MEASURING DEPRESSION USING AN AUTOMATED NATURAL SPEECH ANALYSIS PIPELINE: A POTENTIAL CLINICAL TRIAL ENRICHMENT TOOL

ALEX S COHEN
Louisiana State University
Dept. of Psychology
Center for Computation & Technology
Quantic Innovations

Brian Kirkpatrick
Psychiatric Research Institute
University of Arkansas for Medical Sciences
Quantic Innovations

Mark Opler
WCG Clinical
Quantic Innovations

Snezana Milanovic, Steve T. Szabo, Kenneth S. Koblan, Seth Hopkins
Sunovion Pharmaceuticals Inc

INTRODUCTION

Speech analysis offers an objective approach to quantifying depression-spectrum phenotypes.

"Proof of concept" has been demonstrated, but few studies show replicability/generalization across studies and samples.

The present study applied:
1) speech analysis for potential clinical trial enrichment and
2) used a robust clinical trial dataset as a foundation for replicability/generalizability, specifically by:
   - Focusing on a small number of "face valid" speech features
   - Speech from a standardized clinical interview
   - Developing a clinically-interpretable algorithmic approach.

METHODS OVERVIEW

Audio recordings from Montgomery-Åsberg Depression Rating Scale (MADRS) interviews for a completed Phase 2 clinical trial were evaluated. Participants had a diagnosis of bipolar I disorder and currently in a major depressive episode.

Approximately 1661 recordings were available for 477 participants over baseline, screening, and 3 post-randomization sessions from 7 recruitment countries (15% of data was unusable due to technical issues). All sessions were examined in this study

Depression status of MADRS scores was examined in a binary manner. Analysis employed Decision Tree Learning.

"Face Valid" Speech Features Set:
- Average Latency to Response: Average time for the patient to begin responding to interviewer
- Average Pause Times: Average time between phrases within a patient response.
- Speaking Rate: Average words per second.
- Intonation: Variability in "pitch"
- Emphasis: Variability in "emphasis"

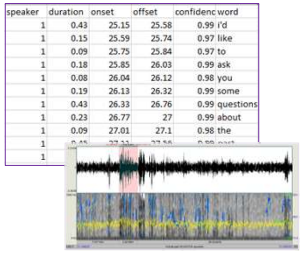
SPEECH ANALYSIS PIPELINE

MADRS Interview

Speaker Detection & Transcription (with Timestamping)

Acoustic Analysis

Predicting Depression (Using Decision Tree Learning)

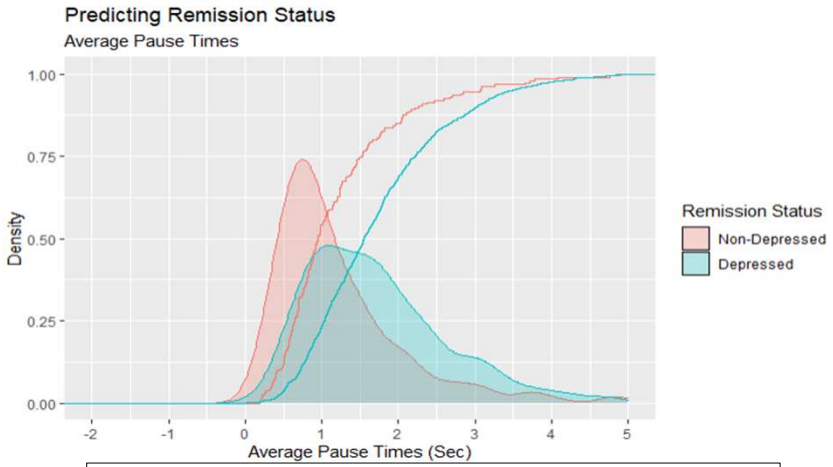


speaker	duration	onset	offset	confidenc	word
1	0.43	25.15	25.58	0.99	i'd
1	0.15	25.59	25.74	0.97	like
1	0.09	25.75	25.84	0.97	to
1	0.18	25.85	26.03	0.99	ask
1	0.08	26.04	26.12	0.98	you
1	0.19	26.13	26.32	0.99	some
1	0.43	26.33	26.76	0.99	questions
1	0.23	26.77	27	0.99	about
1	0.09	27.01	27.1	0.98	the

The Model Predicted Depression, But Required Tuning Across Countries

The model showed high accuracy (0.88) and specificity (0.99), and modest sensitivity (0.22), using data from all countries, and all five features.

Average Pause Time Was Critical



Density plot showing average pause times for all participants from all countries as a function of Depressed and Non-Depressed status.

RESULTS

Are automated speaker detection & transcription accurate? Yes, accuracy with trained human transcribers exceeded 90%.

Did speech measures differ across countries? *Yes, there was some variability across countries.*

Which speech feature was most important? *Average Pause Time was critical for the overall model, was the only feature needed for three countries, and was a predictor in all countries but one.*

How does an average Pause Time of 1 second alone relate to depression status?
High accuracy (acc = 0.74, AUC = 0.59) reflecting very high specificity (0.91) and modest sensitivity (0.27).

CONCLUSIONS

Generalizability: The utility of the feature set generalized across countries. Average Pause Time was the most generalizable feature.

Average pause time was a particularly important feature for predicting depression status.

These findings support the use of an automated speech-analytic pipeline for complementing clinical evaluation and enrichment strategies in trial design.

Modelling depression for some countries was challenged due to imbalanced and limited numbers of nondepressed cases.

Contact: Alex Cohen: acohen@lsu.edu