Automated assessment of mental health symptoms using audio and text features from speech

Daniel M. Low[1,2]; Elyse Shenberger[3,4]; Kate H. Bentley[5,6]; Randy P. Auerbach[7,8]; Stewart Shankman[3,4]; Satrajit S. Ghosh[1,6,9]

[1] Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, MA, USA; [2] Department of Brain and Cognitive Sciences, MIT, MA, USA; [3] Department of Psychology, University of Illinois at Chicago, IL, USA; [4] Department of Psychiatry and Brain Sciences, Northwestern University, IL, USA; [5] Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, MA, USA; [6] McGovern Institute for Brain Research, MIT, MA, USA; [7] Department of Psychiatry, Columbia University, NY, USA; [8] Division of Clinical Developmental Neuroscience, Sackler Institute, NY, USA; [9] Department of Otolaryngology, Harvard Medical School, MA, USA

INTRODUCTION

The majority of individuals with a mental health disorder do not receive clinical care due to cost, stigma, and other barriers. Machine learning technology using speech samples could one day be a biomarker used for remote, low-cost tracking of mental health problems. Herein we present the first systematic review of studies using speech for automated assessments across a broad range of psychiatric disorders [1]. We further show results on a dataset where we train machine learning models on audio and text features from participants' speech to predict a range of symptoms (e.g., depression, anxiety) and personality characteristics (e.g., distractibility, perfectionism).

METHODS

Review. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines and included studies from the past 10 years using speech to identify the presence or severity of mental health disorders within the Diagnostic and Statistical Manual of Mental Disorders (DSM–5). For each study, we describe sample size, clinical evaluation method, speech-eliciting tasks, machine learning methodology, performance, and other relevant findings.

Classification tasks. We evaluated data from a recently completed study comparing healthy (n=21) and remitted depressed (n=24) adults who provided free-form speech over multiple days from a smartphone app. Participants also were administered the Structured Clinical Interview for DSM-IV, Inventory of Depression and Anxiety Symptoms (IDAS), and the Personality Inventory for DSM-5 (PID-5). Scores from the questionnaire subscales were binarized to obtain a uniform distribution between high and low scores removing constructs with less than 20% of participants in the minority class, resulting in 34 constructs to be predicted. Answers across days were grouped for our main analysis resulting in 153 to 167 samples. Models included support vector machines, elastic nets, gradient boosted decision trees, and recurrent neural networks. Acoustic features were obtained from OpenSmile INTERSPEECH 2013 ComParE feature set, while text features included universal sentence encoder, the LIWC features, and TFIDF unigrams and bigrams. Models were trained with and without applying UMAP dimensionality reduction. We measured performance using a 30-sample bootstrapping procedure and tested significance through a permutation test with Bonferroni correction. Hyperparameter tuning was performed with nested bootstrapping. Finally, explainability was analyzed through feature importance with SHAP (SHapley Additive exPlanations), a method that fulfills several theoretically desirable properties that other additive feature attribution methods do not. Since each participant provided speech

samples on multiple days, we tested the consistency of SHAP feature importance across multiple days of recording the same participants and random bootstrapping test sets.

RESULTS

Review. All studies (n=1,395) were screened, and 127 studies met the inclusion criteria. Approximately 85% of studies focused on depression, schizophrenia, and bipolar disorder, and the remaining studies included individuals with post-traumatic stress disorder, anxiety disorders, and eating disorders. We provided an online database with our search results and synthesize how particular acoustic features are abnormal in each disorder (e.g., fundamental frequency is significantly lower in depression than healthy controls). We included over a dozen guidelines for acquiring data and building models with a focus on testing hypotheses, open science, reproducibility, and generalizability.

Classification tasks. We achieved significant results for binary classification (higher versus lower) on the IDAS subscales measuring depressive, PTSD, and social anxiety symptom severity and several personality traits from the PID-5 (distractibility, separation insecurity, anhedonia, grandiosity) from audio or text with mean ROC AUC scores for all constructs ranging between 0.58 and 0.68. We detected if individuals belonged to the remitted depression group or the healthy control group using both audio and text models with a mean ROC AUC score of 0.66 and 0.68, respectively. UMAP dimensionality reduction improved some results. Regarding algorithmic explainability, our results show that SHAP feature importance values vary considerably, which means models use different features depending on the day or random sample on which they were trained.

DISCUSSION

We found multiple significant links between different psychological constructs and audio and text data. We were able to demonstrate that individuals with a history of depression can be detected from their audible or transcribed speech even when they no longer meet the criteria for major depressive disorder, which could be useful for identifying individuals at-risk for future relapse. Our systematic review shows that the field is growing but has not yet determined with precision which acoustic features modulate systematically in distinct psychiatric disorders. When analyzing the generalizability of feature importance in our dataset we confirmed the inconsistency found in the literature. This inconsistency may be due to the episodic nature of some mental health symptoms (i.e., a label may be accurate at one time point but not another), using a small dataset, and to the possible limitations of methods such as SHAP. This finding should emphasize the importance of exercising caution when applying feature importance methods without testing generalization and shows how a more robust feature importance can be reconstructed by measuring which features are important across days or random samples. Advancing the robustness of feature importance is key for both algorithmic transparency and to ultimately achieve generative models of how mental health problems affect speech, which in turn would improve its likelihood of becoming a clinically useful biomarker.

References

1. Low DM, Bentley KH, Ghosh SS. Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review. 2019. doi:10.31219/osf.io/5pwze