Gathering normative speech data for depression research remotely, using online task marketplaces.

Celia Fidalgo, PhD[1], Mengdan Xu, MSc[1], Aparna Balagopalan, MSc[1], Jessica Robin, PhD[1], Liam D. Kaufman, MSc[1], William Simpson, PhD[1,2]

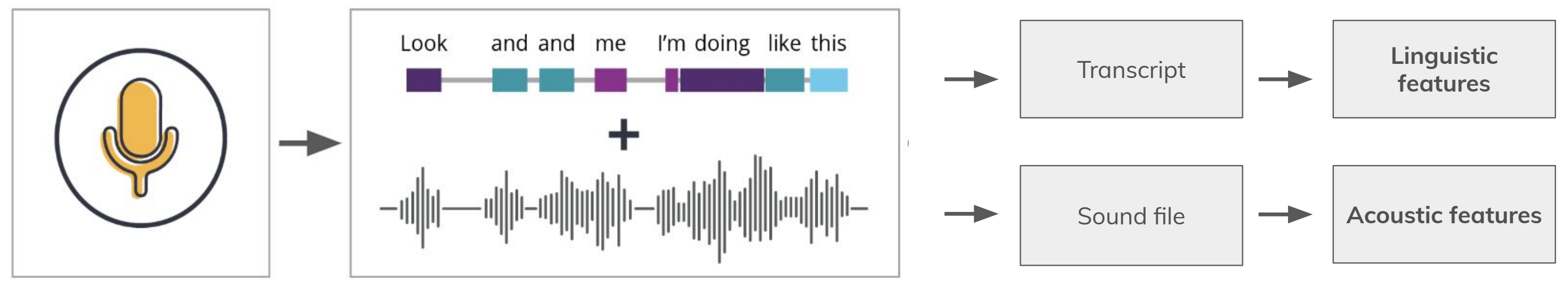
(1) Winterlight Labs, Toronto, ON, Canada
(2) Department of Psychiatry and Behavioural Neuroscience, McMaster University, Hamilton, ON, Canada

Background

Major Depression is currently monitored via subjective symptoms including low mood, changes in sleep and difficulty concentrating, which present heterogeneously and may reflect multiple underlying etiologies. Non-invasive technologies, including the assessment of speech, could help delineate clinically meaningful subtypes and assist with therapeutic development (1). To understand the relationship between speech and depressive symptoms, large representative samples of speech are needed. The objective of this study was to evaluate whether speech data collected through a distributed online platform was of sufficient quality for use as normative data in depression research.

Methods

We used Amazon's Mechanical Turk (mTurk) to collect speech assessments from a normative sample. Participants provided basic demographic information, completed a Patient Health Questionnaire-9 (PHQ-9) and provided speech samples. Speech tasks included sustained vowel phonation, phonemic fluency, picture description, positive fluency (listing positive events that will occur), and prompted narrative. To ensure participants were paying attention to the task, we asked a validation question ("Have you ever had a fatal heart attack while watching TV?"). Sampling was restricted to English speakers, aged 20-60, in the United States. Speech samples were analyzed using signal and natural language processing tools to extract a range of speech features (see schematic below).



We explored the relationship between the features and self-reported ratings of depressive symptoms using non-parametric, partial correlations, controlling for age, sex and level of education.

Figure 1: Global and Age based distributions of PHQ-9 scores

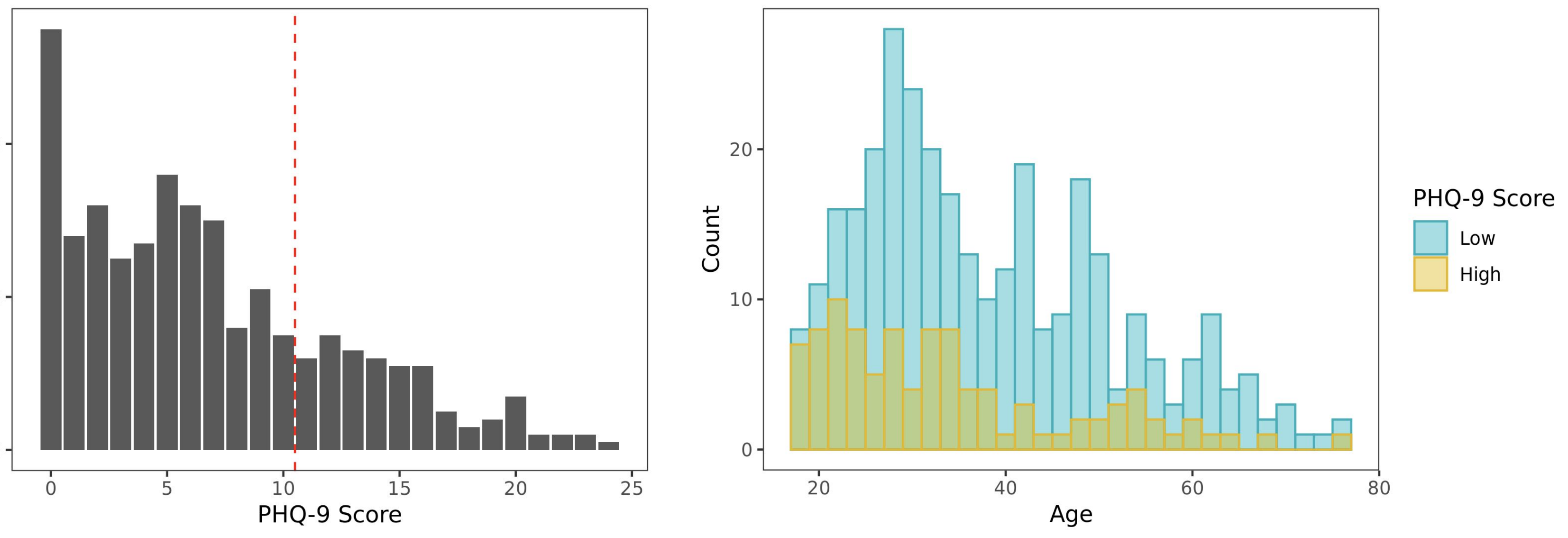


Figure 2: Distribution of PHQ-9 question response times

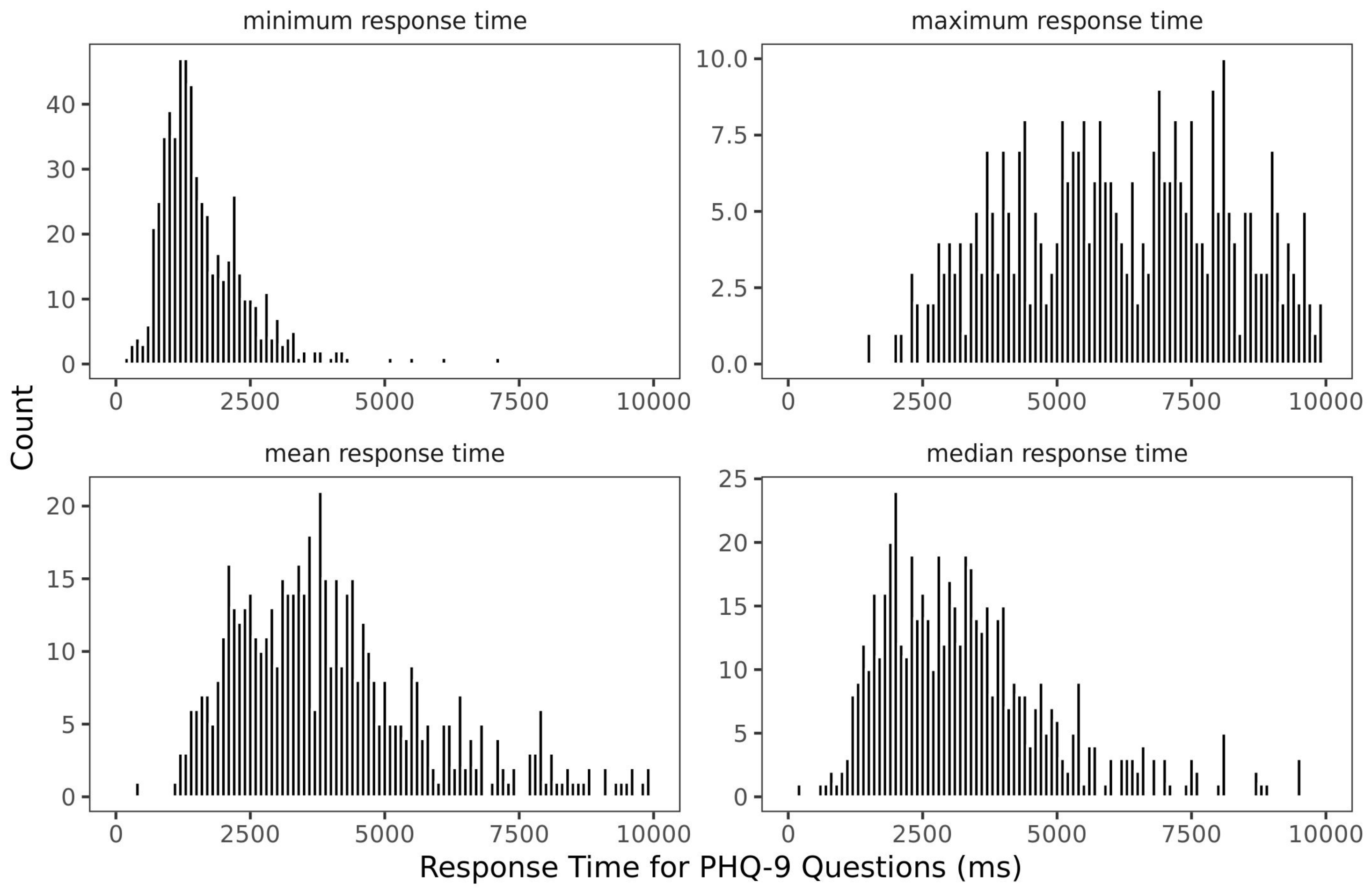
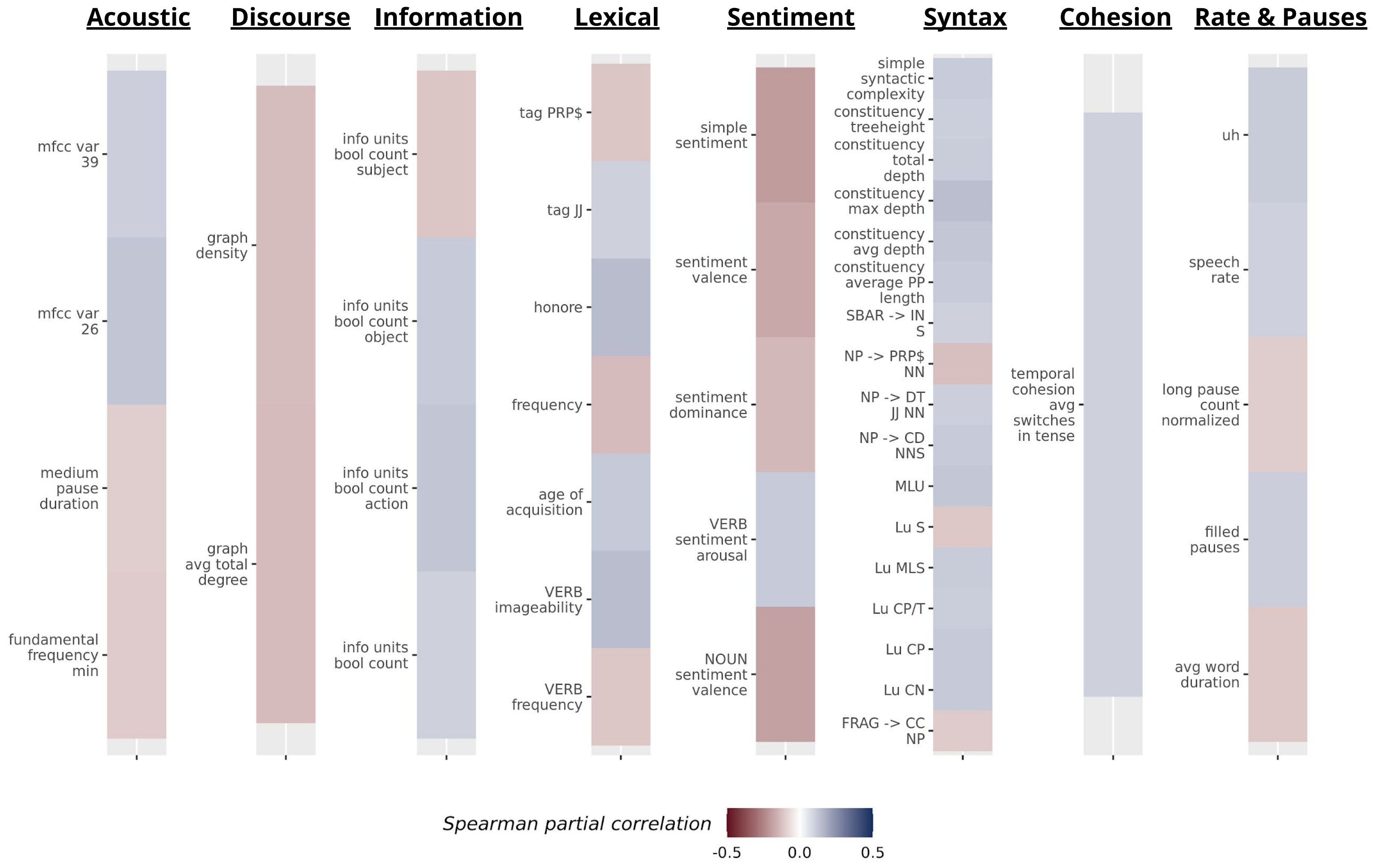


Figure 3: Correlations between speech features and PHQ-9 scores



Speech features were extracted from the picture description speech task. Exploratory Spearman partial correlations (controlling for age, sex and education) were computed between speech features and PHQ-9 scores. Correlations significant at $p < 0.05$ (uncorrected) are displayed.

Results

A total of 2779 samples were collected from 627 individuals. 136 (4.9%) of the samples had no recorded speech. The remaining samples were of varying quality, due to differences in hardware and recording conditions. Based on manual data checks, we identified 30 (1.1%) low audio quality samples, 16 (0.5%) samples where task instructions were not followed, and 7 (<0.1%) samples with overwhelming background noise. Seventeen participants (2.7%) were excluded for failing the validation question. The dataset was collected in 96 hours at a cost of $851.88 USD (<$2USD per participant).

Based on the PHQ-9 cutoff score of 10, 23.9% of the sample met screening criteria for depression (Figure 1). Examination of mean response latencies for PHQ-9 questions revealed a bimodal distribution, with a group of individuals with mean response times < 2500 ms (Figure 2). We hypothesized that very short response latencies were indicative of inattention, and applied a latency cutoff of 2500 ms for mean response times. After these data cleaning steps, the sample size decreased to 1964 samples from 417 participants (66.5% of the original sample).

Exploratory analysis of the relationship between speech features extracted from the picture description task and depressive symptoms (Figure 3) revealed modest correlations between increasing depressive symptoms and use of more negatively valenced words. Similarly modest relationships were seen for fundamental frequency and word duration.

Conclusions

This study demonstrates that remote collection of normative speech data is feasible using online task marketplaces. The rate of depression in participants was higher than expected, possibly indicating selection bias or increased depression relating to the COVID-19 pandemic. Data cleaning procedures identified a number of quality issues, including missing or low quality audio, failed validation questions, and very short response times. Despite these shortcomings, the cost and speed by which data can be collected makes well-designed remote studies of this nature a viable option for collection of normative data, provided that systematic quality checks of the data are implemented. Exploratory analysis of the relationship between speech features and depressive symptoms showed similar results as previously published reports from in clinic and remotely collected samples (2,3).

References

(1) Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology 5, 96–116 (2020).

(2) Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K. & Geralts, D. S. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of Neurolinguistics 20, 50–64 (2007).

(3) Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. Biological Psychiatry 72, 580–587 (2012).

WINTERLIGHT

Presented at ISCTM 2020 Autumn Conference, September 21-25, 2020

McMaster University