds.

The number of weeks of data available varies by participant: participant 1 has 31 weeks of data, while participant 5 has 6 weeks of data. Each participant's data includes euthymic weeks and at least one hypomanic and/or depressive week. Table 1 provides an overview of the collected data for each participant, showing the number of weeks of collected data with categorized assessment labels of euthymic, hypomanic, and depressive.

Table 1. Summary of collected data. $\#E, \#H, \#D$ are the number of weeks in *E*uthymic, *H*ypomanic, and *D*epressive states.

Part. #	**1**	**2**	**3**	**4**	**5**	**6**
#(E:H:D)	22:2:7	9:0:4	21:1:3	10:9:1	2:4:0	3:0:4

3. ANALYSIS SETUP

Our research objective is to use speech data collected in an unobtrusive and unstructured environment to: (1) estimate the clinical assessment made during the participant-clinician weekly evaluation call; (2) determine the feasibility of detecting the mood state assessed during the evaluation call using unstructured personal cell phone recordings from the same day as the evaluation call; and (3) apply this detection to cell phone recordings from days preceding or following the evaluation call. We also conduct feature analyses to identify the speech features that are most informative for mood classification.

This estimation task is a very challenging due to the sparse nature of the data labeling (weekly assessments), the acoustic variability associated with human communication and natural mood fluctuations, and the variability due to uncontrollable environmental factors. A successful result will suggest that speech data collected in uncontrolled and unstructured environments exhibit similar acoustic variations to speech data collected in a structured clinical interaction, which will support both the feasibility of longitudinal wellness monitoring and the feasibility of using clinical data to seed models for deployment in unstructured monitoring.

Datasets: The UM-PAD dataset is partitioned to address the research questions presented above. The partitions are based on proximity to evaluation call. Recall that the evaluation calls are the only recordings that are labeled. Further, the temporal consistency of mania and depression are variable and person dependent. Therefore, it is expected that the labels of the evaluation call are more strongly associated with calls recorded on the day of the evaluation as opposed to the day(s) before or after it.

The data are partitioned into the following *disjoint* datasets. Table 2 describes the per-participant summary of the number of calls assigned each of the three labels. The datasets include:

- *Evaluation calls:* Speech collected during evaluation calls labeled as hypomanic/depressed/euthymic based on the clinical assessment.
- *Day-of calls:* Speech collected from all calls recorded on the day of the clinical assessment, **excluding** the evaluation call.
- *Day before/after (B/A) calls:* Speech collected from all calls made or received *only* on the adjacent day (before or after).

Training Methodology: The classification algorithms are trained using participant-independent modeling, capturing the variations associated with populations of individuals, rather than specific individuals. As the size of the UM-PAD database continues to grow we anticipate leveraging participant-dependent modeling strategies. The goal of **participant independent** modeling is to understand how speech is modulated as a function of mood state while mitigating the

effects of individual variability. We test our models using the leave-one-participant-out cross-validation framework, where each participant is held out for testing and the remaining participants are used for training. The validation set is obtained using leave-one-training-participant out cross-validation within the training set. We train our models using all data from the categories of euthymia, hypomania, and depression. We evaluate the performance of our depression and hypomania classifiers only for participants with at least two weeks of evaluation calls labeled as either depressed or hypomanic.

Table 2. Number of calls assigned each of the categorical labels:

	Part. #	**1**	**2**	**3**	**4**	**5**	**6**
Eval	Euthymic	18	8	21	6	1	2
	Hypomanic	2	0	1	3	3	0
	Depressed	6	4	3	1	0	3
Day-Of	Euthymic	52	227	127	11	10	17
	Hypomanic	13	0	5	14	11	0
	Depressed	22	114	21	1	0	22
Day-B/A	Euthymic	77	202	271	25	5	60
	Hypomanic	7	0	11	22	12	0
	Depressed	29	100	47	2	0	41

4. FEATURES AND CLASSIFIER

It is crucial that we protect the privacy of the participants given the sensitive nature of speech collected from personal phone calls. This is done through the use of statistics extracted from low-level audio features, rather than the features themselves. The statistics are calculated over windows of at least three seconds in length. This windowing obscures the lexical content of the original speech, rendering it extremely challenging to reconstruct the individual words.

Low-level Features: We extract 23 low-level features (LLF) using the openSMILE toolkit [25]. For each recorded call, the speech is windowed into $25ms$ frames overlapping by $15ms$, with the following features extracted per frame:

• Pitch, computed using the autocorrelation/cepstrum method [25], which yields the pitch over voiced windows. For unvoiced windows the pitch is set to 0. Whether a window is voiced is determined by a voicing probability measure, which we also include in the LLF.
• RMS energy, zero-crossing rate, and the maximum and minimum value of the amplitude of the speech waveform.
• Three voiced activity detection (VAD) measures: fuzzy, smoothed, binary. The fuzzy measure is computed using line-spectral frequencies, Mel spectra, and energy. The smoothed measure is the result of smoothing the fuzzy measure using a 10 point moving average. The binary measure, is a $1/0$ feature, by thresholding the fuzzy measure to assess presence of speech.
• The magnitude of Mel spectrum over 14 bands ranging from $50Hz$ to $4KHz$.

Segment Level Features: The VAD measures and voicing probability provide an estimate of the location of speech and silence regions of the input speech waveform. We use these measures to group the speech into contiguous segments of participant speech ranging from 3 seconds to at most 30 seconds. We divide the call into segments by finding non-overlapping regions of at least 3 seconds. We first identify 3 consecutive frames whose energy, voicing probability, and fuzzy VAD are all above the 40^{th} percentile of their values over the whole call. We end a segment when 30 consecutive frames have energy, voicing probability, and fuzzy VAD measures that fall below the 40^{th} percentile of their values over the whole call. If the segment length exceeds 30-seconds before reaching the stopping criteria then the segment is ended and a new one is started; this occurs for less than 3.5% of the segments. Each call has on average 24.3 ± 46.6 segments with a median of 8.

We represent each segment by a 51-dimensional feature vector obtained from the statistics of the LLFs over the segment. This includes 46 mean and standard deviation values of each LLF computed over the segment (for the pitch, these are computed only for frames with voiced speech), the segment length, and 4 segment-level features: relative and absolute jitter and shimmer measures. Each recorded call C_i, is represented by N_i feature vectors, where N_i is the number of segments for call i.

Classifier: The classifier used in the analysis is a support vector machine (SVM) [26] with linear and radial-basis-function (RBF) kernels, implemented using LIBLINEAR [27] and LIBSVM [28], respectively. The RBF kernel parameter were tuned over the range $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ on the participant-independent validation set. The regularization values were tuned for both the linear and RBF implementations over the set $C \in \{100, 10, 1, 0.1, 0.01\}$. The classifiers are trained on the segment-level 51-dimensional features.

For each test call (C_i), we independently classify each of its N_i segments $s_{i,j}$ ($j = 1, ...N_i$). For each segment, we calculate its signed distance to the hyperplane, $d_{i,j}$. We aggregate each distance into a vector D_i. The score for each call is associated with the p^{th} percentile of D_i. The percentile was chosen using the validation set over the range $p \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$.

5. RESULTS AND DISCUSSION

In this section we demonstrate the efficacy of differentiating between hypomanic and euthymic as well as depressed and euthymic speech using a participant-independent training, testing, and validation methodology. Performance is evaluated using the call-level area under the receiver operating characteristic curve (AUC).

Evaluation of Datasets: Table 3 presents the results across the three datasets discussed in Section 3. The results demonstrate that we are able to detect the mood state of individuals for calls recorded during the clinical interactions. We obtain an average AUC of 0.81 ± 0.17 for hypomania and an average AUC of 0.67 ± 0.18 for depression across all participants.

It is expected that the performance of the classification system will decrease, moving from the evaluation call dataset to the day-of and day before/after datasets. The calls recorded on the day-of and the day before/after do not have human assessed labels due to privacy restrictions. We anticipate that calls recorded on the same day as the evaluation call will be well described by the label assigned during the evaluation call. However, it is important to note that clinical interactions are designed or *structured* to uncover underlying mood state, while non-clinical interactions are not in general. It is anticipated that the acoustic expressions in non-clinical or unstructured calls will exhibit mood-modulations more subtle than the clinical calls and that the recognition performance will decrease.

We use two training scenarios for the calls recorded on the day of the evaluation and the days before/after the evaluation (the unstructured datasets): (1) classifier training using only the evaluation call dataset, testing on both unstructured datasets and (2) classifier training over each unstructured dataset individually and testing with held out parts of the same dataset (e.g., training and testing on the day-of assessment calls). Method one asserts that the acoustic modulations that are indicative of mood state in the evaluation call will also be present in the unstructured calls, even if they are more subtle. Method two asserts that even if the symptomology is present in the unstructured calls, the modulations may be different from those exhibited in the evaluation call. Therefore, in order to detect the mood

state, the acoustic patterns in the unstructured data must be modeled directly. If the performance between methods one and two are similar, there is evidence for modulation consistency. If method two outperforms method one, there is evidence for modulation variability.

The results in Table 3 demonstrate that both method one and method two can be used to detect hypomania during the unstructured calls recorded on the day of the evaluation with an AUC of 0.61 ± 0.09 and 0.65 ± 0.14, respectively. The AUC for depression is 0.49 ± 0.08 and 0.59 ± 0.13, for methods one and two respectively. The results suggest that most individuals express mania and depression differently in clinical interactions compared to their personal life.

Table 3. Call-level AUC of binary mood-state classification. Train:Test indicates which dataset (Evaluation (Eval), Day-of (DOf), Day-B/A (DB/A)), was used for training and which for testing:

Part. #	**1**	**2**	**3**	**4**	**5**	**6**	$\mu \pm \sigma$
Train:Test	**Hypomanic vs Euthymic**						
Eval:Eval	.78	–	–	.67	1.0	–	.81±.17
Eval:DOf	.69	–	–	.63	.51	–	.61±.09
DOf:DOf	.66	–	–	.50	.79	–	.65±.14
Eval:DB/A	.48	–	–	.52	.43	–	.47±.05
DB/A:DB/A	.41	–	–	.62	.57	–	.53±.11
	Depressed vs Euthymic						
Eval:Eval	.42	.82	.78	–	–	.67	.67±.18
Eval:DOf	.49	.60	.43	–	–	.43	.49±.08
DOf:DOf	.68	.68	.40	–	–	.60	.59±.13
Eval:DB/A	.5	.47	.42	–	–	.61	.52±.09
DB/A:DB/A	.50	.52	.53	–	–	.34	.52±.13

Most Informative Features: We examine the features that are most informative for classification using feature selection to further our understanding for how speech is affected by hypomanic and depressed mood states. To increase the robustness of the feature selection, we combine the two best performing datasets: evaluation calls and day-of calls, into a single set that contains all calls recorded on the day of the assessment. We perform feature selection using the leave-one-subject-out cross-validation paradigm using greedy forward feature selection for each of the hypomanic vs. euthymic and the depressed vs. euthymic classification problems. The selection only includes features that improve the average and minimum training participant segment-level AUCs and terminates when a further addition no longer yields improvement. The selected features are then used to train a classifier which is evaluated on the held out test participant.

The feature selection process yields different sets of features for each held out participant. Overall, the hypomanic vs. euthymic selection yields an average of 8.3 ± 5.7 features and depressed vs. euthymic 5.2 ± 4.0 features. Of the selected features, the segment-average of the binary VAD was common to all cross-validation folds for both hypomanic and depressed vs. euthymic. An additional three features were common to 3 out of 4 folds of hypomanic classification: standard deviation of the pitch, segment-average of the zero-crossing rate and of the smoothed VAD. While there were two additional features common to 3 of the 5 folds in the depressed classification: absolute jitter and the segment-average of the magnitude of Mel spectrum over the first band. Table 4 presents the resulting call-level AUCs for classifiers trained with only the selected features as well as those trained with all 51 features.

The results demonstrate that with robust feature selection it is possible to separate euthymic speech from hypomanic and depressed using on average approximately $5-8$ features. Feature selection improves our ability to detect depression, while reducing the variance across participants in the detection of hypomania.

The feature selection results highlight the importance of the average binary VAD for the detection of hypomanic and depressed moods. The mean binary VAD is correlated with the vocalization/pause ratio measure, which has been shown in [19] to be lower for depressed speech. Our examination of this measure showed a similar pattern for depressed speech, and also that it tends to be higher for hypomanic speech: we do this by first removing all instances of the feature $\geq 90\%$ since a majority of the segments tend to be significantly voiced regardless of the label, and find that the feature is lowest for depressed ($median(M) = .51$ $\mu \pm \sigma = .46 \pm .32$), higher for euthymic ($M = .63$ $\mu \pm \sigma = .52 \pm .33$), and the highest for hypomanic ($M = .76$ $\mu \pm \sigma = .69 \pm .21$).

Table 4. Call-level AUC of binary mood-state classification using all features or only selected features:

Part. #	**1**	**2**	**3**	**4**	**5**	**6**	$\mu \pm \sigma$
	Hypomanic vs Euthymic						
All Feats	.61	–	–	.37	.84	–	.61±.24
Sel. Feats	.63	–	–	.59	.67	–	.63±.04
	Depressed vs Euthymic						
All Feats	.62	.65	.42	–	–	.65	.59±.11
Sel. Feats	.63	.82	.43	–	–	.67	.64±.16

6. CONCLUSION

This paper presents a new framework for the ecological long-term monitoring of mood states for individuals with BP. We describe our data collection paradigm and our labeling methodology. Our results demonstrate that hypomania and depression can be differentiated from euthymia using speech-based classifiers trained on both structured (the weekly clinical interactions) and unstructured (all other calls) cell phone recordings. The only labels within our data are those associated with structured interactions due to the privacy considerations associated with the continuous recording of cell phone conversational data. We find that our system is most accurate when modeling these structured data. We hypothesize that the relative accuracy of the structured modeling results both from the fact that these calls are the only data directly labeled and the skill with which clinicians evoke underlying mood in their patient interactions. We further demonstrate that the system can detect the presence of hypomania in the unstructured data collected on the same day as the structured interaction. This suggests that the labels assessed during the clinical interactions fit the data recorded during the non-clinical personal interactions in hypomania. We find that our system currently has difficulty detecting the presence of depression outside of the clinical interactions. This may suggest that the acoustic features of depression we studied are associated with the context of questions (asking about depressed moods) compared to hypomanic symptoms.

With the expansion of the UM-PAD, we will gather acoustic data from participants over periods of time up to one year. This will allow us to associate these data with clinical and biological characteristics of the individual and their illness. This additional data will allow us to determine the stability of the observed trends. The ultimate goal is to identify acoustic features that predict impending mood changes with the purpose of providing intervention strategies to prevent mood episodes. Initial results presented in this paper highlight the potential for and the challenges of modeling mood variation in unstructured data collected outside of clinical interactions. The refinement and further development of this technology has the potential to change the manner in which we consider the deployment of health care, particularly in under-resourced communities.

7. REFERENCES

[1] Frederick K Goodwin and Kay Redfield Jamison, *Manic-depressive illness: bipolar disorders and recurrent depression*, vol. 1, Oxford University Press, 2007.

[2] Alan D Lopez, Colin D Mathers, Majid Ezzati, Dean T Jamison, and Christopher JL Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[3] Leah S Kleinman, Ana Lowin, Emuella Flood, Gian Gandhi, Eric Edgell, and Dennis A Revicki, "Costs of bipolar disorder," *Pharmacoeconomics*, vol. 21, no. 9, pp. 601–622, 2003.

[4] Benjamin J Sadock, V. A. Sadock, and P. Ruiz, *Kaplan & Sadock's Comprehensive Textbook of Psychiatry (2 Volume Set)*, lippincott Williams & wilkins, 2009.

[5] Brian S. Helfer, Thomas F. Quatieri, James R. Williamson, Daryush D. Mehta, Rachelle Horwitz, and Bea Yu, "Classification of depression state based on articulatory precision," in *Interspeech, 2013*, 2013.

[6] Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, Michael Breakspear, and Roland Goecke, "Modeling spectral variability for the classification of depressed speech," in *Interspeech, 2013*, 2013.

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

[8] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[9] Theodora Chaspari, Emily Mower Provost, and Shrikanth S. Narayanan, "Analyzing the structure of parent-moderated narratives from children with asd using an entity-based approach," in *Interspeech, 2013*, 2013.

[10] Jan PH Van Santen, Emily T Prud'hommeaux, Lois M Black, and Margaret Mitchell, "Computational prosodic markers for autism," *Autism*, vol. 14, no. 3, pp. 215–236, 2010.

[11] Mohammed E Hoque, Joseph K Lane, Rana El Kaliouby, Matthew Goodwin, and Rosalind W Picard, "Exploring speech therapy games with children on the autism spectrum," 2009.

[12] Daniel Bone, Matthew P Black, Chi-Chun Lee, Marian E Williams, Pat Levitt, Sungbok Lee, and Shrikanth Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist.," in *INTERSPEECH*, 2012.

[13] Frans Sluis, Egon L Broek, and Ton Dijkstra, "Towards an artificial therapy assistant: Measuring excessive stress from speech," 2011.

[14] Egon L van den Broek, Frans van der Sluis, and Ton Dijkstra, "Telling the story and re-living the past: How speech analysis can reveal emotions in post-traumatic stress disorder (ptsd) patients," in *Sensing Emotions*, pp. 153–180. Springer, 2011.

[15] S Tokuno, G Tsumatori, S Shono, E Takei, G Suzuki, T Yamamoto, S Mituyoshi, and M Shimura, "Usage of emotion recognition in military health care," in *Defense Science Research Conference and Expo (DSR), 2011*. IEEE, 2011, pp. 1–5.

[16] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd," in *Interspeech, 2013*, 2013.

[17] Steven F Warren, Jill Gilkerson, Jeffrey A Richards, D Kimbrough Oller, Dongxin Xu, Umit Yapanel, and Sharmistha Gray, "What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism," *Journal of autism and developmental disorders*, vol. 40, no. 5, pp. 555–569, 2010.

[18] DK Oller, P Niyogi, S Gray, JA Richards, J Gilkerson, D Xu, U Yapanel, and SF Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13354–13359, 2010.

[19] 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.

[20] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear, "Characterising depressed speech for classification," in *Interspeech, 2013*, 2013.

[21] Elliot Moore, Mark A Clements, John W Peifer, and Lydia Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 1, pp. 96–107, 2008.

[22] Scott A Langenecker, Erika FH Saunders, Allison M Kade, Michael T Ransom, and Melvin G McInnis, "Intermediate: Cognitive phenotypes in bipolar disorder," *Journal of Affective Disorders*, vol. 122, no. 3, pp. 285–293, 2010.

[23] M Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*, pp. 143–152. Springer, 1986.

[24] RC Young, JT Biggs, VE Ziegler, and DA Meyer, "A rating scale for mania: reliability, validity and sensitivity.," *The British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] Corinna Cortes and Vladimir Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[27] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[28] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.