Use of a voice-based digital biomarker in patients with depression

William Simpson [1,2], Aparna Balagopalan [1,3], Liam D Kaufman [1], Anthony Yeung [3], Adam Butler [4]

(1) Winterlight Labs, Toronto, ON, Canada, (2) McMaster University, Hamilton, ON, Canada,
(3) University of Toronto, Toronto, ON, Canada, (4) Radix Consulting

Methodological Question

Could acoustic and linguistic analysis of speech be used to build a voice-based digital biomarker to detect depression?

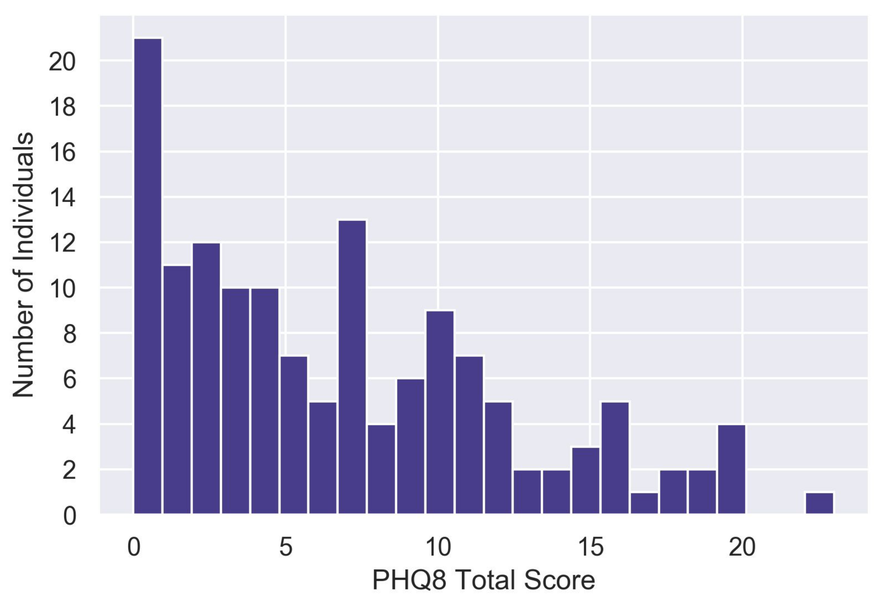
Introduction

- Screening for depression has primarily relied on the use of subjective clinician-rated and patient-reported outcome measures.
- Previous research has identified several components of speech and language that are characteristic of depression; including reduced amplitude, greater use of emotionally negative words and higher use of personal pronouns (Taguchi et al., 2018).
- Speech is not only a potentially rich data source, but it is simple to collect, and could enable effective high frequency, remote assessment. A digital biomarker could improve participant selection and outcomes of clinical trials in depression and other mood disorders.
- We had previously developed and reported on an automated speech analysis platform which examined both the acoustics (i.e. properties of the sound wave) and content (i.e. detailed linguistic analysis) of speech with a focus on building digital biomarkers for Alzheimer's disease.
- The objective of this proof-of-concept study was to examine the comparative advantages of adding linguistic features to voice-based digital biomarkers for identifying individuals with depression.

Methodology

- Analysis was conducted on the DAIC-WOZ dataset (USC Institute for Creative Technologies, http://dcapswoz.ict.usc.edu/). This dataset (n=189 individuals) is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014) that contains transcripts of clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder using a computer agent (DeVault et al., 2014). A variety of other measures including PHQ-8 scores were collected.
- Audio samples and transcripts were processed using speech analysis software developed by Winterlight Labs. This software produces a set of ~500 linguistic and acoustic variables per sample.
- DAIC-WOZ is separated into standard training (n=107) and dev (n=35) sets. We trained and tested a variety of machine learning classifiers using these standard datasets (with feature selection) to predict the binary label of depressed / not depressed contained within the dataset. This label was based on a threshold PHQ-8 score of 10.
- Training dataset: 30 depressed, 77 non-depressed, dev dataset: 9 depressed, 26 non-depressed.
- Classifiers included: Neural Networks (NN), Support Vector Machines (SVM), and Random Forests (RF).
- Accuracy, sensitivity and specificity were computed for each model. To determine whether the results were specific to the 'training' and 'dev' sets, we also trained models using 10-fold cross validation on the full dataset and compared the results.
- The distribution of PHQ-8 scores is depicted in Figure 1, indicating a skew towards subclinical and mild symptoms with ~ 3:1 not depressed to depressed ratio. It is expected that models trained on this dataset would have lower sensitivities due to fewer symptomatic participants.

Figure 1: Distribution of PHQ-8 scores in the DAIC-WOZ dataset



Results

- Previous studies, using individual participant responses to questions, have achieved an accuracy threshold of 77%
- Using the standard 'training' and 'dev' sets with all individual participant responses aggregated into a single transcript, we replicated this threshold using a linguistic only model, (Table 1).
- Adding in the available acoustic features further increased the accuracy to 79.3%, at the expense of the balance between sensitivity and specificity (Table 1).
- Using the full dataset and 10-fold cross validation produced comparable levels of accuracy for models based on linguistic features (Table 2), though the balance between sensitivity and specificity was poorer.
- Models which included linguistic features had higher accuracies and a better balance of sensitivity and specificity compared to acoustic only models.

Table 1: Accuracy, Sensitivity and Specificity of best performing machine learning classifiers for different feature groups, using the standard 'training' and 'dev' datasets

Model	Feature Set	Accuracy	Sensitivity	Specificity
NN/RF	Acoustic	51.7%	36.4%	61.1%
NN	Linguistic	75.8 %	72.7%	77.8%
NN	Acoustic + Linguistic	79.3%	63.6%	88.9%

Table 2: Accuracy, Sensitivity and Specificity of best performing machine learning classifiers for different feature groups, using 10-fold cross validation

Model	Feature Set	Accuracy	Sensitivity	Specificity
NN	Acoustic	65.9%	42.0%	74.1%
NN	Linguistic	76.3 %	60.2%	89.1%
RF	Acoustic + Linguistic	63.5%	53.3%	61.2%

Conclusions & Next Steps

- A model using linguistic variables derived from DAIC-WOZ transcripts achieved the threshold classification accuracy seen in previous work, further replicating that there is a detectable speech signal in patients with depression.
- Models based on linguistic features consistently outperformed acoustic only models, highlighting the potential diagnostic importance of the content (and not just the sound) of speech.
- Linguistic analysis could provide more clinically meaningful and interpretable results in a conversational sample of speech, though further exploration in other datasets is required.
- Discrepancies between models trained on standard datasets and those which were cross validated indicate that subsequent studies with larger data sets are required to assess the generalizability of these results.
- The use of machine learning and artificial intelligence to assist in clinical decision-making represents an important opportunity to improve the conduct of clinical trials in mood disorders.
- Subsequent experiments examining the dataset at response level (thereby dramatically increasing training set size) and expanding the dataset to include those with higher PHQ-9 scores are currently in progress.

References

Gratch J, et. al. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128)

DeVault, D. et al. (2014). "SimSensei kiosk: A virtual human interviewer for healthcare decision support". In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14), Paris

Taguchi T, et al. Major depressive disorder discrimination using vocal acoustic features. JAD. 2018;225:214-220.

WINTERLIGHT

Presented at ISCTM 2019, Copenhagen, Denmark. September 5-7, 2019.

McMaster University