


Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood

Ellen W. McGinnis, Steven P. Anderau, Jessica Hruschak, Reed D. Gurchiek, Nestor L. Lopez-Duran, Kate Fitzgerald, Katherine L. Rosenblum, Maria Muzik, and Ryan S. McGinnis , *Member, IEEE*

Abstract—Childhood anxiety and depression often go undiagnosed. If left untreated these conditions, collectively known as internalizing disorders, are associated with long-term negative outcomes including substance abuse and increased risk for suicide. This paper presents a new approach for identifying young children with internalizing disorders using a 3-min speech task. We show that machine learning analysis of audio data from the task can be used to identify children with an internalizing disorder with 80% accuracy (54% sensitivity, 93% specificity). The speech features most discriminative of internalizing disorder are analyzed in detail, showing that affected children exhibit especially low-pitch voices, with repeatable speech inflections and content, and high-pitched response to surprising stimuli relative to controls. This new tool is shown to outperform clinical thresholds on parent-reported child symptoms, which identify children with an internalizing disorder with lower accuracy (67–77% versus 80%), and similar specificity (85–100% versus 93%), and sensitivity (0–58% versus 54%) in this sample. These results point toward the future use of this approach for screening children for internalizing disorders so that interventions can be deployed when they have the highest chance for long-term success.

Index Terms—Machine learning, speech analysis, mental health, anxiety, depression.

I. INTRODUCTION

ANXIETY and depression can emerge in children as young as four years old, but symptoms are often overlooked until children can more clearly express their discomfort given the abstract emotions involved [1], and communicate their impairment with help-seeking adults. The current gold standard diagnostic assessment in young children is to conduct a 60–90 minute semi-structured interview with a trained clinician and their primary caregiver. Limitations such as waiting lists and insurance burden may slow the assessment process, and poor parental report of internal child emotions may also prevent many children from receiving appropriate referrals and diagnoses. Many signs of anxiety or depression at this early age are unrecognized by well-intentioned but unknowing parents or are dismissed as transient. However, we know that nearly 20% of children experience an internalizing disorder during childhood [2], [3]. This psychopathology impairs a child's functioning and development [4]–[8] and predicts serious health problems later in life if left untreated (e.g., substance abuse [9]; development of comorbid psychopathology [10], [11]; increased risk for suicide [12]). Thus, there are high individual and societal burdens associated with internalizing disorders [13] that highlight the need for effective early assessment. New tools that can feasibly and objectively screen children for these internalizing disorders during routine pediatric well-visits would support surrounding adults in understanding the intensity and chronicity of their child's distress, and connect them with interventions early in development, when neuroplasticity and potential for symptom improvement is greatest [14].

Mood induction tasks have been increasingly used in research contexts to “press” for anxious, frustrating, joyful, or saddening affect. A child's behavioral and physiological response to these tasks is recorded using a variety of technologies (i.e., video-recorded and coded behaviors and directly measured cortisol, heart rate variability, electrodermal activity, movement), manually processed, and studied in relation to theory-driven expectations (e.g., see [15]–[20]). The Trier-Social Stress Task (TSST) is one such mood induction task meant to induce performance anxiety by having a participant give a short, improvisational

Manuscript received November 6, 2018; revised February 26, 2019; accepted April 23, 2019. Date of publication April 26, 2019; date of current version November 6, 2019. The work of E. W. McGinnis was supported in part by the Biomedical and Social Sciences Scholar Program and in part by the Blue Cross Blue Shield of Michigan Foundation Grant (1982.SAP). The work of K. Fitzgerald was supported in part by the Brain Behavior Research Foundation and in part by NIMH under Grant R03MH102648. The work of K. L. Rosenblum was supported by the NIMH under Grant R03MH102648. The work of M. Muzik was supported in part by NIMH under Grant K23-MH080147 and in part by the Michigan Institute for Clinical and Health Research under Grant UL1TR000433. (E. W. McGinnis and S. Anderau contributed equally to this work.) (Corresponding author: Ryan S. McGinnis.)

E. W. McGinnis is with the Department of Child Psychiatry, University of Vermont, Burlington, VT 05405 USA (e-mail: ellen.mcginis@med.uvm.edu).

S. P. Anderau, R. D. Gurchiek, and R. S. McGinnis are with the Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, VT 05405 USA (e-mail: steven.anderau@uvm.edu; reed.gurchiek@uvm.edu; ryan.mcginis@uvm.edu).

J. Hruschak, K. Fitzgerald, K. L. Rosenblum, and M. Muzik, are with the Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jlhrus@med.umich.edu; krd@med.umich.edu; katier@med.umich.edu; muzik@med.umich.edu).

N. L. Lopez-Duran is with the Department of Clinical Psychology, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: nestorl@umich.edu).

Digital Object Identifier 10.1109/JBHI.2019.2913590

speech to a confederate audience pretending to be thoroughly bored and critical. Behavioral coding and physiological measures have not only been associated with task affect, but also with mental illness (anxiety and depression) more generally [21]–[23]. While useful for research, results from objective assessment of this task are not generally meant for immediate clinical utility due to long and costly processing requiring specialized equipment.

However, recently researchers have identified voice audio signal characteristics that enable the detection of affect [24] and signs of mental health diagnoses (depression, anxiety, bipolar disorder) in task-specific and daily life recordings [25]–[27]. The goals of these projects are often to use voice analysis to detect the onset of mania [27] and/or suicidal intent [28] in adults and ultimately aid individuals in connecting with urgent intervention. When taken with results from the TSST, these outcomes point toward development of a similar approach for detecting childhood anxiety and depression based on voice audio recordings during mood induction tasks.

To identify children with an internalizing disorder using data from such tasks, a model of the complex relationship between objective voice audio signal characteristics and diagnosis must be established. A data-driven approach, like machine learning, is ideally suited for this task. In fact, our previous work has explored the use of wearable inertial sensors for identifying children with internalizing psychopathology based on machine learning analysis of their movement patterns during a fear induction task [29]–[31]. This approach allows for the realization of much more complex relationships than would be possible from theory-based modeling alone. While these efforts form one facet of the burgeoning field of digital medicine [32], [33], the use of these techniques for improving childhood mental health is just beginning.

The space for mental illness detection via fast, low cost, and feasible technologies has the potential to give voice to populations who have challenges understanding and expressing their own distress and seeking appropriate help. In adults, this is salient for individuals with mania and suicidal intent. However, these methods can also serve any child under the age of eight who has more difficulty understanding and communicating their abstract emotions. While some groundbreaking work has been done in identifying affect in children between the ages of 10 and 13 [34] and in capturing conversational reciprocity in children with autism [35], the community has yet to employ the analysis of vocal features within child speech epochs for *identifying* children with underlying psychopathology.

Herein, we employ speech analysis of child voice recordings during a 3-minute speech task and machine learning to detect clinically-derived anxiety and depressive diagnoses in children between the ages of 3 and 7 years old. In the following sections, we describe the experiment, data processing, model development and evaluation approaches (Section II), present performance characteristics of a common parent-report questionnaire and the machine learning models for identifying children with internalizing disorders (Section III), and discuss these results and their implications for screening young children for internalizing psychopathology (Section IV).

II. METHODS

A. Procedure

Studies had approval from the University of Michigan Institutional Review Board (HUM00091788; HUM00033838). Child and caregiver were brought into the university-based laboratory and provided written consent to complete a battery of tasks. Caregivers completed self- and parent-report questionnaires and a diagnostic interview to assess for child psychiatric diagnoses while children underwent a series of behavioral tasks in an adjacent room. Behavioral tasks were designed to elicit fear responses and positive affect. Experimenters also conducted a home visit where additional behavioral tasks were performed. Participants were compensated for their time.

Herein, we consider a subset of data from the larger study by examining participant response to the Speech Task, a behavioral task designed to elicit an anxiety response, as well as questionnaires and a diagnostic interview used to assess internalizing symptoms and diagnoses.

B. Clinical Measures

The Speech Task is an adapted version of the Trier Social Stress Task for children (TSST-C), which has been shown to induce anxiety in children 7 and older [36]. This task, which was conducted during the home visit, is standardized, and all research assistants were trained to carry out the task according to protocol including displaying flat affect through the duration of the task. In the Speech Task, participants are instructed to prepare and give a three-minute speech and are told that they will be judged based on how interesting it is. They are given three minutes to prepare, and then begin their three-minute speech. A buzzer is used to interrupt the participant's speech with 90 and 30 seconds remaining in the task. At each interruption, the experimenter informs the participant of the time remaining in the task using a standardized script. The experimenter responds to participant questions as necessary. Each speech was recorded using a standard video camera, truncated to include just the three-minute speech task, and the audio was extracted for further analysis.

A structured clinical interview 1–2 hours in duration was conducted by master's level psychology students with the caregiver. The Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime Version (K-SADS-PL) is validated for children 5–18 with evidence to suggest validation for preschoolers [37], [38]. Percent agreement for all diagnoses ranges from 93% to 100%, test-retest reliability ranges from .63 to .90. Final diagnoses were derived via clinical consensus using the best-estimate procedures [39] based on the child and parent report, family history, and other self-report symptom checklists. K-SADS is considered gold standard, but not commonly completed in typical practice, which more often than not includes an unstructured 1-hour interview with a psychologist.

The Child Behavior Checklist 1.5–5 and 6–18 [40] is a 15-minute parent-report questionnaire designed to screen for externalizing and internalizing problem behaviors in children

ages 1.5 to 18 in both clinical and research environments [41]. The scale consists of 120 items related to behavior problems across multiple domains. Responses result in global scores for externalizing, internalizing, and total problems, as well as disorder-based subscales (Anxiety, Depression, Attention Deficit/Hyperactivity, Oppositional Defiant Disorder). Only scales available in both versions (ages 1.5–5 and 6–18) were used in subsequent analyses. The CBCL has well established validity and reliability (see [40]). Externalizing and internalizing broadband scores standardized into ‘T scores’ for the respective child ages and gender. A T score of 65–69 indicates “borderline” and a T score of 70 or greater indicates “above threshold” and thus likely clinical risk.

C. Audio Data Processing

Audio data from the speech task were sampled at 48 kHz and processed via a voice activity detector (VAD) that discriminates instances of speaking from background noise. The VAD operates on signal energy and has been designed to have a high sensitivity towards speech. Speech epochs were identified when energy within a sliding window was above the baseline noise. Identified raw speech epochs, which included full sentences, phrases, phonemes, and high energy noise, were smoothed using a median filter with window length of 0.21 seconds. This ensured that natural pauses in speech were contained within a single speech epoch and that short-duration phonemes and noise were removed. Due to the realities of collecting data from children in the home, many recordings had low signal-to-noise ratios (SNRs) and were corrupted by significant harmonic background noise. Thus, each audio file and its detected speech epochs were screened manually for quality. In this process, two research assistants manually labeled each speech epoch detected by the VAD as containing audio from the participant, clinician, buzzer, or another source. Discrepancies between labels were resolved by a third researcher. All audio files containing participant speech were then subjectively classified by a single researcher into one of four categories: 1) High Quality, defined as having “moderate to very strong representation of speech content and frequency” ($N = 43$, Age: 74 ± 13 months, Sex: 26-F, 17-M, 13 with internalizing disorders), 2) Low Quality, defined as having “very poor to poor representation of speech content and frequency” ($N = 28$, Age: 68 ± 13 months, Sex: 19-F, 9-M, 7 with internalizing disorders), 3) Participant Deviation, where the participant did not speak during the task ($N = 9$, Age: 66 ± 11 months, Sex: 4-F, 5-M, 6 with internalizing disorders), or 4) Protocol Deviation, where the task was not conducted as planned (e.g., task did not last 3 minutes, buzzers were not played). Data from only the High Quality and Low Quality groups are considered in the following analyses.

D. Participants

Data in the High Quality and Low Quality groups were collected from 71 children (63% female) and their primary caregivers (95% mothers). Participants were recruited from either an ongoing observational study (Bonding Between Mothers and

Children, PI: Maria Muzik) or from flyers posted in the community and psychiatry clinics to obtain a sample with a wide range of symptom presentations. Eligible participants were children between the ages of 3 and 8 who spoke fluent English and whose caregivers were 18 years and older. Exclusion criteria were suspected or diagnosed developmental disorder (e.g., autism), having a serious medical condition, or taking medications that affect the central nervous system. The resulting sample of children were aged between 3 and 7 years ($M = 5.25$ $SD = 1.14$), was 65% White non-Latinx (19% Multiracial, 11% African American, 5% Other), and 83% lived in two-parent households. Annual household income was \$65–70k (27% greater than \$100k, and 6% less than \$25k).

Multimodal assessments, including diagnostic interviews, were conducted between August 2014 and August 2015. Based on these multimodal assessments and consensus coding, $N = 20$ participants were identified as having an internalizing diagnosis (current ($N = 17$), past ($N = 3$)) according to DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 4th. Edition).

E. Audio Features

To characterize the ability of the proposed approach for identifying children with an internalizing disorder, we first partitioned each three-minute speech task into three phases, the boundaries of which were defined by the buzzer interruptions inherent to the task. To parameterize the audio signal within each phase, we computed the following features for each speech epoch: speech epoch duration, zero crossing rate (ZCR) of the audio signal, Mel frequency cepstral coefficients (MFCC), dominant frequency, mean frequency, perceptual spectral centroid (PSC), spectral flatness, skew and kurtosis of the power spectral density (PSD), ZCR of the z-score of the PSD (ZCR zPSD) for all speech epochs, the first, second, and third formants, and the percentage of signal energy above 200, 500, 700, 1000, and 2000 Hz. We also extracted the mean, median, standard deviation, maximum, and minimum ZCR zPSD from sliding windows in the time and frequency domains within each speech epoch (ZCR zPSD_{sw}). Descriptive statistics (i.e., mean, standard deviation, median, maximum, minimum) were computed for each feature within each phase. We also computed the number of speech epochs completed by the patient and the clinician yielding a total of 164 features for each of the three phases. Signal processing and feature extraction were performed in MATLAB (Mathworks, Natick, MA, USA). Many of these features have been proposed previously in the literature for identifying anxious and depressive symptoms in adults [24]–[27].

F. Model Development and Analysis

Binary classification models (logistic regression – LR, support vector machine with a linear kernel – SL, support vector machine with a gaussian kernel – SG, random forest – RF) relating the audio signal features from each phase to internalizing disorder determined via K-SADS-PL with clinical consensus were trained using a supervised learning approach on the data classified as High Quality. Classifier performance was estab-

lished using leave-one-subject-out (LOSO) cross validation. In this approach, data from all but one participant ($N = 42$) were partitioned into a training dataset and converted to z-scores prior to performing Davies-Bouldin Index [42] based feature selection. This yields the eight features with zero mean and unit variance that best discriminate between diagnostic groups. Thus, 42 observations of these eight features were used to train the binary classification models for predicting internalizing diagnosis. The same eight features were extracted, converted to z-scores based on parameters (e.g., mean, variance) from the training set, and used as input to the model for predicting the diagnosis of the one remaining test subject. The score threshold to determine diagnosis was set for each iteration using the Number Needed to Misdiagnosis criteria [43] based on the ROC curve of the training data. This measure, which is maximized to find the appropriate threshold, is an estimate of the number of patients who need to be tested in order for one to be misdiagnosed. This process was repeated 42 times until the diagnosis of each subject had been predicted.

The resulting predicted diagnoses were used to compute standard measures of classification performance including accuracy, sensitivity, and specificity. We also computed the area under the receiver operating characteristic (ROC) curve (AUC) to comment on the general discriminative ability of the classifiers. We further examined the eight features selected most often for discriminating between participants with and without an internalizing disorder to provide an indication of the speech cues indicative of underlying psychopathology.

A permutation test was used to determine if the classification error rates (error rate = number of incorrect predictions/total number of predictions = $1 - \text{classification accuracy}$) were significantly different from what we would observe due to random chance. To complete this test, we approximated the distribution of possible error rates for each logistic regression model as a beta distribution parameterized by the number of incorrect predictions and the total number of observations, as indicated in [44], [45], and randomly sampled 100 possible error rates from this distribution. Next, we repeated the model training process outlined previously for 100 random permutations of the diagnostic labels, computing the classification error rate for each. Finally, a Mann-Whitney U-test was used to identify the temporal phases that yield classification models with error rates significantly different from those expected by chance from this dataset.

To place the results from the high-quality data in context, we conducted two additional analyses. First, we trained classification models on all of the data labeled as High Quality and used it to predict subject diagnoses from the data labeled as Low Quality. The resulting predictions were used to compute accuracy, sensitivity, specificity, and AUC for comparison to the results from the High Quality data. We also examined the utility of the CBCL as a screening tool for internalizing disorders (according to the K-SADS-PL with clinical consensus) in this sample using previously established clinical cutoffs ($T \text{ score} \geq 70$) for manualized use [41] as well as a more conservative cutoff ($T \text{ score} \geq 55$) suggested for improving screening efficiency [46].

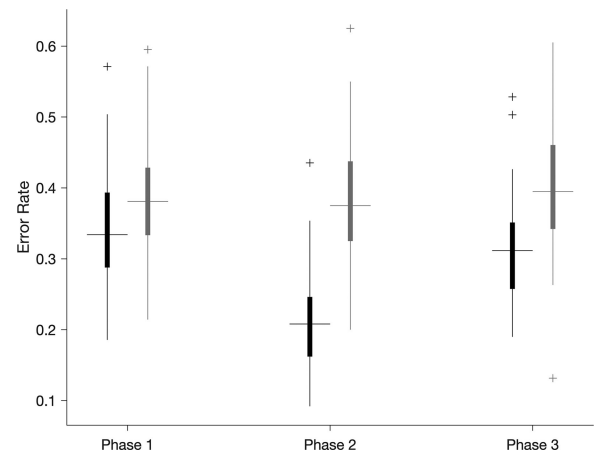


Fig. 1. Boxplots of error rates for logistic regression models trained to detect children with internalizing disorders as determined via K-SADS-PL with clinical consensus (black) compared to those due to chance (gray) for each temporal phase of the task. All models perform better than random chance.

TABLE I
PERFORMANCE CHARACTERISTICS FOR HIGH- AND LOW-QUALITY DATA

Data Model	High Quality				Low Quality			
	LR	SL	SG	RF	LR	SL	SG	RF
ACC	0.80	0.80	0.68	0.70	0.57	0.53	0.76	0.67
AUC	0.75	0.78	0.72	0.74	0.43	0.36	0.55	0.49
SEN	0.54	0.62	1.00	0.15	0.69	0.69	1.00	0.88
SPE	0.93	0.89	0.00	0.96	0.20	0.00	0.00	0.00

III. RESULTS

Figure 1 illustrates the results of permutation testing for logistic regression models developed using data from each phase of the task. The distribution of error rates expected from the actual models are shown in black while those expected due to random chance are shown in gray. Models trained on data from each phase have statistically lower error rates than expected by chance ($p < 0.01$). Only results from models trained on data from Phase 2 (best performing) will be considered in the analyses that follow.

Table I reports classification performance in terms of accuracy, sensitivity, specificity, and AUC as determined using the LOSO approach on high-quality data as well as on the low-quality data.

For comparison, we also report the performance in terms of accuracy, sensitivity, and specificity for the internalizing broadband (Int) and two DSM-oriented scales (anxiety problems (Anx), depressive problems (Dep)) of the CBCL using previous established clinical ($T \text{ score} \geq 70$) and more conservative ($T \text{ score} \geq 55$) cutoffs in Table II.

Next, we examine the eight features used as input to the model trained on data from Phase 2 of the task. The boxplot of Fig. 2 illustrates the distribution of each of the eight features across subjects with (gray) and without (black) an internalizing disorder and the directionality of any difference between groups.

Across the 43 iterations of the LOSO cross validation, the eight features reported in Table III were selected a median of

TABLE II
PERFORMANCE CHARACTERISTICS OF THE CBCL

Cutoff	Accuracy		Sensitivity		Specificity	
	70	55	70	55	70	55
Int	0.67	0.74	0.00	0.42	0.96	0.89
Anx	0.72	0.69	0.17	0.33	0.96	0.85
Dep	0.77	0.77	0.25	0.58	1.00	0.85

Accuracy, sensitivity, and specificity of the internalizing broadband (Int), anxious problems (Anx), and depressive problems (Dep) subscales of the CBCL with clinical thresholds (T score ≥ 70) for manualized use and a more conservative cutoff (T score ≥ 55) suggested for improving screening efficiency relative to diagnosis established via K-SADS-PL with clinical consensus.

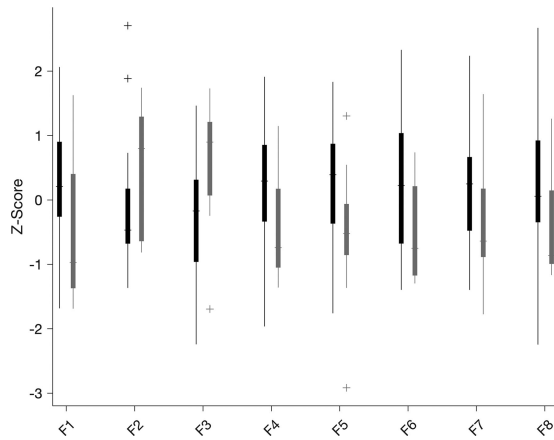


Fig. 2. Boxplots of selected features from subjects with high quality data. Illustrates separation and directionality of differences between subjects with (gray) and without (black) an internalizing diagnosis.

TABLE III
FEATURES SELECTED FOR DISCRIMINATING BETWEEN DIAGNOSTIC GROUPS

Label	Feature	High	Low
F1	Initial speech epoch MFCC (267 – 400 Hz)	Low pitch tone	High pitch tone
F2	Minimum, average, standard deviation of ZCR zPSDsw	Higher pitch variance within a speech epoch	Lower pitch variance within a speech epoch
F3	Mean MFCC (667 – 800 Hz)	Low pitch tone	High pitch tone
F4	Maximum, standard deviation, average of ZCR zPSDsw	Chaotic speech inflection and content across epochs	Repetitive speech inflection and content across epochs
F5	Mean speech epoch PSC	Broad, full timbre	Breathy, light timbre
F6	Median MFCC (733 – 867 Hz)	High pitch tone	Low pitch tone
F7	Median, average, standard deviation of ZCR zPSDsw	High pitch variance within speech epoch	Low pitch variance within speech epoch
F8	Mean ZCR zPSD	Chaotic inflection and content within speech epoch	Repetitive inflection and content within speech epoch

Features from the second phase of the task selected most often by the feature selection process for discriminating between subjects with and without an internalizing disorder. Descriptions of the speech characteristics of subject's with high and low z-score values are included for each to ease interpretation of the model.

38.5 times. Four of the selected features (F2, F4, F7, F8) are derived from the ZCR of the zPSD, and three are derived from the MFCC (F1, F3, F6). A qualitative analysis of how each of these features relate to the observed speaking patterns is also reported in [Table III](#).

IV. DISCUSSION

Here we explore the use of audio data from a three-minute mood induction task and machine learning for identifying young children with internalizing psychopathology. We examine the performance of several statistical models for this task, explore the effect of low data quality on this performance, and compare the performance to that from the best available tool that could be used to screen for childhood psychopathology. Finally, we examine the eight features selected by the feature selection process to lend clinical interpretability to the model.

As reported in [Table I](#), logistic regression and support vector machine models were able to identify children with an internalizing disorder with 80% accuracy (LR: 54% sensitivity, 93% specificity; SL: 62% sensitivity, 89% specificity), and with moderate [47] AUCs of 0.75 and 0.78, respectively. Notably, in the case of the low-quality data, the logistic regression and support vector machine models were only able to identify subjects with an internalizing disorder with 57% and 53% accuracy (LR: 69% sensitivity, 20% specificity; SL: 69% sensitivity, 0% specificity), and with failing AUCs of 0.43 and 0.36, respectively. Data were identified as low-quality if there was significant background noise present during the speech, if siblings often interrupted the task, or if the subjects engaged in a non-speaking activity like playing an instrument. This collection procedure was not designed to yield high quality audio data, and thus it is not surprising that almost 40% of the recordings were considered low quality. These results highlight the need for refining the protocol for this assessment so that it provides a suitable environment for collecting high-quality audio data while minimizing the distractions (e.g., siblings, musical instruments, background noise) with which a subject can engage during the task. This could be achieved by making the recordings in more constrained settings like a physician's office, or by using recording equipment designed for capturing high-quality audio data.

In comparing performance characteristics across models in [Table I](#), the linear models (LR, SL) had very similar performance. However, the non-linear models (SG, RF) did not perform as well as their linear counterparts on the high-quality data, achieving accuracies of only about 70%. While these models demonstrated superior performance on the low-quality data (SG: 76% accuracy; RF: 67% accuracy), they had very low specificity (0% for each), further supporting the need to refine the protocol to yield higher-quality audio data.

These results compare favorably to our previous work, where k-nearest neighbor ($k = 3$) and logistic regression models were able to achieve a maximum accuracy of 81% (67% sensitivity, 88% specificity), using data from the first 20-seconds of a fear induction task. In this phase, an experimenter led a child into an unknown and dimly lit room and their motion was measured

with a wearable inertial sensor [29]–[31]. The logistic regression employed herein achieves a slightly higher specificity (93%), but with slightly lower sensitivity (54%) and accuracy (80%). The complementary nature of these performance characteristics suggests that these objective measures could be combined to form an assessment battery, and the model outputs fused to yield a composite prediction of childhood psychopathology that inherits the best qualities from each.

The results also compare favorably to the psychometric properties of the questionnaire-based, parent-reported CBCL on child internalizing disorder as determined via K-SADS-PL with clinical consensus. CBCL-derived models for both cutoffs (55 and 70) exhibited lower classification accuracy (67–77 vs. 80%), and similar specificity (85–100 vs. 93%) and sensitivity (0–58 vs. 54%) compared to the model developed using machine learning (see results of Table II). CBCL subscale psychometrics in our study are similar to those from much larger studies (e.g., see [47], [48]). Notably, both in our study and paralleled in these previous studies is the varied sensitivity of the CBCL, with some samples exhibiting sensitivities as low as 0–38% [49] and some as high as 44–86% [50]. Overall, CBCL internalizing psychometrics across studies suggest room for improvement in internalizing screening efficiency [48]. The results presented herein yield improvements in accuracy and sensitivity with comparable specificity and based solely on a 3-minute mood induction task. These results suggest that this objective data may be especially helpful for increasing screening efficiency while minimizing the bias of parent-report.

The feature selection process identified features that related to both the child's reaction to specific prompts within the task (F1), their vocal tone and timbre (F3, F5, F6) and the tonal complexity (F2, F4, F7, F8). Children with an internalizing disorder generally had higher values of F2 and F3 and lower values of F1 and F4–F8 (see boxplots of Fig. 2). Low values of F1 were exhibited by subjects with a higher pitched voice in the first speech epoch directly following the first time they were interrupted by the buzzer (start of Phase 2). High values of F2 were exhibited by subjects who had high pitch variance within their speech epochs. Subjects with high values of F3 had more bass present in their vocal tone on average. Subjects with low values of F4 and F8 exhibited repeatable speech inflections and content across speech epochs. Low values of F5 were exhibited by subjects who had breathy or light timbres. Low values of F6 corresponded to subjects with lower pitch across epochs. Interestingly, low values for F7 were exhibited by subjects who had low pitch variance within their speech epochs. This apparent disagreement may be explained by considering the boxplots of Fig. 2, which illustrate the significant overlap between diagnostic groups for F2 which is less apparent for F7 suggesting that perhaps subjects with an internalizing disorder who exhibit low F2 are still able to be identified because of similarly low F7. This qualitative assessment of audio features begins to paint a picture of the sound of childhood internalizing disorders with affected children generally exhibiting low-pitch (F3, F6) voices with repeatable speech inflections and content (F4, F8) and high-pitched response to surprising stimuli (F1).

These speech patterns parallel previous relationships between depression and anxiety and other physiological metrics (e.g., behavior, cognition, cortisol, electromyography). For instance, behaviorally, one study has mothers speak in a low, monotone voice to mimic depression (known for flat affect) to their infants [51], and cognitively, depression and anxiety are known for repetitive thinking or rumination [52], which may be reflected in their speech patterns. Finally, children with depression and anxiety have been shown to have heightened response to stress (cortisol [21]) and startle (electromyography [53]). However, more evidence is needed to confirm these theorized speculations. Nevertheless, these results provide, for the first time, valuable insight into the association between child speech patterns and anxiety and depression.

Overall, this paper describes a methodology requiring very limited computational resources (e.g., compute 8 features from a small subset of 3 minutes of audio data, use as input to a logistic regression) which points toward future deployment of this technique for identifying young children with probable internalizing diagnoses using resource-constrained but ubiquitous devices like mobile phones. This new approach reduces the time required for screening while also establishing high accuracy, which can help to reduce barriers and better alert families to the need for child mental health services. While these results can likely be improved and extended, and should be replicated, this is an important first step in connecting often overlooked children [54], [55] to the help they need to both mitigate their current distress and prevent subsequent comorbid emotional disorders and additional negative sequelae [10], [11], [56].

This study is not without limitations. Future research should replicate and investigate our claims in a larger study with equal gender representation, and with subjects at varying levels of risk for developing an internalizing disorder. Additionally, a larger sample size would allow examination of specific internalizing disorders to explore whether one disorder type yields different speech patterns than another.

V. CONCLUSION

The results presented herein suggest that a machine learning analysis of child speaking patterns during a short anxiety induction task is able to identify children with internalizing psychopathology. This statistical classification model outperforms clinical thresholds on parent-reported child symptoms collected with the CBCL, indicating its potential as an objective screening tool in this population. A detailed analysis of the audio features selected for this classification indicated that affected children exhibit low-pitch voices, with repetitive inflection and content, and high-pitched response to surprising stimuli.

ACKNOWLEDGMENT

The authors would like to acknowledge E. Bilek, K. Ip, D. Morelen, J. Lawler, K. Khosravi, and A. Dewoolkar for their efforts on the larger study that has led to this work.

REFERENCES

- [1] T. E. Chansky and P. C. Kendall, "Social expectancies and self-perceptions in anxiety-disordered children," *J. Anxiety Disorders*, vol. 11, no. 4, pp. 347–363, Aug. 1997.
- [2] H. L. Egger and A. Angold, "Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology," *J. Child Psychol. Psychiatry*, vol. 47, no. 3/4, pp. 313–337, Apr. 2006.
- [3] S. J. Bufferd, L. R. Dougherty, G. A. Carlson, and D. N. Klein, "Parent-reported mental health in preschoolers: Findings using a diagnostic interview," *Comprehensive Psychiatry*, vol. 52, no. 4, pp. 359–369, 2011.
- [4] J. L. Luby, A. C. Belden, J. Pautsch, X. Si, and E. Spitznagel, "The clinical significance of preschool depression: Impairment in functioning and clinical markers of the disorder," *J. Affect. Disorders*, vol. 112, no. 1–3, pp. 111–119, Jan. 2009.
- [5] N. R. Towe-Goodman, L. Franz, W. Copeland, A. Angold, and H. Egger, "Perceived family impact of preschool anxiety disorders," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 53, no. 4, pp. 437–446, Apr. 2014.
- [6] A. C. Belden, M. S. Gaffrey, and J. L. Luby, "Relational aggression in children with preschool-onset psychiatric disorders," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 51, no. 9, pp. 889–901, Sep. 2012.
- [7] S. J. Bufferd, L. R. Dougherty, G. A. Carlson, S. Rose, and D. N. Klein, "Psychiatric disorders in preschoolers: Continuity from ages 3 to 6," *Amer. J. Psychiatry*, vol. 169, no. 11, pp. 1157–1164, Nov. 2012.
- [8] D. J. Whalen, C. M. Sylvester, and J. L. Luby, "Depression and anxiety in preschoolers: A review of the past 7 years," *Child Adolescent Psychiatry Clin. North Amer.*, vol. 26, no. 3, pp. 503–522, Jul. 2017.
- [9] S. N. Compton, B. J. Burns, L. E. Helen, and E. Robertson, "Review of the evidence base for treatment of childhood psychopathology: Internalizing disorders," *J. Consulting Clin. Psychol.*, vol. 70, no. 6, pp. 1240–1266, Dec. 2002.
- [10] A. Bittner, H. L. Egger, A. Erkanli, E. Jane Costello, D. L. Foley, and A. Angold, "What do childhood anxiety disorders predict?," *J. Child Psychol. Psychiatry*, vol. 48, no. 12, pp. 1174–1183, Dec. 2007.
- [11] D. A. Cole, L. G. Peeke, J. M. Martin, R. Truglio, and A. D., "A longitudinal look at the relation between depression and anxiety in children and adolescents," *J. Consulting Clin. Psychol.*, vol. 66, no. 3, pp. 451–460, 1998.
- [12] M. S. Gould *et al.*, "Psychopathology associated with suicidal ideation and attempts among children and adolescents," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 37, no. 9, pp. 915–923, Sep. 1998.
- [13] A. Konnopka, F. Leichenring, E. Leibling, and H.-H. König, "Cost-of-illness studies and cost-effectiveness analyses in anxiety disorders: A systematic review," *J. Affect. Disorders*, vol. 114, no. 1, pp. 14–31, Apr. 2009.
- [14] J. L. Luby, "Preschool depression: The importance of identification of depression early in development," *Current Dir. Psychol. Sci.*, vol. 19, no. 2, pp. 91–95, May 2010.
- [15] E. W. McGinnis *et al.*, "Wearable sensors detect childhood internalizing disorders during mood induction task," *PLOS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0195598.
- [16] E. McGinnis *et al.*, "Movements indicate threat response phases in children at-risk for anxiety," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 5, pp. 1460–1465, Sep. 2017.
- [17] N. L. Lopez-Duran, N. J. Hajal, S. L. Olson, B. T. Felt, and D. M. Vazquez, "Individual differences in cortisol responses to fear and frustration during middle childhood," *J. Exp. Child Psychol.*, vol. 103, no. 3, pp. 285–295, Jul. 2009.
- [18] N. A. Fox, H. A. Henderson, P. J. Marshall, K. E. Nichols, and M. M. Ghera, "Behavioral inhibition: Linking biology and behavior within a developmental framework," *Annu. Rev. Psychol.*, vol. 56, pp. 235–262, 2005.
- [19] D. C. Fowles, G. Kochanska, and K. Murray, "Electrodermal activity and temperament in preschool children," *Psychophysiology*, vol. 37, no. 6, pp. 777–787, Nov. 2000.
- [20] S. D. Calkins, S. E. Dedmon, K. L. Gill, L. E. Lomax, and L. M. Johnson, "Frustration in infancy: Implications for emotion regulation, physiological processes, and temperament," *Infancy*, vol. 3, no. 2, pp. 175–197, 2002.
- [21] N. L. Lopez-Duran, E. McGinnis, K. Kuhlman, E. Geiss, I. Vargas, and S. Mayer, "HPA-axis stress reactivity in youth depression: Evidence of impaired regulatory processes in depressed boys," *Stress*, vol. 18, no. 5, pp. 545–553, Sep. 2015.
- [22] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, "Depression and cortisol responses to psychological stress: A meta-analysis," *Psychoneuroendocrinology*, vol. 30, no. 9, pp. 846–856, Oct. 2005.
- [23] C. Garcia-Leal *et al.*, "Anxiety and salivary cortisol in symptomatic and nonsymptomatic panic patients and healthy volunteers performing simulated public speaking," *Psychiatry Res.*, vol. 133, no. 2, pp. 239–252, Feb. 2005.
- [24] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10, pp. 787–800, Oct. 2007.
- [25] N. Cummins, "An investigation of depressed speech detection: Features and normalization," in *Proc. INTERSPEECH*, 2011, pp. 6–9.
- [26] P. Laukka *et al.*, "In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech," *J. Nonverbal Behav.*, vol. 32, no. 4, Jul. 2008.
- [27] Z. N. Karam *et al.*, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *Proc. 2014 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4858–4862.
- [28] M. E. Larsen *et al.*, "The use of technology in suicide prevention," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, vol. 2015, pp. 7316–7319.
- [29] R. S. McGinnis *et al.*, "Wearable sensors and machine learning diagnose anxiety and depression in young children," in *Proc. IEEE Int. Conf. Biomed. Health Inf.*, 2018, pp. 410–413.
- [30] R. S. McGinnis *et al.*, "Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 3983–3986.
- [31] R. S. McGinnis *et al.*, "Rapid detection of internalizing diagnosis in young children enabled by wearable sensors and machine learning," *PLOS ONE*, vol. 14, no. 1, p. e0210267, Jan. 2019.
- [32] E. Elenko, L. Underwood, and D. Zohar, "Defining digital medicine," *Nature Biotechnol.*, vol. 33, pp. 56–461, May 12, 2015. [Online]. Available: <https://www.nature.com/articles/nbt.3222>, Accessed on: May 21, 2018.
- [33] E. J. Topol, S. R. Steinhubl, and A. Torkamani, "Digital medical tools and sensors," *JAMA*, vol. 313, no. 4, pp. 353–354, Jan. 2015.
- [34] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH 2009, 10th Ann. Conf. Int. Speech Commun. Association*, Brighton, U.K., ISCA, ISCA, Sep. 2009, pp. 312–315.
- [35] S. F. Warren *et al.*, "What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism," *J. Autism Develop. Disorder*, vol. 40, no. 5, pp. 555–569, May 2010.
- [36] A. Buske-Kirschbaum, S. Jobst, A. Wustmans, C. Kirschbaum, W. Rauh, and D. Hellhammer, "Attenuated free cortisol response to psychosocial stress in children with atopic dermatitis," *Psychosom. Med.*, vol. 59, no. 4, pp. 419–426, Aug. 1997.
- [37] B. Birmaher *et al.*, "Schedule for affective disorders and schizophrenia for school-age children (K-SADS-PL) for the assessment of preschool children- A preliminary psychometric study," *J. Psychiatry Res.*, vol. 43, no. 7, pp. 680–686, Apr. 2009.
- [38] J. Kaufman *et al.*, "Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): Initial reliability and validity data," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 36, no. 7, pp. 980–988, Jul. 1997.
- [39] M. Maziade *et al.*, "Reliability of best-estimate diagnosis in genetic linkage studies of major psychoses: Results from the Quebec pedigree studies," *Amer. J. Psychiatry*, pp. 1674–1686, 1992.
- [40] T. M. Achenbach and L. Rescorla, *ASEBA School-Age Forms & Profiles*. Burlington, VT, USA: Univ. Vermont Res. Center Children, Youth, Families, 2001.
- [41] T. M. Achenbach, C. T. Howell, H. C. Quay, and C. K. Conners, "National survey of problems and competencies among four- to sixteen-year-olds: Parents' reports for normative and clinical samples," *Monographs Soc. Res. Child Develop.*, vol. 56, no. 3, pp. v–120, 1991.
- [42] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
- [43] F. Habibzadeh and M. Yadollahie, "Number needed to misdiagnose: A measure of diagnostic test effectiveness," *Epidemiol. Cambridge Mass.*, vol. 24, no. 1, p. 170, Jan. 2013.
- [44] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artif. Intell. Rev.*, vol. 44, no. 4, pp. 467–508, Dec. 2015.
- [45] A. Isaksson, M. Wallman, H. Göransson, and M. G. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognit. Lett.*, vol. 29, no. 14, pp. 1960–1965, Oct. 2008.

- [46] J. J. Hudziak, W. Copeland, C. Stanger, and M. Wadsworth, "Screening for DSM-IV externalizing disorders with the child behavior checklist: A receiver-operating characteristic analysis," *J. Child Psychol. Psychiatry*, vol. 45, no. 7, pp. 1299–1307, Oct. 2004.
- [47] N. de la Osa, R. Granero, E. Trepal, J. M. Domenech, and L. Ezpeleta, "The discriminative capacity of CBCL/1½-5-DSM5 scales to identify disruptive and internalizing disorders in preschool children," *Eur. Child Adolescent Psychiatry*, vol. 25, no. 1, pp. 17–23, Jan. 2016.
- [48] E. M. Warnick, M. B. Bracken, and S. Kasl, "Screening efficiency of the child behavior checklist and strengths and difficulties questionnaire: A systematic review," *Child Adolescent Mental Health*, vol. 13, no. 3, pp. 140–147, Sep. 2008.
- [49] C. W. Rishel, C. Greeno, S. C. Marcus, M. K. Shear, and C. Anderson, "Use of the child behavior checklist as a diagnostic screening tool in community mental health," *Res. Soc. Work Pract.*, vol. 15, no. 3, pp. 195–203, May 2005.
- [50] S. G. Aschenbrand, A. G. Angelosante, and P. C. Kendall, "Discriminant validity and clinical utility of the CBCL with anxiety-disordered youth," *J. Clin. Child Adolescent Psychol.*, vol. 34, no. 4, pp. 735–746, Dec. 2005.
- [51] J. F. Cohn and E. Z. Tronick, "Three-month-old infants' reaction to simulated maternal depression," *Child Develop.*, vol. 54, no. 1, pp. 185–193, 1983.
- [52] S. Nolen-Hoeksema, B. E. Wisco, and S. Lyubomirsky, "Rethinking rumination," *Perspectives Psychol. Sci.*, vol. 3, no. 5, pp. 400–424, Sep. 2008.
- [53] T. V. Barker *et al.*, "Contextual startle responses moderate the relation between behavioral inhibition and anxiety in middle childhood," *Psychophysiology*, vol. 52, no. 11, pp. 1544–1549, Nov. 2015.
- [54] M. N. Pavuluri, S. L. Luk, and R. McGee, "Help-seeking for behavior problems by parents of preschool children: A community study," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 35, no. 2, pp. 215–222, Feb. 1996.
- [55] N. D. Mian, "Little children with big worries: Addressing the needs of young, anxious children, and the problem of parent engagement," *Clin. Child Family Psychol. Rev.*, vol. 17, no. 1, pp. 85–96, Mar. 2014.
- [56] P. C. Kendall, S. Safford, E. Flannery-Schroeder, and A. Webb, "Child anxiety treatment: Outcomes in adolescence and impact on substance use and depression at 7.4-year follow-up," *J. Consult. Clin. Psychol.*, vol. 72, no. 2, pp. 276–287, Apr. 2004.