Correlates and Prediction of Generalized Anxiety Disorder from Acoustic and Linguistic Features of Impromptu Speech

by

Bazen Gashaw Teferra

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright 2022 by Bazen Gashaw Teferra

Correlates and Prediction of Generalized Anxiety Disorder from Acoustic and Linguistic Features of Impromptu Speech

Bazen Gashaw Teferra
Doctor of Philosophy

Graduate Department of Electrical and Computer Engineering
University of Toronto
2022

Abstract

The diagnosis and tracking of Generalized Anxiety Disorder (GAD) require frequent interaction with mental health professionals, but there are not enough professionals to serve the demand. The ability to automatically detect anxiety disorders through samples of speech could be a valuable complement to conventional treatment and provide greater access to it.

In this work, we explore acoustic and linguistic features of speech that correlate with anxiety and use those to predict above or below the screening threshold of GAD. A large number of participants ($N = 2,000$) participated in a single online study session where they completed the Generalized Anxiety Disorder-7 (GAD-7) assessment and provided an impromptu speech sample in response to a modified version of the Trier Social Stress Test (TSST). Acoustic and linguistic speech features were a-priori selected based on the existing speech and anxiety literature, together with related features. Associations between speech features and anxiety levels were assessed using sex, age, and personal income included as covariates. The amount of speech had the most significant correlation with GAD-7 ($r = -0.12$; $P < .001$), indicating that participants with higher anxiety scores spoke less. Linguistic features acquired using Linguistic Inquiry and Word Count (LIWC) were also significantly ($P < .05$) associated with anxiety.

Using these acoustic and LIWC features to predict above or below a screening threshold of 10 for GAD, a logistic regression model achieved a mean $AUROC = 0.57$, $SD = 0.03$. The mean AUROC increased to 0.62 ($SD = 0.03$) when demographic

information (age, sex, and income) was included indicating the importance of demographics when screening for anxiety disorders.

The LIWC linguistic features are based on single-word counts and do not account for the surrounding word context. We attempted the same prediction based on greater context using a transformer-based neural-network model (pre-trained on large textual corpora) and fine-tuned on the speech transcripts. This model, which only uses the textual input, achieved an AUROC value of 0.64. When acoustic, LIWC, and the predicted output of the fine-tuned transformer-based model are combined into one model, the AUROC increased to 0.67 and further increased to 0.68 when demographic information was included. The results suggest that there is a signal of anxiety within such impromptu speech that may be useful as part of a system to screen for anxiety, detect relapse, or monitor treatment.

To my family

Acknowledgments

First and foremost, I want to thank GOD for all the blessings in my life.

I would like to express my deepest gratitude to my supervisor, Professor Jonathan Rose, for his guidance, patience, and support throughout my thesis. His thoughtful comments, constructive criticisms, and unwavering encouragement have been instrumental in helping me reach this milestone. I would also like to thank my wife, friends, and family for their support and understanding during this time. Without their love and patience, this would not have been possible. Furthermore, I would like to thank all the members of my research group for their invaluable help and assistance during the research and writing processes. Their insight and feedback helped me to stay focused and motivated. Finally, I would like to thank Dr. Daniel Di Matteo, Professor Ludovic Rheault, Professor Sophie Borwein, Dr. Danielle D DeSouza, and Dr. Bill Simpson for their energy and assistance with the launch of this project and the data collection and advice.

This work would not have been possible without the support of these individuals. Thank you all!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Poor mental health causes suffering and affects a person's ability to function in society. It affects all Canadians at some time in their lives, either themselves, in a family member, a friend, or a colleague [1]. Early detection and treatment of mental health disorders is therefore crucial in order to improve the quality of life for those affected and the people around them. However, it is widely acknowledged that there is insufficient effort and resources allocated to the diagnosis and treatment of mental health disorders [2].

One issue in the diagnosis and treatment of mental health illness is the availability of psychiatrists [3]. The insufficient number of psychiatrists may, in part, be due to the high cost of training practitioners [4]. It might be required to introduce some type of automation that could help with the cost reduction. The long-term vision that drives this specific work is to look for ways that computer-based automation can assist conventional treatment by automating some parts of the screening, diagnosis, and treatment.

In addition to cost reduction, computer-based automation provides the ability to monitor passively, remotely, and frequently. Furthermore, the data collected can be considered to be objective, in that it is not interpreted by a human. In contrast, an interview with a psychiatrist or a psychologist is influenced by the subjective opinion of the clinician. This has been shown to cause a varying diagnosis outcome in patients [5].

We anticipate the following scenario for a system that includes the capability to automatically predict a mental health disorder from speech: such a system would sample a sequence of the subject's speech throughout the day and produce multiple predictions. Depending on the accuracy of the individual predictions, the system could use multiple predictions to increase the overall accuracy of the final screening result. Note that such an approach would use passively collected speech which gives rise to its own challenges - including that there would need to be a process of speaker identification to select language spoken by the subject and also ensure the privacy and confidentiality of the subject is protected.

1.1 Focus & Goals

In this work, we focus on anxiety disorders and, more specifically, on Generalized Anxiety Disorders (GAD). Anxiety disorders are one group of mental illnesses and are among the most common types of mental health disorders. GAD is characterized by excessive feelings of nervousness, anxiety, and even fear. According to a 2013 Canadian Community Health Survey, 3 million Canadians (11.6% of the population) aged 18 years or older reported that they had a mood and/or anxiety disorder [6]. This high prevalence among mental illnesses has led us to focus on anxiety disorders.

To help people suffering from anxiety disorders, our broader goal is to automate the detection of anxiety disorders. If this capability existed, it might have an application in the following areas:

- It may be used for early detection of anxiety, perhaps used by the majority of the population in a monitoring mode.
- It could be used to track a patient's level of anxiety during conventional treatment in a far more granular way than is currently possible.
- After a patient achieves remission, it could be used to detect relapse.
- Finally, such frequent monitoring, together with other time-series information, could be used to help understand the mechanisms of the GAD itself.

A related project in our research sought more broadly to automate mental health illness detection. This work, titled Project Arrow, used smartphones to passively collect patients' data (such as text, GPS location, audio etc...) over a certain duration of time to infer their mental health state. Project Arrow had some success in finding a relationship between passively collected data with mental health illness [7] [8] [9] [10] [11]. In that project, the audio sample stream yielded among the most useful correlates of depression and social anxiety, but did not succeed in finding significant correlates with generalized anxiety. The current study focuses only on audio data and explores the feasibility of detecting and measuring generalized anxiety disorder from speech.

1.2 Contributions

The present work makes the following contributions:

- **Collection of larger-scale speech data labelled with GAD-7**

  We have collected one of the largest-ever sample sizes ($N = 2000$) of impromptu speech labelled with the GAD-7 scale. Although our ethics protocol prevents us from publishing the raw data, it does permit us to publish, for use by others, the computed features of the data for use by others, in addition to being the full basis for the present work.

- **Validation of the association of previously suggested acoustic and linguistic features of speech with anxiety severity**

  Here, we explored the acoustic and linguistic features that were suggested to be associated with anxiety disorders in previous studies and see if they still hold up with a much larger dataset. Several of both the acoustic and linguistic features were significantly ($P < .05$) correlated with anxiety, but there were also some that were not significantly associated with anxiety.

- **Created a model to predict anxiety from acoustic and linguistic features of speech**

  The overall goal of this work is to detect generalized anxiety disorder from speech. The inputs are the selected speech features (from the previous contribution), and the output is either the presence or absence of anxiety in that speech sample. A logistic regression model using the significant acoustic and linguistic features achieved a mean AUROC of 0.57 and further increased to 0.62 when demographic information (age, sex, and income) was included.

- **Exploring Context-Aware multi-word patterns of speech in the prediction of anxiety**

  Most prior work on linguistic features of speech in mental health focused on using single words as features, even though the full context of individual words within sentences can easily change the sense and meaning of those words. We have trained and tested a model that does consider the multi-word patterns and shows that the context does indeed matter. It makes use of very recent large language models that are pre-trained and then fine-tuned on our data.

1.3 Organization

The remainder of the thesis is organized as follows: Chapter 2 presents background information on mental health disorders, speech processing, and machine learning models, and provides a review of prior related work on prediction. Chapter 3 describes the methods used to collect speech samples from participants as well as the anxiety severity label associated with each participant and then provides a summary of the collected data and the demographics of the recruited participants. Chapter 4 explores the acoustic and linguistic features of speech suggested by prior work to have an association with anxiety and validates if the association still exists with our much larger dataset. Chapter 5 presents the performance of a model that predicts the presence of GAD from the acoustic and linguistic features of speech validated in Chapter 4. Chapter 6 focuses on the prediction of anxiety from the transcript only, using a context-aware model based on a Transformer-based neural network. In Chapter 7, we merge the prediction models presented in Chapters 5 & 6 into one single model and discuss the performance gain achieved. Finally, Chapter 8 concludes this thesis by summarizing our work and discussing its implication for future research on anxiety detection.

Chapter 2

Background and Related Work

This chapter begins with background information on the condition we are focusing on, Generalized Anxiety Disorder (GAD). We then cover related technical backgrounds in speech processing and machine learning. Finally, there is a review of prior work related to the detection of anxiety disorder.

2.1 Generalized Anxiety Disorder

Anxiety is a very normal and common emotion that is a person's natural feeling of fear about what is to come. Its evolutionary purpose is to keep us away from danger and to make us prepare for future events [12]. While this is very important, it can also be harmful when it is in excess and limits us from doing our daily tasks. The mental health issue where there is excessive anxiety is referred to as an *anxiety disorder*.

Anxiety disorders are among the most common type of mental health illnesses and can be further classified into different types. Some of these anxiety disorders include: Social Anxiety Disorder (SAD) - fear of being negatively judged by others; Panic disorder - fear of losing control; Phobia - fear of something bad is about to happen over a specific object or situation. One type where this work will focus is Generalized Anxiety Disorder (GAD). GAD is characterized by an intense and persistent feeling of anxiety and fear of what is to come, limiting the performance of a person's day-to-day activities [13]. Out of all the patients in primary care who have anxiety problems, 22% suffer from GAD [14]. GAD is a risk factor for several social problems, such as living alone, low level of education and unemployment [15]. However, it is still rarely diagnosed and treated [16] - a reminder of the need to improve the ability to identify, diagnose and treat patients with GAD, the long-term goal of this work.

2.1.1 Current Diagnosis Method for GAD

The current diagnosis method for GAD requires a one-on-one interview with a psychiatrist. The psychiatrist will study and analyze the patient's behaviour to identify the possible symptoms of GAD. Psychiatrists follow a specific set of guidelines around symptoms or actions (referred to as *diagnostic criteria*) to diagnose GAD. One commonly used set of diagnostic criteria is the Diagnostic and Statistical Manual of Mental Disorders (DSM). The following are listed as the diagnostic criteria for GAD according to the Diagnostic and Statistical Manual of Mental Disorders, version 5 (DSM-V)[17]:

- Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, about a number of events or activities (such as work or school performance).
- The person finds it difficult to control the worry.
- The anxiety and worry are associated with three or more of the following six symptoms (with at least some symptoms present for more days than not for the past six months).
  - Restlessness or feeling keyed up or on edge
  - Being easily fatigued
  - Difficulty concentrating or mind going blank
  - Irritability
  - Muscle tension
  - Sleep disturbance (difficulty falling or staying asleep, or restless unsatisfying sleep)
- The anxiety, worry, or physical symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
- The disturbance is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition (e.g., hyperthyroidism).
- The disturbance is not better explained by another mental disorder (e.g., anxiety or worry about having panic attacks in panic disorder, negative evaluation in social anxiety disorder [social phobia], contamination or other obsessions in obsessive-compulsive disorder, separation from attachment figures in separation anxiety disorder, reminders of traumatic events in posttraumatic stress disorder,

gaining weight in anorexia nervosa, physical complaints in somatic symptom disorder, perceived appearance flaws in body dysmorphic disorder, having a serious illness in illness anxiety disorder, or the content of delusional beliefs in schizophrenia or delusional disorder).

2.1.2 Generalized Anxiety Disorder 7 item Questionnaire (GAD-7)

The Generalized Anxiety Disorder seven-item questionnaire – also known as GAD-7 in short – is a fast questionnaire (that usually takes less than two minutes to fill) to measure the severity of GAD. It is used in primary care as a screening tool. Although a clinician or a psychiatric diagnosis is considered as the gold standard for diagnosing GAD, the GAD-7 is fairly accurate with sensitivity of 89% and specificity of 82% [18].

There are seven questions on the GAD-7 questionnaire (see Appendix A) where each of the seven items has four options for how often a participant has been bothered by a particular problem. These options with their numerical representations are: 0-not at all, 1-Several days, 2-more than half the days, and 3-nearly every day. The final GAD-7 score is a summation of all these values giving a severity measure for GAD in the range of 0-21. The GAD-7 score threshold for diagnosing GAD with optimum sensitivity and specificity is 10 [18]. Furthermore, scores of 5, 10 and 15 can be used as the cut-off threshold for mild, moderate and severe anxiety, respectively.

2.2 Speech Processing

Speech processing is the task of analyzing the audio signal produced from human speech aiming to study and extract useful information for further study. In the subsection below, two classes of speech processing - audio and text - that are relevant to this work are discussed.

2.2.1 Speech Audio Processing

In speech audio processing, the task is to extract and produce features from the waveform of the audio. It can be considered as compressing the raw audio signal into a small amount of information relevant to the overall goal, and discarding the rest. In the foregoing, we will refer to these features as acoustic features. Below, we will describe a number of acoustic features that have been explored in previous related studies and have been suggested to have an association with anxiety. The methodology of validating the correlation of these set of acoustic features with anxiety using our dataset will be described in Chapter 4.

Mel Frequency Cepstral Coefficients (MFCC)

The Mel Frequency Cepstral Coefficients, or MFCC for short, are one of the most widely-used features extracted from speech for different tasks such as Speech-To-Text (STT) [19], Speaker Recognition [20], and also the prediction of different mental health illnesses [21].

There are multiple steps of processing to obtain the final MFCC values. Given a certain window of speech audio, the first step is to transform the audio signal from a time-domain representation to a frequency-domain representation using a Discrete Fourier Transform (DFT), producing a frequency spectrum. Then, the spectrum will be filtered using mel-scale, which mimics the response of the human ear. Then the log of the mel-scale signal is taken and then another DFT is applied. This last step produces a 1-D array representing the spectrum of the log of the spectrum and is known as the *Cepstrum*. The MFCCs are the amplitude of the first 13 values (usually, but can also use more depending on the task) of the Cepstrum.

Linear Prediction Cepstral Coefficient (LPCC)

While the MFCC come from the mel-scale filtering, LPCC are coefficient derived from the linear prediction cepstral representation of an audio signal. All the steps of the extraction of LPCC are similar to MFCC (described above) except for the filtering step. During the MFCC extraction, the frequency spectrum are filtered using mel-scale. However, during the extraction of LPCC, the frequency spectrum are filtered on a linear scale.

ZCR zPSD

ZCR zPSD is an abbreviated form for the Zero Crossing Rate (ZCR) of the z-score of the Power Spectral Density (PSD). The PSD is a measure of a signal's power content versus frequency. The z-score of the PSD (zPSD) measures how different the values are from the mean in terms of the standard deviation - i.e., if the z-score is 0 then the data point's score is identical to the mean and a z-score of 1 indicates a value that is one standard deviation from the mean. It could also be both positive and negative. The zero crossing rate (ZCR) is the rate at which the zPSD changes from positive to negative and vice versa.

Amount of speech

The amount of time a person is speaking and related metrics such as the percentage of silence. We will use two features to measure the amount of speech: the amount of time, in seconds, that speech was present, and the total number of words present in a speech-to-text transcript.

Articulation rate

The articulation rate indicates how fast a person is speaking. We use a software library called My Voice Analysis [22] to extract this feature from a speech audio.

Fundamental frequency

The fundamental frequency (F0) of speech is the rate at which the glottis vibrates, and is also known as the voice *pitch*. This glottis vibration creates a periodic wave for which a particular signal $s(t)$ in a single period t would be equal to the signal at $s(t+T)$ where T is the period of the wave. The fundamental frequency is the inverse of the smallest possible period T.

A review conducted in 2006 [23] indicates that there are more than 70 ways of extracting F0, which shows the task of extracting F0 is not an easy one. For this study, we will be using a popular and widely cited tool known as Praat [24]. By default, Praat calculates the F0 value for every 10 ms window producing 100 F0 values per second. This means that the F0 value will vary (as indeed pitch does vary during speech) over time. Therefore, descriptive statistics such as the mean F0 and the standard deviation of F0 can be used for further analysis of the fundamental frequency.

F1, F2, F3

These are the first, second and third formants [25] which are the spectral maximum from an acoustic resonance of the vocal tract [26]. They are frequency peaks in a spectrum which have a high degree of energy. Typically, the first formant is around 500Hz, the second is around 1500Hz and the third is around 2500Hz.

Jitter

In signal processing, jitter is the cycle-to-cycle F0 variation in the sound wave. i.e., given N F0 computations from certain length audio (for example, 100 F0 sampled per second using the default Praat settings), jitter would be the average absolute difference between successive F0 samples. Praat will also be used for the estimation of jitter.

$$Jitter = \frac{1}{N} \sum_{i=1}^{N-1} |F0_i - F0_{i+1}| \tag{2.1}$$

Shimmer

Shimmer is the cycle-to-cycle amplitude variation of the sound wave. In equation B.2 below A_i and A_{i+1} are amplitudes of consecutive periods.

$$Shimmer = \frac{1}{N} \sum_{i=1}^{N-1} |A_i - A_{i+1}| \tag{2.2}$$

Intensity

Intensity is often interpreted as the loudness of sound. The intensity for a given sound frame is calculated as the squared mean of the amplitude of a sound wave within that given frame.

2.2.2 Speech Text Processing

Speech text processing usually comes after the speech audio has been converted to text through a speech-to-text system. We explore the transcripts because the language and the words used are associated with a person's mental state [27].

Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count (LIWC) [28] creates metrics based on the counts of words in different pre-set categories. The recent LIWC version (LIWC2015) is made up of a Dictionary of almost 6,400 words. Each of these words belongs to one or more word categories. Given a certain transcript from a participant, LIWC will categorize each of the words in that transcript into one or more categories based on which category that word belongs to. For example, (example taken from LIWC2015 manual), the word 'cried' belongs to six different categories - sadness, negative emotion, overall affect, verbs and past focus. If this word is found in a certain transcript, the counts of all six of these categories will be incremented.

There are a total of 93 word categories in LIWC; hence 93 feature values can be extracted from a single transcript. However, since the transcript we use is produced using a speech-to-text system, we will not be using some categories, like a category that includes informal language words (e.g. lol, btw, etc...) and some punctuation (e.g. colons, quotation marks, parenthesis, etc...). After removing these irrelevant categories, we are left with a total of 80 LIWC features.

2.3 Machine Learning

The present work seeks to explore the relationship between different features extracted from speech and anxiety, and to use those features in a prediction of anxiety. One way to achieve these goals is to learn from data that exhibits a relationship between anxiety and speech. This is the task of learning from data. In the foregoing, we achieve the goals in two phases: First, exploring significant correlations between speech and anxiety, and second, training models to predict anxiety.

2.3.1 Correlations

Correlation quantifies the statistical relationship between two random variables. The measure of correlation ranges between -1 to 1, where zero signifies no correlation. When learning from data, we use correlations to explore features that have a significant relationship with the labels and also to restrict our feature set to a proper subset which can reduce over-fitting and speed up the training process [29]. The significance of a correlation can be measured using P-value, which indicates whether the correlation between two random variables occurred by chance or not. P values below a low threshold – usually $<.05$ or $<.001$ – indicate a significant correlation.

Two commonly used correlation metrics are Pearson's correlation and Spearman's correlation. Given two arrays of random variables X and Y with N elements in each, Pearson's correlation (shown in equation 2.3) is a measure of linear dependence and does not measure slope nor non-linear relationships. Spearman's correlation (shown in equation 2.4) can measure the non-linear relationship between the two random variables. This non-linear relationship can be measured by taking the rank instead of the actual value of the random variable. For example, the rank of the third-highest value, $R(X_i)$ will be 3.

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \tag{2.3}$$

$$rho = \frac{\sum_{i=1}^{N}(R(X_i) - \overline{R(X)})(R(Y_i) - \overline{R(Y)})}{\sqrt{\sum_{i=1}^{N}(R(X_i) - \overline{R(X)})^2}\sqrt{\sum_{i=1}^{N}(R(Y_i) - \overline{R(Y)})^2}} \tag{2.4}$$

2.3.2 Machine Learning Models

A machine learning model is a computational system that improves prediction performance or accuracy after being trained on a collected data. The idea behind machine learning is based on finding a pattern in a set of data during a training stage and looking for a similar pattern in other unseen data in order to make a prediction. In the coming sub-sections, multiple examples of machine learning models will be described.

Logistic Regression

A binary logistic regression model is trained on data in order to achieve the best accuracy at classifying between two classes. As shown in Equation 2.5, for an input X with D dimensions, the output is the sum of weighted inputs $(X_i * W_i)$ plus a bias (b), passed through a logistic function. The most widely used logistic function is the sigmoid since the output is in the range of 0 to 1 and so can be interpreted as a probability. During training, the weights and bias are adjusted (through, for example, stochastic gradient descent algorithm [30]) so that the output Y approaches the target value.

$$Y = sigmoid(b + \sum_{i=1}^{D} W_i * X_i) \tag{2.5}$$

In this work, our initial baseline models will use logistic regression when we are trying to classify between anxious and nonanxious because logistic regression models are easier to analyze. We can view the weights W_i and see which corresponding input X_i has more importance, assuming that the inputs themselves are normalized.

Support Vector Machine

Support Vector Machines (SVMs) [31] can be thought of as an extension to the Logistic regression method. In the case of Logistic regression, the task is to find a line or a hyperplane that can separate between two classes. An SVM seeks a hyperplane that separates the two classes but further optimizes this separation by maximizing the width of the gap between the two classes. By doing this, SVMs may be able to more correctly classify unseen data that are closer to the decision boundary.

Artificial Neural Networks

Artificial Neural Networks (ANNs), originally inspired by biological neural networks [32], can be thought of as a logistic regression model but with multiple layers. In other words, a Logistic regression is a neural network with a single neuron. There are also various types of activation functions beyond the Sigmoid one mentioned above - for example, Rectified Linear Unit, ReLu; Hyperbolic Tangent Function, tanh; Softmax. Each of these can be used with neural networks, although a sigmoid activation function is usually used on the final output layer for classification tasks in order to get a probabilistic output for binary classifiers. Figure 2.1 shows the structure of a typical neural network.

Unlike Logistic regression, ANNs are not as easy to analyze and investigate since there are multiple layers, each with its own sets of weight, and therefore ANNs are assumed to be black box models. The process of training (achieved through, for example, backpropagation [33]) is also very compute resource intensive, compared to logistic regression models, since the weights have to be updated at every layer and the bigger the neural network, the more time and computing resource the training takes. However, through their many layers, ANNs can approximate more complex functions.

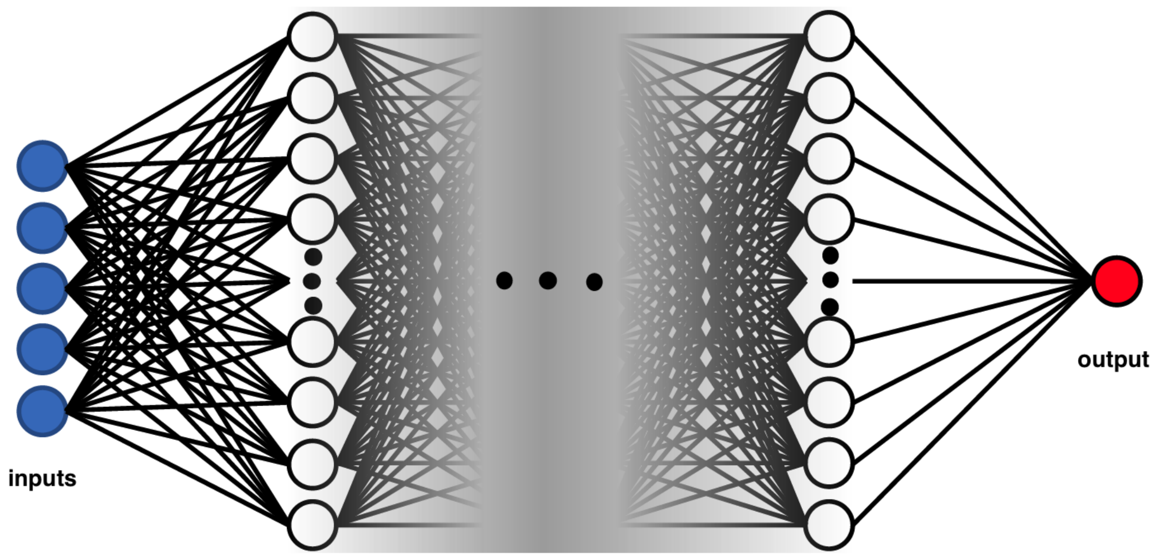


Figure 2.1: Artificial Neural Network

Transformer Neural Network Models and Natural Language Processing

Over the last five years, there have been significant advances of the capabilities in the field of Natural Language Processing (NLP) [34]. A key step was the invention of limited-size word vectors or embeddings, in which it has been shown that a small size (from 50-300) sized vector of real numbers was capable of representing the meaning of individual words or parts of words [35][36][37]. Note that sometimes words are divided into sub-parts and then converted into tokens, which can represent either a full or a partial word. These word/token vectors make it possible to know if two words have similar meanings through a numerical comparison, as well as other encapsulations of meaning through calculation. It also permitted the use of neural networks to process language in a far more effective way and has led to significant advances in the subfields of speech recognition, natural language understanding, question answering, and language generation [35][38].

Another important step that has dramatically improved the state of the art in these fields is the advent of the transformer-based neural network architecture [39][40][41][42][43]. These so-called 'large' language models are trained using massive corpora of text often obtained from the Internet. More specifically, the 'learning' (in the machine learning sense [44]) is done by either predicting the next word in sequence, or predicting intentionally missing words. The architecture of a transformer-style neural network has two important properties: 1) It 'transforms' a sequence of words or parts of words, represented as vectors, into another sequence of vectors. The output vectors account for additional meaning inferred from the full sequence of words, and so create a sequence of so-called 'contextual embeddings' that better encapsulate the meaning of the full input sequence. 2) It makes use of an important neural-network mechanism known as 'attention' [39][45]. Here a part of the network learns several different ways in which parts of a sentence or paragraph are related to other parts of the sentence. For example, a certain word/meaning may typically be connected to specific other words in a sentence. A transformer learns many such relationships, which makes it capable of classifying the broader meaning of a sentence or paragraph. It is this capability that we leverage in the present work (in Chapter 6), to look for patterns of language that indicate the presence of anxiety.

There now exist many such large language models that have already been fully *pre-trained* [40] on massive corpora of text gathered from a number of sources on the internet and elsewhere [40][42][43]. A common use case in the field of deep learning and NLP is to take such pre-trained models, and *fine-tune* [46] them for a specific prediction task that takes language as input. To 'fine-tune' a model means to train it on a (typically much smaller) dataset to learn the task at hand. The task in our current study is the classification of participants into anxious or nonanxious categories.

2.4 Related Work on Anxiety Detection

There have been various prior studies on the automatic detection of different kinds of mental health disorders, including anxiety disorders. In this section, we present related works in the area of the automatic detection of anxiety. This begins with work on the detection of anxiety based on behavioural and physiological measurements and then move to work that focuses on qualities of speech used to measure and detect anxiety.

2.4.1 Detection of Anxiety from Physiological Measurements

The detection of anxiety from physiological measurements requires specialized equipment to measure a physiological signal. Each of the following studies used a different measurement approach.

Pavlidis et al. sought to detect anxiety for the purpose of determining if a person was engaged in illegal activities [47]. Their underlying assumption was that such people would be more anxious. For their study, they recruited six participants and were asked to go through multiple tasks, where one of those tasks was meant to temporarily increase anxiety. The anxiety-producing task was a startle stimulus induced by a sudden loud (60dB) noise created very close to the participant. A thermal camera was used to capture the face of the participants. Their result shows a pixel value change (corresponding to a temperature change) after the startle stimulus, compared to before, in the periorbital area, cheeks and neck area - indicating the changes in facial temperature during anxiety.

Frick A et al. [48] explored the possibility of classifying patients with social anxiety disorder (SAD) from healthy controls using functional magnetic resonance imaging (fMRI). For their study, they recruited 14 SAD patients with a psychiatric diagnosis assessed using the structural clinical interview for DSM-IV and 12 healthy controls. All the participants were male so as to reduce the bias that may be introduced due to gender. fMRI scans were collected while the participants were asked to look at an alternating fearful and neutral faces and were also asked to identify the sex of the face. From the fMRI scans, regional gray matter volume was calculated, and the results were used to train a Support Vector Machine (SVM) machine learning model. A model trained strictly on the fMRI scans of the fear network of the brain to classify between SAD and the healthy controls achieved a 72.6% classification accuracy.

Similarly, there are multiple studies (e.g., see [49] [50] [51]) that used physiological measurements (i.e., heart rate variability, blood volume pulse, galvanic skin response, respiration) to detect anxiety. Although these works show promising results in the automatic detection of anxiety, they are not meant to be utilized for immediate clinical services due to costly processing and requiring specialized equipment.

2.4.2 Detection of Anxiety from Behavioural Responses

There have also been some works that aimed to detect anxiety from behavioural responses, which are a change in behaviour in response to a stimulus. A study published in 2017 [52] explored methods of detecting stress and anxiety from the behavioural response of facial cues. They conducted a study on 23 participants (16 male, 7 female) performing a neutral, relaxing, and stressful/anxious tasks while recording eye movements, mouth activity, head motion, and heart rate, all estimated from a video. Results showed that it was possible to classify between stress/anxiety and neutral state from features extracted from the video recording. The best classification accuracy they were able to achieve was 91.6%. The task that achieved this accuracy was using the different features estimated from a video (mentioned above) and under a social exposure task that classifies between a neutral task (sitting for 1 minute without doing anything) and an anxious task (participants describing themselves in a foreign language).

A study conducted on 228 college students, [53] found a correlation between social anxiety levels and several mobility features acquired from Global Positioning System (GPS) location. The location data, together with the Social Interaction Anxiety Scale (SIAS) score, was collected using an app installed on the participant's smartphone for two weeks. They explored different computational features based on the GPS location over time, including the resident time at a specific location, distribution of visits over time, frequency of transition between locations, and location entropy. The location entropy is a measure of how the time spent at each location is different on different days for a single participant. During weekdays, location entropy had the highest correlation (r=-0.64, P = .001) among all the features. A neural-network-based binary classification model was trained, using several of these extracted features as input. It predicted either a high or low SIAS scores (using score ranges of 0-80 with a cut-off at 34) and achieved a classification accuracy of 85%.

Recent work [8] explored the relationship between passively collected audio data (also collected using a smartphone) with anxiety and depression. The study ran for two weeks, where 84 participants installed an app on their smartphone that collected the volume of sounds and the presence or absence of speech in the environment. From that, they were able to engineer and extract four environmental audio-based features, including daily similarity, sleep disturbance (on all nights and on weeknights only), and speech presence ratio. The participants also completed a self-report measure of anxiety (Liebowitz Social Anxiety Scale, LSAS and the 7-item Generalized Anxiety Disorder scale, GAD-7), depression (8-item Patient Health Questionnaire scale, PHQ-8), and functional impairment (Sheehan Disability scale, SDS). In this work, the audio metrics were shown to correlate with Social Anxiety and depression scales but not with the generalized anxiety disorder scale. Specifically, there was a significant correlation with depression (PHQ-8): daily similarities (r=-0.37, P<.001), sleep disturbance on weeknights (r=0.23, P=.03) and speech presence (r=-0.37, P<.001). Functional impairment measured by SDS also showed a significant correlation (r=-

0.29, P=.01) with speech presence.

2.4.3 Detection of Anxiety from Speech

We will first describe prior work that explored the association between speech features and anxiety. Then we cover prior work on the prediction of anxiety from speech.

Association between Anxiety and Speech Features

Özseven et al. [54] conducted a study that used speech as input with 43 adults between the ages of 17 and 50. Of these 43 adults, 21 were clinically diagnosed with GAD, 2 were diagnosed with panic disorder, and 20 were healthy controls. They collected audio recordings of the following from each participant: 1) Participants reading a neutral passage, 2) Participants reading an "anxious" passage and 3) The patient group describing their most recent anxiety attack and the healthy controls describing a stressful situation all recorded in the same room and environment. A total of 122 acoustic features derived from the participant's speech were explored to determine the level of correlation between these features and anxiety.

They aimed to answer three research questions:

1 Which of these features correlate with anxiety on the basis of data from the two different groups?

2 How do the features change for the anxious group?

3 Is there a relationship between sociodemographics and anxiety?

On the first question, they showed that 30 features from the anxiety/stress description audio were correlated with anxiety (P<.05) and 33 features from the "anxious" passage were correlated with anxiety.

On the second question, within the anxious group, they calculated the correlation between the Beck anxiety inventory (BAI) score [55] collected from the participants and each of the features and found the highest correlation with F0 mean (they didn't present correlation value). On the third research question, they compared the correlation between the different demographics (age, sex, height, BMI, etc..) and a binary representation of anxiety and healthy controls and found that Body mass Index (BMI) had the highest correlation (r=0.319, P=.03).

A similar study found a relationship between anxiety and the changes in voice [56] by finding a link between the vocal pitch - characterized by the fundamental frequency - and social anxiety disorder (SAD). They collected an impromptu speech samples from 46 undergraduate psychology students, 25 of whom had received a SAD diagnosis (together with the severity score based on the Beck Anxiety Scale [24]) while the remainder were healthy controls. There was no significant difference in demographics between the two groups. The Impromptu speech task required the participant to speak about any subject of their choice for 4 minutes in front of an audience member. Their results suggested that male participants' mean fundamental frequency was positively correlated (r=0.72, P = .002) to the SAD patients' diagnostic severity level. In contrast, the correlation for the female participants was not that strong (r=0.02, P = .92), indicating there might be a difference between gender and the effect of anxiety disorder on voice. The authors were also able to identify an optimal cut-off of the mean F0 value that differentiates between the male SAD patients and male healthy controls to be at F0=121.96Hz.

Di Matteo et al. [9] was a continuation of [8] (presented in section 2.4.2 above) where they used the passively collected audio data and applied a speech-to-text in order to capture the words spoken. For privacy reasons, they did not collect the whole transcript but rather the list of words in a random order so that they won't be able to reconstruct the original transcript. They used 67 LIWC word categories [28] as described above, and calculated the correlations with four self-report measures: social anxiety disorder, generalized anxiety disorder, depression, and functional impairment. The highest correlation they found was between depression and words related to death (r=0.41, P<.001), and there was a significant correlation between words related to vision with social anxiety disorder ($r = 0.31, P = .003$) and words related to rewards with generalized anxiety disorder ($r = -0.29, P = .007$).

In a similar study that used LIWC features, Anderson et al. [57] recruited 42 participants diagnosed with SAD and 27 healthy controls to explore the differences in the words used between these two groups. The participants were asked to write a distinct autobiographical and socially painful passage. They used LIWC to extract word counts in each of the LIWC categories, such as first-person singular, anxiety-related words and fear-related words. Their results showed that SAD patients used more first-person singular pronouns (I, me, mine), anxiety-related words, sensory/perceptual words, words denoting physical touch, and fewer references to other people.

Hofmann et al. [58] examined the association between linguistic features and social anxiety disorder (SAD). They recruited 24 participants diagnosed with SAD and 21 healthy controls. The participants were asked to give a speech on any topic of their choice for a total of 4 minutes in front of one experimenter while being video recorded. To induce stress and anxiety in the participants, they were told that a panel of judges would conduct a post-rating of their speech regarding poise, social confidence, and general presentation skills. Later the speech was transcribed, and LIWC was used to extract the count of the words in different pre-set categories. More specifically, they obtained the counts for first-person pronouns, negative emotion words, and positive emotion words. Their results showed that SAD patients used more positive emotion words compared to the healthy controls. The authors did not observe any significant difference for the other explored LIWC categories.

Sonnenschein et al. [59] explored the transcripts from a recorded therapy session of 85 patients. These patients were categorized into three groups: diagnosed with anxiety but not depression, depression but not anxiety, and both anxiety and depression. The LIWC system was used to obtain counts in four categories: first-person singular, Sad, Anxiety, and Filler. Their results showed that the depressed but not anxious group showed a higher usage of Sad words compared to the anxious but not depressed group. The anxious but not depressed group, on the other hand, showed a higher usage of anxiety words compared to the depressed but not anxious group. The "both anxious and depressed" group also showed a higher usage of "sad" words than the anxious but not depressed group.

Prediction of Anxiety from Speech

A study published in 2019 [60] describes a machine learning approach to detect what is collectively known as internalizing disorder - anxiety and depression - in children. To build a machine learning model, they collected speech data from 71 children, where 43 children had been diagnosed as having an internalized disorder after a clinical interview with their primary caregivers. The data collected was a three-minute speech task based on the Trier Social Stress Task for children (TSST-C) [61]. Here they were told to prepare and present a 3-minute speech that will be judged based on how interesting it is. The researchers selected a set of features previously proposed in the literature for identifying anxious and depressive symptoms in adults including: zero-crossing rate, MFCC, dominant frequency, mean frequency, perceptual spectral centroid, spectral flatness, skew and kurtosis of the power spectral density. These features were extracted from the audio samples. Using a Davies-Bouldin-Index-based feature selection method to select the top 8 features (the highest was the MFCC 267-400Hz frequency range), built a series of prediction models using logistic regression, SVM, and random forest. Using the Leave-One-Out Cross Validation (LOO-CV), the highest accuracy they achieved was 80% on both a Logistic regression and an SVM model.

Continuing from their previous study (described in Section 2.4.3 above) on finding a relationship between anxiety and alteration in voice [56], Weeks et al. attempted to classify between men with social anxiety disorder and healthy controls [62]. They chose to study only men since they found a significant positive correlation between the mean fundamental frequency and anxiety. Using a mean fundamental frequency value of 122.78 Hz, they were able to achieve a sensitivity of 89% (8/9 male SAD patients correctly classified) and a specificity of 100% (4/4 male healthy controls correctly classified).

Salekin et al. [63] explore methods of detecting social anxiety and depression from an audio clip of speech. Their dataset includes a 3-minute speech from each of the 105 participants describing the things they liked and disliked about college or their hometown. Participants were asked to report their peak level of anxiety during the speech. The authors present and use a novel feature modelling technique called NN2Vec that can identify the relationship between a participant's speech and affective states. Using the features from NN2Vec and a bi-directional Long Short Term Memory Multiple Instance Learning (BLSTM-MIL) network, they were able to detect speakers with high social anxiety with an F-1 score of 90.1% and speakers with depression symptoms with an F-1 score of 85.4%.

Baird et al. [64] explored the effect of anxiety on speech by attempting to predict anxiety from sustained vowels. Their dataset comprises 239 speakers (69 males) aged 18 to 68 performing various vocal exercises, which included sustained vowels. Each participant filled out a Beck Anxiety Inventory (BAI) [55], which is a scale used to screen for GAD. They used four classes of sustained *a* vowels from each participant: a sad phonation, a smiling phonation, a comfortable phonation, and a powerful/loud phonation. From the sustained vowels, they extracted acoustic features such as standard deviation of F0, intensity, and harmonic-to-noise ratio. Using these features and a BAI label, they trained a Support Vector Regressor (SVR) with a linear kernel. To evaluate the performance of their model, they used Spearman's correlation between the predicted and the actual label. Using all the data and splitting it into training and test data, they achieved a Spearman's correlation of 0.243 on the test data. However, they had a better performance (Spearman correlation of 0.59) when they only considered the group with High BAI scores indicating symptoms of anxiety are more observable in individuals with high anxiety.

Rook et al. [65] aimed at predicting GAD from linguistic patterns under the hypothesis that worrying behaviour in GAD comes from a verbal-linguistic process. A total of 142 undergraduate participants (56 men and 86 women) were recruited. Each were asked to recall and write down an anxious experience during their university life, and to fill out the GAD-7 scale, as well as a BIS/BAS (behavioural inhibition/behavioural approach) scale [66]. The LIWC features [28] were extracted from the texts written by the participants. Another set of features was also used by combining the LIWC features with the BIS/BAS scores, which is somewhat unusual, as these merge the labels with the data. Several machine learning models were explored, including Support Vector Machine with the linear kernel (SVM), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF). Their results show that all models built using the LIWC features performed significantly better than a random model and performance generally improved (somewhat unsurprisingly) when the LIWC features were used together with the BIS/BAS scores as input features.

In their third study using the data collected from the 84 participants, Di Matteo et al. [10] attempted to predict generalized anxiety disorder (GAD), social anxiety disorder (SAD), and depression from the smartphone-collected data. The features used in this study include daily similarity, speech presence, weeknight sleep disturbance, death-related words, number of locations visited, number of exits from home, screen use, and time in darkness. Although the models built on these features achieved an above-random prediction accuracy for social anxiety disorder and depression, they did not observe above-random prediction accuracy for GAD.

2.4.4 Summary

All the prior research work reviewed shows promising results in the measurements and detection of anxiety. However, the acquisition of data for analysis might make some better than others. For example, the acquisition of data from video recording as described in [52] might be a preferred choice of input compared to acquiring data through fMRI scans as in [48], since the latter would be more expensive. An even better solution would be passively collecting data provided by smartphone sensors (location data, for example, presented in [53]) since the individual does not have to be actively providing the required data for processing (as in a video recording). But this might be an issue for some individuals who do not want their privacy being invaded by collecting data in the background. A somewhat middle-ground solution would be to collect speech data. It is not as invasive as collecting your GPS location and/or all the data provided by your smartphone sensor, yet it would be relatively cheaper to collect. The individual should be made aware that only speech data would be collected and therefore allow this. We, the researchers, should have the ethical responsibility to respect the privacy and confidentiality of the individual.

Overall, prior studies suggested multiple features that may allow us to detect anxiety from speech. However, among these studies that attempted to find an association between speech features with anxiety and predict anxiety from speech, the largest sample size was 239, which limits the potential for generalizability to the larger population; a necessary step before considering the deployment of technologies for passive anxiety monitoring. In this study, we recruit a substantially larger cohort (N = 2,000) to explore features of speech from previous findings at a greater scale.

Chapter 3

Data Collection

The exploration of methods to detect anxiety in speech requires the collection of speech samples together with relevant labels. This chapter describes the data collection methodology and design used to collect speech samples from participants and those participants' anxiety severity levels. Here we describe the recruitment inclusion criteria, the protocol for collection, the software system built to implement the protocol and to ensure data quality, as well as the data inspection process.

3.1 Overview

Participants from a nonclinical population were recruited for a 10-to-15-minute task implemented through a custom website. Self-report measures of anxiety were collected once at the beginning of the study and at the end of each of two specific tasks. In the sub-sections below, we describe the recruitment of participants, the data collection procedure, and the assessment of anxiety and speech measures.

3.2 Recruitment and Inclusion Criteria

A total of 2,000 participants were recruited using Prolific [67] - an online human subject recruitment platform. This number was chosen to both be two orders of magnitude greater than most of the prior work described in Chapter 2, and to be a tractable number within our time and cost constraints.

Prolific maintains a list of registered participants and, for each participant, many characteristics, including age, income, sex, primary language spoken, country of birth, and residence.

The inclusion criteria for this study were:

1 Age range 18–65 years.

2 Fluency in English.

3 English as a first language

4 At least ten previous studies completed on Prolific.

5 A minimum of 95% of previous Prolific tasks completed satisfactorily.

The dataset was also balanced for sex (50% Female, 50% Male).

Participants who completed the study were paid £2 (approximately $3.41 CAD). They were able to complete the entire study remotely, using their personal computers.

3.3 Study Procedure

Subjects were presented with the opportunity to participate in this study on Prolific if they met the inclusion criteria given above. Those who wished to participate clicked on the study link, which brought them to a consent form that described the procedure and goals of the study and provided information on data privacy. After they gave consent, a hyperlink brought participants to an external web application (a screenshot of which is shown in Appendix C) that implemented the tasks described below.

Participants were first asked to fill out the standard GAD-7 questionnaire [18] described in more detail in Section 3.4 below. Then, they were asked to complete two speech tasks, which were recorded using their computer's internal microphone. Note that our protocol also recorded a video of the participants' faces during both speech tasks. Although that video is not used in the work reported in this study, the fact that the video was requested may have influenced the set of participants willing to continue participation, as discussed later.

For the first speech task (Task 1), participants were asked to read aloud a specific passage called "My Grandfather," which is a public domain passage that contains nearly all the phonemes of American English [68]. The full script of this passage appears in Appendix D. This passage is not intended to induce stress or anxiety, but to provide a baseline speech sample for each participant. It was used in this work to test the quality of the speech-to-text transcription.

For the second speech task (Task 2), the participant followed a modified version of the widely-used Trier Social Stress Test (TSST) [69], for the purpose of inducing a moderate amount of stress. We have chosen to base our anxiety stimulus on the TSST because previous studies ([70][71]) have shown a higher activation in participants with relatively higher anxiety after exposure to moderate stress induced by the TSST.

In this modified version of the TSST, participants were told to imagine that they were a job applicant for a job that they really want (their 'dream' job), and they were invited for an interview with a hiring manager. They were given a few minutes to prepare – to decide what their ‘dream’ job is – and how they would convince an interviewer that they are the right person for the position. Participants were also told that the recorded video would be viewed by researchers studying their behaviour and language. Participants were then asked to speak for five minutes, making the case for themselves to be hired for that dream job.

Note that, in the original TSST [69], participants would normally deliver their speech in front of a live panel of judges. If a participant finished their delivery in less than five minutes, the judges in the original TSST design would encourage the participant to keep speaking for the full five minutes. An example statement of encouragement is: “What are your personal strengths?” In our modified TSST, we implemented a similar method to encourage participants to speak for the full five minutes: When our software detects silence (the absence of speech for more than 6 seconds), it will display several different prompts, which are reproduced in Appendix E, inviting participants to keep speaking on different topics relating to the task. Finally, note that the modified TSST only made use of the first part of the original TSST, and not the second task involving mental arithmetic.

3.4 Anxiety Measures

To measure the severity of GAD, we used the GAD-7 [18] scale, which is a seven-item questionnaire that asks participants how often they were bothered by anxiety-related problems during the previous two weeks. While the two-week time period suggests that the GAD-7 measures a temporary condition, this seems in contradiction with the fact that a GAD diagnosis requires six months duration of symptoms [72][17]. However, the GAD-7 has been validated as a diagnostic tool for GAD (using a value of 10 as the cut-off threshold) with a sensitivity of 89% and a specificity of 82% [18]. Thus, we choose to use the GAD-7 to obtain a binary label of GAD (using the same threshold of 10) as our main indicator of anxiety.

Each of the seven questions on the GAD-7 has four options for the participant to select from, indicating how often they have been bothered by the seven problems in the scale. These options and their numerical ratings are: 0-Not at all, 1-Several days, 2-More than half the days, and 3-Nearly every day. The final GAD-7 score is a summation of the values for each question, giving a severity measure for GAD in the range from 0 (no anxiety symptoms) to 21 (severe anxiety symptoms).

We also employ a second informal anxiety measure in this study to serve as an internal check to measure how much, on average, the modified TSST (Task 2) has induced stress and anxiety compared to Task 1 (the reading/speaking of the Grandfather passage). Here we use a single question to measure self-reported levels of anxiety on a four-point scale. We ask participants how anxious they felt during the task and to choose from the following numerical rating: 0-Not anxious at all, 1-Somewhat anxious, 2-Very anxious, and 3-Extremely anxious. This question is deployed immediately after the first and second tasks.

3.5 Development of Recruitment Software

This section describes the construction and deployment of a software used to collect participants' data.

3.5.1 Platform Selection

The first design choice in building a system that can be used to collect data is to decide which platform to deploy the software on. There are three main categories where we can deploy software:

1 **Desktop/laptop application:** a computer software that is run locally on a computer

2 **Mobile/smartphone/tab app:** installed on a mobile device and has the ability to access mobile resources, such as a camera, GPS, accelerometer, etc.

3 **Website:** Accessed via the internet through both a desktop machine or a mobile phone.

While each of these has its own advantages and disadvantages, we wanted to select the platform that is appropriate to our task. These are the criteria we listed that our software should include:

(a) Must be able to capture speech data.

(b) Should not be Operating System specific.

(c) Must be able to reach the majority of the population throughout the world.

(d) No need to download and install any software.

(e) Easy to update (centrally without the need to reinstall).

A custom-built website software satisfies all of the above criteria (as shown in Table 3.1), so designed one of these to perform the data collection.

Table 3.1: Platform comparison

	Desktop app	Mobile app	Website
(a)	✓	✓	✓
(b)	×	×	✓
(c)	✓	✓	✓
(d)	×	×	✓
(e)	×	×	✓

3.5.2 Software Development

The website used to recruit participants was developed using Google Web Designer [73]. Google Web Designer is an advanced web application that allows developers to design and build HTML5 web content. It has a drag-and-drop page builder, which makes building websites faster and easier for novice developers. We used Google Web designer to develop the front-end participant-facing web page.

The data from the participants was stored in the Google Firebase Realtime Database [74]. A JavaScript language application was written as part of the front-end to connect, authenticate, and send participants' data to the back-end Firebase database.

3.5.3 Software Details

In this sub-section, we describe some of the details of the software implementation that dealt with issues that arose during the study. These improvements allowed us to successfully recruit a very large amount of participants in a very short amount of time.

Silence Detector

During the TSST task, participants would usually describe their dream job within a minute and then stay silent for the remaining amount of time (which was 5 minutes in total). In the original TSST, however, the panel of judges would ask questions to the participants to encourage them to keep speaking. The speech encouragement questions are presented in Appendix E. We implemented an equivalent version of this functionality. To monitor if participants were speaking, we used a speech-to-text system. The speech-to-text system uses Web Speech API [75] to get the transcripts in real-time, and it is not the speech-to-text software we used to get the final transcript. If that system produced a transcript in real-time, then we could conclude that the participant was speaking. Conversely, if there is no transcript observed for more than six seconds, then silence has been detected, and a speech encouragement statement will be displayed.

Language Translation Detection

We observed that for a few study participants, we noticed that some participants spoke a different language even though we specified (on Prolific) that we wanted English-speaking participants. Even the "Grandfather passage" that was written in English was being read in another language by some of the participants. We determined that some participants' web browsers were set to perform translation based on their location and the language spoken at that location. So, it was necessary to add a restriction against automatic translation in the code on the front-end and so force all languages to be displayed in English as it had been written. This eliminated the problem.

Non-Functional Video and Audio

There was also an issue with submissions received that had no working video or muted audio. To prevent this, we included a page at the beginning of the front-end sequence to ensure that the participants' audio and video were working correctly. For the video, we turned on their webcam and asked the participants if they could see themselves clearly with sufficient lighting. We also double-checked, after submission, if the participant's face could be clearly seen by using the OpenCV [76] face detection library to check and see if a face has been detected. For the audio, we again used the speech-to-text method described above - participants were asked to read a short sentence aloud, and if that produced a transcript, we could safely conclude that their audio was working. If not, the participants were asked to check if their microphone was installed correctly and if they could check their settings.

3.6 Ethics and Privacy

Since this study includes human participants ethics approval was required and received. The study was approved by the University of Toronto Research Ethics Board (REB) under protocol #37584.

The approved ethics protocol by the REB is provided in Appendix G

3.7 Potential Conflicts

A collaborator in the study, Dr. Bill Simpson from Winterlight Labs, works for Winterlight Labs and so the knowledge and information gained from this work could benefit Winterlight and therefore Dr. Simpson. Indeed, the goals of Winterlight are very similar to the goals of this research project, and that is the reason for the collaboration.

3.8 Recruitment Outcome and Overview of Participant Data

This section discusses the results of the data collection study described in prior sections of this Chapter. We review the recruitment yield, provide an overview of the collected data, and the demographic characteristics of the participants.

3.8.1 Recruitment and Data Inclusion

Recall that the study recruitment occurred on the Prolific [77] platform, and that we sought a total of 2,000 participants. Recruitment was launched in phases, with roughly 250 participants gathered at one time, in approximately 8 phases over the period of November 23, 2020 to May 28, 2021. (We note that recruitment began 8 months after the beginning of the global COVID-19 pandemic.) During that time, a total of 4,542 participants accepted an offer from Prolific recruitment platform to participate in the study. From those, 2,212 participants finished the study, giving a recruitment yield of 49%.

Of the 2,212 participants who completed the study, 2,000 provided acceptable submissions (and thus received payment), giving a submission-to-approval yield of 90%. To be clear, the recruitment continued until 2,000 acceptable submissions were received. The reasons that submissions were deemed unacceptable include: a missing video, a missing or grossly imperfect audio, or a failure to complete one or both tasks. These *submission* acceptability criteria are distinct from those used in a subsequent review of audio quality that is described below.

In addition to the above submission acceptability criteria, we reviewed the input data and audio for a higher standard of acceptability, after the recruitment phase, using the following procedure: We first computed the acoustic and linguistic features described in Section 2.2.1 and Section 2.2.2. Recordings with poor quality were selected for manual (human) review based on the following criteria:

(a) A Task 2 word count lower than 125.

(b) A speaking duration for Task 2 lower than 60 seconds (as compared to the full five minutes).

(c) Any other feature value being more than three standard deviations from the mean, in either direction.

A total of 193 participant recordings were flagged based on these criteria. For each of these, a researcher listened to the associated Task 2 audio recordings. The researcher discarded any samples that were deemed, subjectively, to be of insufficient audio quality, or those whose response to Task 2 was not responsive to the task itself. A total of 123 out of the 193 flagged participants were rejected through this manual review, leaving at total of 1,877 samples.

Finally, the samples were checked for missing data, with 133 participants having missing demographic info. Consequently, the final number of participants included in our analysis is 1,744. The flow chart of the study recruitment and quality control is given in Figure 3.1.

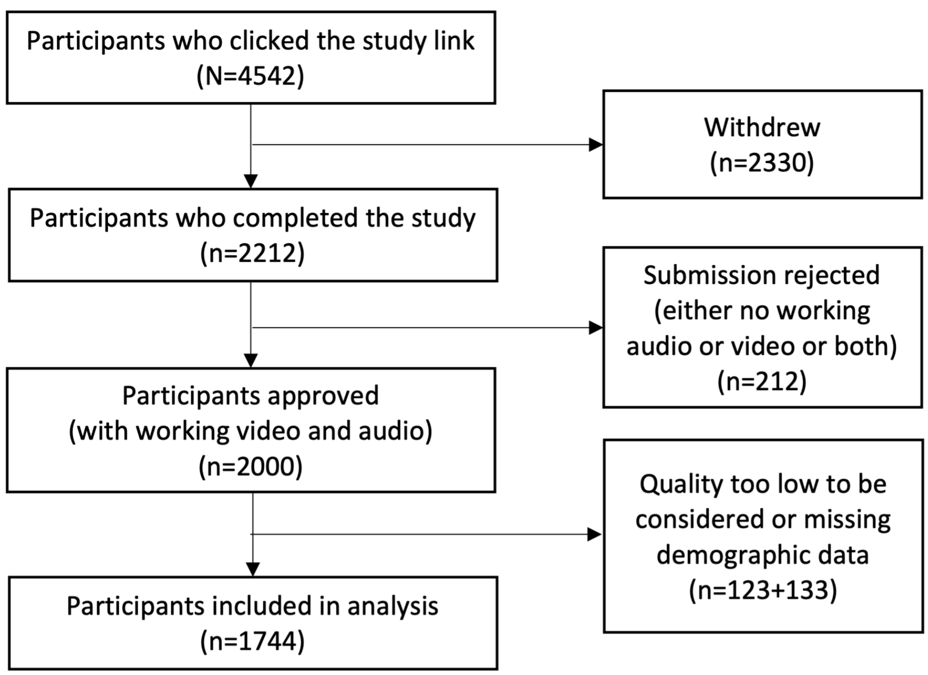


Figure 3.1: Study recruitment flow chart

We have also explored correlations on the 256 samples that were excluded from the study with the GAD-7, often referred to as *missingness* analysis, and it is presented in Appendix F. Later in Chapter 4 we will compute the correlations between the extracted speech features and the GAD-7. The results indicates that excluding the 256 samples did not affect the correlation results since we observed similar significant positive and negative correlations in both the included and excluded data.

Figure 3.1, the study recruitment flow chart, shows that the recruitment yield was 49%. For the 2,330 participants who dropped out after accepting the study, we can only speculate as to why. Some may have been unwilling to have their words recorded in audio or their full video recorded, and even though the consent form makes this task apparent, it may be that the participants who dropped out only really understood this when seeing their faces on the computer screen.

3.8.2 Data Overview and Demographics of Participants

Table 3.2 shows how many of the 1,744 participants are above and below the GAD-7 screening threshold of 10 [18]. In the foregoing, we will refer to those participants who have a GAD-7 score ≥ 10 as the anxious group, and those with a GAD-7 score < 10 as the nonanxious group.

GAD-7 Category	
Above threshold (≥10) Anxious Group	540 (31%)
Below threshold (<10) Non-Anxious Group	1204 (69%)

Table 3.2: GAD classification of 1,744 accepted participants

Table 3.2 shows that the proportion of participants in the anxious group (those above the GAD-7 screening threshold of 10) is 31%, which is much higher than the general population rate of roughly 10% [78]. This result, that English speakers recruited from Prolific have elevated rates of anxiety and depression, is consistent with prior studies using recruits from Prolific and suggests that this population exhibits higher incidence of anxiety [9][8][10]. Table 3.3 does include data that is consistent with this difference, as the proportion of participants self-reported on their Prolific profile that they have an ongoing mental health condition is 34.9%.

Table 3.3 shows participants' demographics, obtained from the Prolific recruitment platform. Columns 1 and 2 of the table shows the name of demographic attributes and each category, while columns 3 and 4 give the number (and percentage) of participants with that attribute in the anxious and nonanxious groups, respectively. Column 5 gives the P-value for a chi-square test of the null of independence, to determine if there is a significant difference between the anxious and nonanxious groups, for each categorical factor.

The demographic data listed in Table 3.3 provides several interesting insights on the recruited cohort, with respect to the presence or absence of above-threshold GAD. First, there is a significantly larger proportion of females in the anxious group com-

Demographic Factors		Anxious (N=540)	Non-anxious (N=1204)	P value
Sex	M	229 (26.0%)	653 (74.0%)	<.001 (χ^2)
	F	311 (36.1%)	551 (63.9%)	
Self-reported ongoing mental health illness/condition	Yes	297 (48.8%)	311 (51.2%)	<.001 (χ^2)
	No	243 (21.4%)	893 (78.6%)	
Personal Income (GBP)	Less than £10,000	181 (39.2%)	281 (60.8%)	<.001 (χ^2)
	£10,000 - £19,999	112 (35.0%)	208 (65.0%)	
	£20,000 - £29,999	92 (26.2%)	259 (73.8%)	
	£30,000 - £39,999	60 (24.6%)	184 (75.4%)	
	£40,000 - £49,999	36 (24.8%)	109 (75.2%)	
	£50,000 - £59,999	20 (21.3%)	74 (78.7%)	
	>£60,000	39 (30.5%)	89 (69.5%)	
Age	18-19	27 (38.0%)	44 (62.0%)	<.001 (χ^2)
	20-29	239 (38.7%)	379 (61.3%)	
	30-39	162 (32.7%)	334 (67.3%)	
	40-49	67 (23.4%)	219 (76.6%)	
	50-59	39 (22.8%)	132 (77.2%)	
	>60	6 (5.9%)	96 (94.1%)	

Table 3.3: Demographic characteristics of participants in the group with anxiety and the nonanxious group (N=1744)

pared to the males. This is consistent with previous findings suggesting that anxiety is more prevalent in females than males [79]. We feel that this confirms that it is useful to consider separate female-only and male-only datasets to avoid the bias introduced by sex when exploring features that may correlate with GAD-7. For example, pitch (F0) would typically be higher for females, and as a result, sex effects could easily confound the association between pitch and anxiety.

The rows of Table 3.3 that show the proportion of anxious/nonanxious participants by income suggest that there is a relationship between income and anxiety: the two very lowest categories of income show a disproportionately higher amount of anxiety. There is a downward trend of anxiety with income until the very last category, which is £60,000 and above. It is interesting that above a certain income level, anxiety seems to increase, although this is consistent with prior studies on anxiety and income [80].

Similarly, with respect to age, younger participants are more likely to be in the anxious group, which is consistent with previous work [81].

3.8.3 Post-Task Self-Report Anxiety Measure

As described in Section 3.4, participants were asked to rate their state anxiety after each task on a scale from 0 to 3, where 3 is the highest level of anxiety. Table 3.4 gives the average value of this informal rating after Task 1 and Task 2. We report a paired t-test to assess the difference between the two measurements. A histogram plot is shown in Figure 3.2. The test validates that the modified TSST task successfully induced some anxiety in participants, with the average score on the self-reported state anxiety measure increasing from 0.5 to 1.6 ($P<.001$), before and after completing Task 2.

	Task 1	**Task 2**	**Paired t-test *P*-value**
Average Post-Task Self-Reported Anxiety measure (SD)	0.5 (0.6)	1.5 (0.9)	$< .001$

Table 3.4: Post-task self-report anxiety measure and paired t-test

As described in Section 3.4, we used the post-task, self-reported anxiety measure as an internal check to see if Task 2 (the modified TSST task) induced more self-reported anxiety as compared to Task 1. Table 3.4 gives the paired t-test conducted on the two informal ratings of anxiety of the two tasks. It had a P-value $<.001$, indicating a significant difference and that Task 2 induced greater anxiety. Recall that most of the prior work discussed in Section 2.4 also used mood induction tasks.

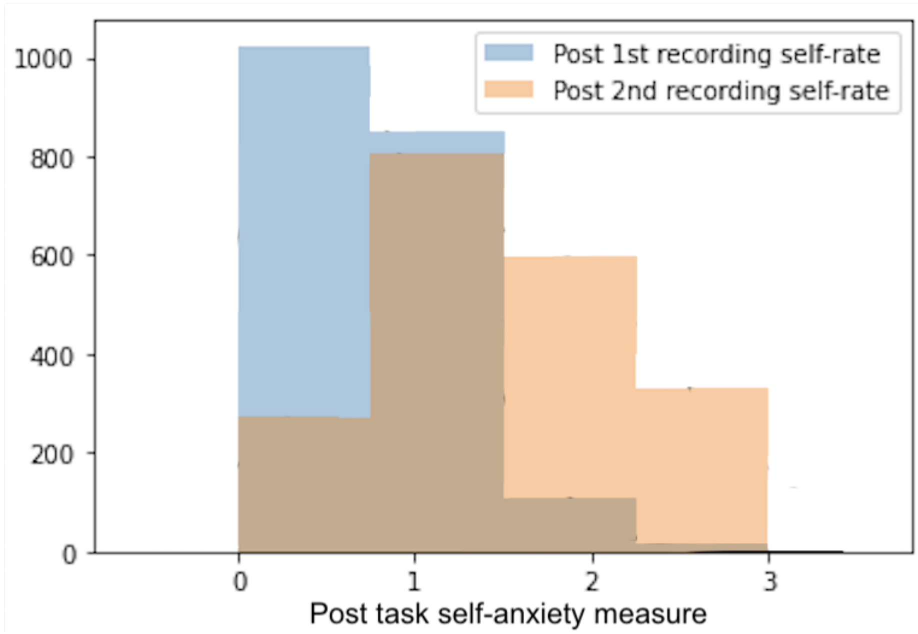


Figure 3.2: Histogram plot of the post 1st recording self-rate and the post 2nd recording self-rate

3.9 Summary

This chapter started with the discussion of the data collection study, which was deployed to recruit participants and collect speech samples from those participants along with the GAD-7 score. We described the recruitment software that was used, the participants' recruitment platform, the inclusion criteria for our study, the study procedure, and the ethics protocol procedures that allowed us to run the study.

This chapter also presents and discusses the outcome of the participant recruitment study. One contribution of this work is the creation of a very large dataset that can be used to train models to predict the presence or absence of anxiety. The number of the participants recruited for this study is two orders of magnitude greater than prior recruited participants for automatic prediction of anxiety studies. This has the advantage of having a model built, on this data, to have more generalizability than previous studies.

Chapter 4

Acoustic and Linguistic Features of Speech and their Association with Anxiety

In this chapter, the acoustic and linguistic features of speech suggested by prior work to have an association with anxiety will be described and we will validate if the association with anxiety can be verified with our larger-than-prior work's sample size. This work was published in [82].

4.1 Overview

As described in Chapter 3, a large number of participants (N = 2,000) were recruited using the Prolific [77] web-based human participant recruit platform, and participated in a single online study session. Participants completed the Generalized Anxiety Disorder-7 item scale (GAD-7) assessment, read aloud a specific passage titled *My Grandfather*, and provided an impromptu speech sample in response to a modified version of the Trier Social Stress Test. Acoustic and linguistic speech features were a-priori selected based on the existing speech and anxiety literature, together with related features. Associations between speech features and anxiety levels will be assessed using age and personal income included as covariates.

4.2 Methods

4.2.1 Selection of Acoustic and Linguistic Features

Prior work suggested that information about the mental state of a person may be acquired from the signals within speech acoustics [83] as well as the language used [27]. We refer to each kind of this extracted information as a feature using the terminology employed in the field of machine learning.

In this work, we consider specific acoustic and linguistic features which are described in the following sections. These features were extracted from each of the 5-minute speech samples in which the subject responded to the modified TSST task. Note that all the participants are prompted to speak for the full five minutes, as described in Section 3.3, although the total speech duration of each participant may vary.

Acoustic Features

Previous research has identified several acoustic features that are correlated with anxiety, as described in Section 2.4. Using these previous findings as a reference point, we select the following acoustic features for our empirical analysis. These features are also described in Appendix B.

- `First 13 Mel Frequency Cepstral Coefficients (MFCC)`: These are the coefficients derived from a mel-scale cepstral representation of an audio signal, as described in Section 2.2. We include 13 MFCCs, a common set of acoustic signals that are designed to reflect changes in perceivable pitch. The MFCC features were shown to be related to anxiety in [84][60][54]. Descriptive statistics (mean and standard deviation) of the 13 MFCC features were used in the current study. Note that not all MFCC features included in the current study were determined to be significant in prior work; however, these 13 are most commonly assessed together so we included them all as features of interest. The parameters we used when extracting these 13 MFCC features are: window length = 2048 samples; length of FFT window = 2048 samples; samples advance between successive frames = 512 samples; Window type = Hanning; Number of Mel bands = 128.

- `First 13 Linear Prediction Cepstral Coefficients (LPCC)`: These are the coefficients derived from a linear prediction cepstral representation of an audio signal. The LPCC features were shown to be related with anxiety in [54]. Descriptive statistics (mean and standard deviation) of the 13 LPCC were used in the current study.

- `ZCR zPSD`: The Zero Crossing Rate (ZCR) of the z-score of the Power Spectral Density (PSD). In [60], ZCR-zPSD was one of the top features selected using Davies-Bouldin index based feature selection [85] for an anxiety prediction task.

- `Amount of Speech`: The amount of speech and related metrics such as the percentage of silence were shown to be related with anxiety in [86] [87][84]. The specific features we will use are: the amount of time, in seconds, that speech was present, and total number of words present in a speech-to-text transcript.

- `Articulation rate`: Indicates how fast the participant spoke. Work by [88] suggests that patients with panic disorder spoke significantly slower during autobiographical talking as compared to reading a script.

- `Fundamental frequency (F0)`: is the frequency at which the glottis vibrates or, also known as *pitch* of the voice. Several studies have shown F0 to be one of the acoustic features that are affected by anxiety [54][56][86][88]. The fundamental frequency varies throughout a person's speech, so both the mean and standard deviation of F0 are used as features.

- `F1, F2, F3`: The first, second and third formants [25]. The work in [54] shows a significant relation with anxiety. The mean and standard deviation of each format were used as features.

- `Jitter`: The cycle-to-cycle F0 variation of the sound wave, also known as *jitter* have been shown to be an indicator of anxiety [54][89][90].

- `Shimmer`: The cycle-to-cycle amplitude variation of the sound wave, also known as *shimmer* has been shown to be related with anxiety severity [54].

- `Intensity`: The mean squared of the amplitude of the sound wave within a given frame (1,764 samples), also known as *intensity* has been shown to be related with anxiety [86]. Since the amplitude of a sound wave varies during speech, the mean and standard deviation were used as features.

The features listed above were extracted using the following software packages: My Voice Analysis [22], Surfboard [91], and Librosa [92].

Linguistic Features

A transcript of the audio recordings was produced using a commercial speech-to-text system from Amazon Web Services [93]. Linguistic features were extracted using the Linguistic Inquiry and Word Count (LIWC) software [28], which places words into dictionaries based on semantic categories, as described in Section 2.2.2. Recall that one category is called 'negemo' and contains words that relate to negative emotions, such as hurt, ugly and nasty. Another category is called 'health' and contains words such as clinic, flu and pill. There is also a category called 'anxiety' which includes words such as anxiety and fearful. Some categories are contained within others - for example anxiety is contained within negemo.

The LIWC software simply counts the number of words that belong to each category, and each count becomes a feature. There are a total of 93 categories in the LIWC, but not all are relevant for a speech-to-text transcript. We have removed those features that are not relevant - for example informal language words such as 'lol' and 'btw.' Other excluded categories include those relating to some punctuation (e.g., colons, quotation marks, parentheses). Removing these, a total of 80 linguistic features remained. Prior work [9][57] that was discussed in Section 2.4 has shown that LIWC categories related to perceptual processes (see, hear, feel), words related to reward, the use of first-person singular pronoun, and anxiety related words were associated with anxiety.

4.2.2 Separation of Data for Analysis

The overarching objective of this study is to gain an understanding of which features of speech – both acoustic and linguistic – are correlated with the GAD-7 scale. It is known, however, that certain demographic attributes are directly indicative of anxiety. For example, sex is known to influence prevalence of anxiety [79]. Also, both age [81] and income [80] influence anxiety which suggests the need to control for these demographics. An additional reason to control for the demographics is that both age and income have been shown to be related to speech features [94][95]. Due to the strong effect of sex on GAD-7 score, we created a separate sample-set for analysis of female and male samples, in addition to the combined dataset. We chose to do this rather than correcting for sex computationally, as it leaves the data intact. Note that also the tasks undergone by the participants to perform may activate different pathways in the brain for different sexes, giving rise to different types of information coming from male and female participants.

4.2.3 Statistical Analysis

The partial Pearson's correlation coefficients [96] was computed between each of the features and the GAD-7 (controlling for the effect of age and personal income). Correlations were examined for three versions of samples: the entire sample data set, and separately by sex for male and female participants. We have considered a result statistically significant at a P value significance level of .05. The P values were not corrected to account for the large number of tests, since we are attempting to use features that were determined to be significant in previous works.

4.3 Results

Section 4.2.1 describes the set of acoustic and linguistic features that were selected. These were features that were reported as significant in prior work on anxiety and speech, as well as closely associated features. These features were computed on the speech samples of participants performing Task 2, the modified TSST. The subsections below summarize the main empirical results. Correlation between demographics and the acoustic/linguistic features is presented in Appendix I and inter-correlation between the significant features is presented in Appendix J, K, and L for the all-sample, female sample and male samples, respectively.

4.3.1 Amount of Speech

The features with one of the highest correlations, for both the male and female samples, are those related to how much the participant spoke during Task 2. Two specific features used to estimate speech length are speaking duration (the number of seconds of speech present within the five-minutes speech task) and the word count derived from a speech-to-text transcript. Table 4.1 gives the correlation for the all sample (a) (controlling for sex, age, and income) and for separated female (b) and male samples (c) (controlling for age and income). Figure 4.1 presents a scatter plot of speaking duration vs. GAD-7 as well as the distribution of both variables, for all three data sets. The scatter plot is coloured to give a better sense of the density of data points. Figure 4.2 provides the same kind of scatter plots/distributions for the word count metric of Task 2 for the all samples (a), male samples (b) and female samples (c).

a) All Participants (N=1744)

Feature	**r**	***P***
Speaking Duration	-0.12	<.001
Word Count	-0.12	<.001

b) Female Participants (N=862)

Feature	**r**	***P***
Word Count	-0.13	<.001
Speaking Duration	-0.11	<.001

c) Male Participants (N=882)

Feature	**r**	***P***
Speaking Duration	-0.13	<.001
Word Count	-0.12	<.001

Table 4.1: Correlation of amount of speech features with the GAD-7

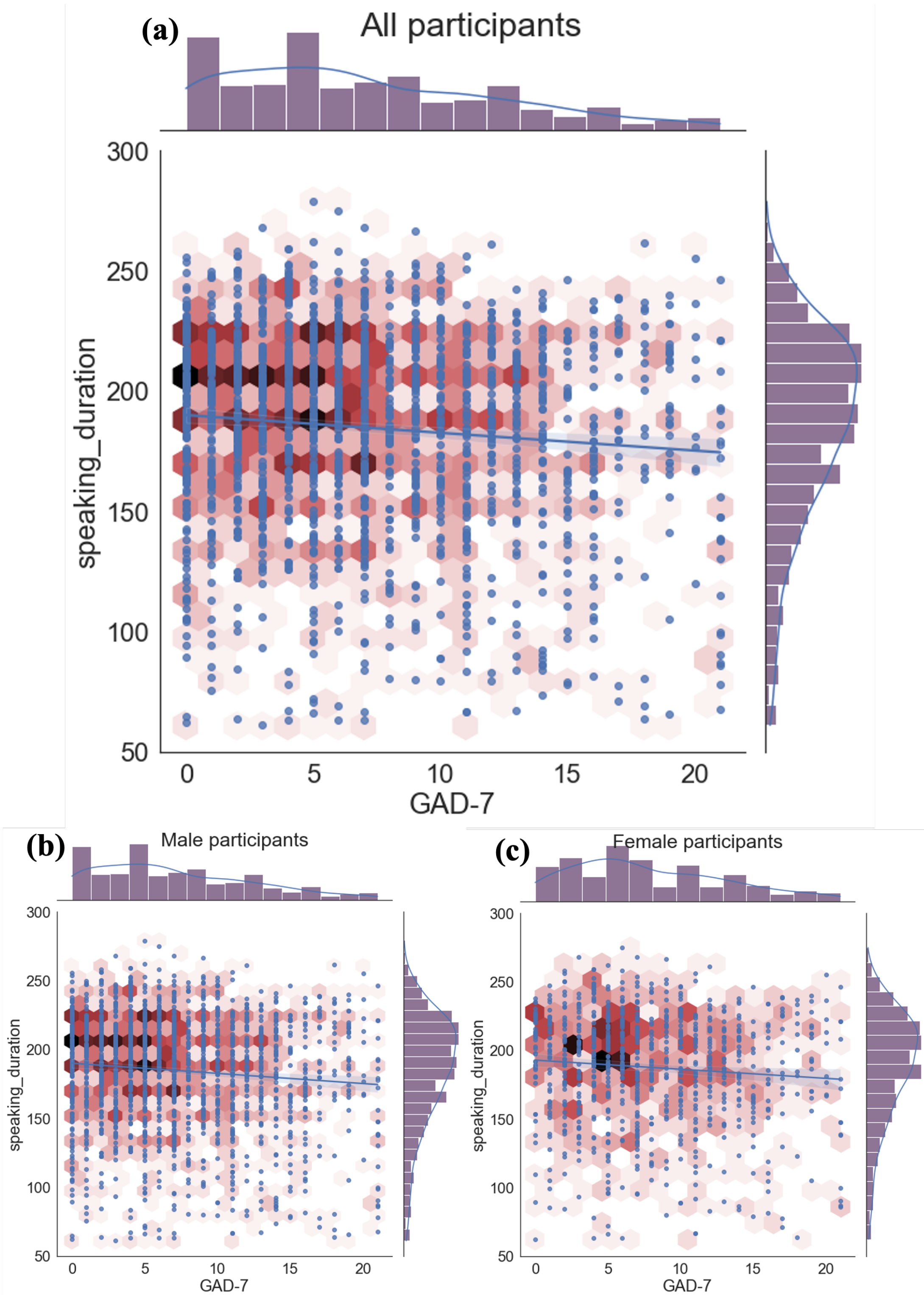


Figure 4.1: Speaking duration vs. GAD-7 scatter plot and distributions

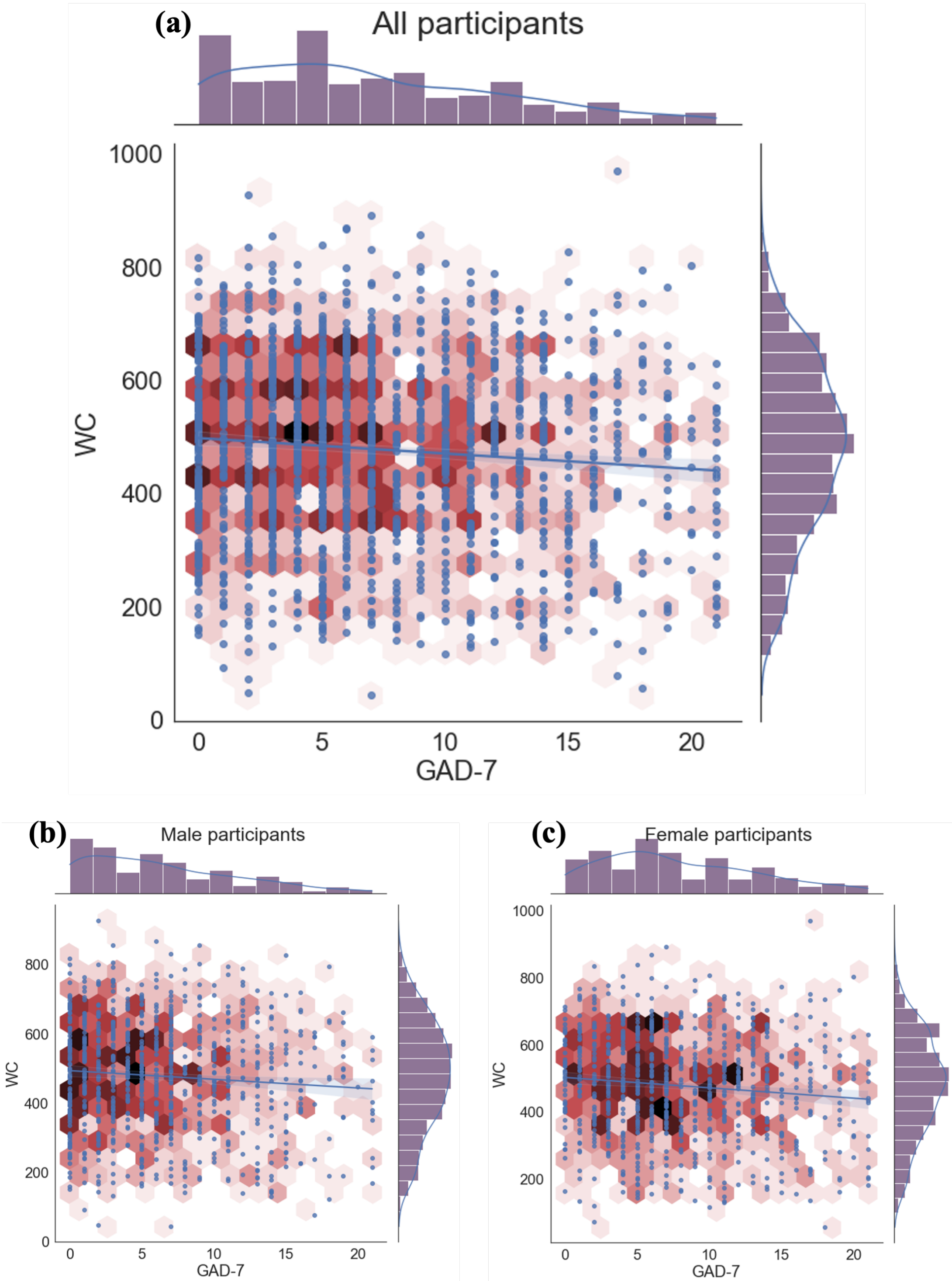


Figure 4.2: Word count vs. GAD-7 scatter plot and distributions

4.3.2 Acoustic Feature Correlation with GAD-7

Table 4.2 presents the correlation and P-values for all of the acoustic features (presented in section Acoustic Features) that had P-values above the 95% confidence level for the three data sets – all participants, and then female-only and male-only participants. Again, note that all correlations were computed after controlling for age and personal income, while the calculations involving all participants also controlled for sex. Table 4.3 reports results for features that previous work found to be statistically significant, but for which we find no correlation in our sample. In our results, these features were not significantly associated with anxiety in any of the three data sets: all, female, and male.

Table 4.4 makes a direct comparison between previous work on the specific features (and their relation to anxiety) and results from the present study.

a) All Participants (N=1744)

Feature	**r**	***P***
Shimmer	0.08	< .001
mfcc_std_2	-0.08	0.002
mfcc_std_3	-0.07	0.002
mfcc_mean_2	-0.07	0.004
f0_std	0.06	0.01
mfcc_std_5	-0.06	0.01
mfcc_std_4	-0.05	0.03

b) Female Participants (N=862)

Feature	**r**	***P***
mfcc_std_3	-0.10	0.002
Shimmer	0.10	0.004
lpcc_std_6	-0.09	0.008
lpcc_std_4	-0.09	0.008
mfcc_mean_2	-0.09	0.01
intensity_mean	-0.09	0.01
mfcc_mean_1	-0.09	0.01
lpcc_std_10	-0.07	0.03
intensity_std	-0.07	0.03
lpcc_std_12	-0.07	0.04
mfcc_mean_8	0.07	0.04
lpcc_mean_4	0.07	0.049

c) Male Participants (N=882)

Feature	**r**	***P***
mfcc_std_2	-0.09	0.005
mfcc_std_5	-0.09	0.01
mfcc_mean_5	-0.08	0.01
f0_std	0.07	0.03
mfcc_std_4	-0.07	0.04
Shimmer	0.07	0.04
mfcc_std_11	-0.07	0.046
f1_mean	0.07	0.047

Table 4.2: Correlation of significant acoustic features with GAD-7

4.3.3 Linguistic Feature Correlation with GAD-7

The quality of the transcript produced using the speech-to-text (STT) system from Amazon [93] was analysed by comparing the transcript produced from Task 1 audio to the actual Grandfather Passage. The Word Error Rate of the Grandfather passage, averaged across all transcripts is 7% (SD 4.6%).

Table 4.5 presents the set of linguistic features (described in Section 4.2.1) that had P-values lower than .05 for the three data sets – all participants, male-only and female-only. Each table is sorted in decreasing order of absolute value of correlation. As before, the partial correlations are controlled for age and personal income across

Feature	Previous works	Current study					
		All samples		Female samples		Male samples	
		r	P	r	P	r	P
Jitter	Showed a significant increase from a neutral state to anxious state [54]	0.03	0.18	-0.01	0.76	0.06	0.06
ZCR_zPSD	ZCR_zPSD was one of the top selected features using Davies-Bouldin Index based feature selection [60]	0.01	0.67	-0.04	0.29	0.05	0.14
Articulation rate	Patients with panic disorder spoke significantly slower (P < .001) during autobiographical talking compared to a script talking [88]	-0.01	0.64	-0.05	0.12	0.02	0.55
F1 std	Showed a significant change between a neutral state and anxious state [54]	-0.03	0.18	-0.02	0.53	-0.04	0.25
F2 mean		0.004	0.85	0.04	0.26	-0.04	0.22
F2 std		0.01	0.59	0.03	0.38	-0.02	0.60
F3 mean		0.02	0.49	0.04	0.21	-0.01	0.72

Table 4.3: Correlation of acoustic features not found significant

Feature	Previous work				Current study					
					All samples		Female samples		Male samples	
	r	***P***	**Ref**	**N**	r	P	r	P	r	P
Speaking duration	-0.36	<.01	[86]	71	-0.12	5E-07	-0.11	0.0007	-0.13	0.0001
MFCC_std_1	-0.36	<.05	[84]	45	0.01	0.54	0.02	0.61	0.02	0.52
F0_mean	Female:0.02; Male:0.72	Female: 0.92; Male:0.002	[56]	46	0.02	0.37	-0.03	0.33	0.06	0.06
F0_std	-0.24	<.05	[86]	71	0.06	0.01	0.03	0.30	0.07	0.03
Intensity mean	-0.2	-	[86]	71	-0.04	0.13	-0.09	0.01	0.01	0.72

Table 4.4: Comparison of previous work correlations with present study

all samples, and we also control for sex in the full data set.

4.4 Discussion

Our central objective is to test specific acoustic and linguistic features of impromptu speech for their association with anxiety and to do so with a larger number of participants than in prior work. In this section, we discuss the implications of the findings presented in the previous section, as well as the limitations of the study.

4.4.1 Principal Findings

The results presented above quantified the relationship between features computed from recorded speech and the self-reported GAD-7 scale, using Pearson correlation coefficients, controlling for age and income. There are several significant correlations between features extracted from speech and anxiety, which can help to inform future efforts in the automatic monitoring of anxiety. We discuss these below.

a) All Samples (N=1744)

Feature	**r**	***P***
AllPunc	0.13	<.001
Period	0.12	<.001
assent	0.10	<.001
negemo	0.10	<.001
relativ	-0.09	<.001
motion	-0.08	<.001
swear	0.08	<.001
anger	0.08	<.001
focusfuture	-0.07	.003
adverb	-0.07	.004
time	-0.07	.004
function	-0.07	.005
negate	0.07	.006
prep	-0.06	.007
WPS	-0.06	.007
anx	0.06	.008
hear	0.06	0.01
death	0.06	0.01
ipron	-0.06	0.01
see	-0.06	0.01
affect	0.06	0.02
i	0.05	0.02
family	0.05	0.02
sad	0.05	0.03
ppron	0.05	0.03
space	-0.05	0.04
article	-0.05	0.04
leisure	0.05	0.04
friend	0.05	0.047

b) Female Samples (N=862)

Feature	**r**	***P***
Period	0.16	<.001
AllPunc	0.14	<.001
adverb	-0.11	<.001
negemo	0.11	<.001
anger	0.11	.002
motion	-0.10	.003
assent	0.10	.004
see	-0.09	.006
relativ	-0.09	.006
sad	0.08	.01
Dic	-0.08	.02
power	0.07	.03
WPS	-0.07	.03
death	0.07	.04
percept	-0.07	.046

c) Male Samples (N=882)

Feature	**r**	**P**
AllPunc	0.13	<.001
assent	0.11	.001
relativ	-0.10	.002
leisure	0.10	.002
hear	0.10	.003
swear	0.10	.004
time	-0.10	.004
Apostro	0.09	.005
power	-0.09	.01
ppron	0.09	.01
Sixltr	-0.09	.01
anx	0.08	.01
negate	0.08	.01
negemo	0.08	.01
article	-0.08	.01
Period	0.08	.02
prep	-0.08	0.02
focusfuture	-0.08	0.02
family	0.08	0.02
ipron	-0.07	0.04
affect	0.07	0.04
motion	-0.07	0.048

Table 4.5: Correlation of significant LIWC linguistic features with the GAD-7

Amount of Speech

The amount of speech that the participants delivered in response to Task 2 had one of the highest correlations with the GAD-7 scale across all the features explored in this work. The Speaking Duration and Word Count features, as shown in Table 4.1 (and their inter-correlation with each other is shown in Multimedia Appendix J). In all cases the negative direction of the correlation suggests that participants who spoke more, tend to have lower GAD-7 scores. This result is consistent with previous work, as shown in the first data row of Table 4.4), but the present result has a lower Pearson's correlation than prior work ($r = -0.12; P < .001$ in this study, vs. $r = -0.36; P < .01$ for [86]). The present result does achieve a higher significance level than the prior work. We speculate that the more anxious a person is, the less confidence they would have about their speech, and so perhaps they speak less.

Acoustic Features

The main purpose of this work was to explore how acoustic features relate to anxiety. We wanted to determine if associations found in previous studies still hold with the larger sample size. Table 4.2 lists the features that have significant correlations, with $P < .05$, across all three data sets. The features with the strongest correlation in this set were shimmer on the all-samples data set and the standard deviation of the 2nd and 3rd MFCCs for the male and female samples, respectively. We note that there are multiple parameters used in the extraction of MFCC features, so a direct comparison of the specific MFCC features of our study with specific features of previous work is not possible since the prior work does not specify the exact parameters used to compute the MFCCs. The parameters used in the present study are given in Section 4.2.1. In prior research, the 4th MFCC was the most significant from the 13 MFCC features in [54] and the standard deviation of the 1st MFCC in [84] had a significant correlation ($r = -0.36$; $P < .05$) with an anxiety scale. These results, from both the present study and previous work, suggests that signals of anxiety are present in the MFCC features.

The following features, suggested as relevant to anxiety in prior work, did not show significant correlations with GAD-7 in the present work: the second and third formant; jitter; zero-crossing rate of the z-score of the power spectral density; and the articulation rate. Table 4.3 shows prior work associations with anxiety on these features and the correlation values obtained in the present study. It is important to note that, in previous research, these features were noted as significant or relevant, but no correlations with an indicator of anxiety were given. This makes it difficult to compare directly with the correlations obtained in our study.

Linguistic Features

Correlations between linguistic features extracted using the LIWC dictionaries [28] and the GAD-7 were presented in Section 4.3. These had a higher correlation than the acoustic features, as shown in Table 4.5. The top LIWC category with the highest correlation in all the data sets is the count of punctuation. This includes the count of periods, which would indicate the number of separate sentences. The positive correlation of count of periods and a negative correlation of words per sentence (WPS) indicates that the use of shorter sentences is positively associated with anxiety.

The other LIWC categories with high correlation in the all-sample data set are Negative emotion ("negemo"; e.g., hurt, ugly, nasty), Anger ("anger"; e.g., hate, kill, annoyed), Anxiety ("anx"; e.g., worried, fearful), and Sad ("sad"; e.g., crying, grief, sad). The Anger, Anxiety, and Sad categories are constituent subsets of the negative emotion ("negemo") category – i.e., words counted under one of the Anger, Anxiety, or Sad categories will also be counted for the negemo category. The high inter-correlation of these features is given in a table in Multimedia Appendix J. The negemo count had a higher correlation than these individual sub-categories, suggesting that words relating to anger, anxiety and sad are capturing different dimensions of self-reported anxiety.

An LIWC category with a significant correlation that is present in the male samples but not in the female samples is the use of apostrophes (Apostro), indicating words with contractions (such as I'll) were positively associated with GAD-7. Also, only for males, function words, including personal pronouns (ppron) had a significant positive correlation with anxiety. We speculate that anxious male individuals might use personal pronouns (which includes I, me, mine) to divert their attention from the anxiety-inducing event and focus on themselves. More generally, the increased use of personal pronouns has been shown to occur in individuals with depression [97], a highly comorbid mental health illness with GAD (but not only for males).

Another differentiation between males and females occur in the LIWC feature for words related to "power" (e.g., superior, bully). The "power" count had a positive correlation with GAD-7 for females and a negative correlation for males. We speculate that the negative correlation is somehow related to the stereotypical dominance behaviour associated with males.

In prior work studying associations between LIWC scores and anxiety, words related to anxiety and first-person singular pronouns were shown to be significantly associated with social anxiety [57], similar to our results. The same work [57] has also shown that perceptual process words (see, hear, feel) are significantly associated with anxiety, which does not align with our results. In our results, the LIWC category for see has a negative correlation in the all the sample ($r = -0.6; P = .01$) and the female ($r = -0.9; P = .006$) samples (as shown in Table 4.5). However, in [9], the category see had a positive correlation ($r = 0.31; P = .02$) with a social anxiety measure. We speculate that the use of perceptual process words (See) might be a differentiating factor between social anxiety and generalized anxiety disorder since it was positively correlated with social anxiety and negatively correlated with generalized anxiety.

The LIWC category for the perceptual process "hear", on the other hand, had a positive correlation in both the all-sample and the male samples (also shown in Table 4.5). Notice that both see and hear are perceptual processes, but the category for see is significant for females while the category for hear is significant for males.

Also, in prior work, death-related words were shown to have positive correlation with anxiety [9]. Our results (as shown in Table 4.5) show a similar trend where death-related words had a significant positive correlation in the female and all-sample data set. However, a significant correlation was not observed in the male samples.

The fact that there are several single-word categories that have significant correlations suggests that techniques that are able to look at multiple word meanings may have greater potential in making predictions. We explore that concept in Chapter 6.

4.4.2 Conclusion

We have presented results from a large-N=1,744 study examining the relationship between speech and generalized anxiety disorder. Our data collection relies on participants using recording devices in their home, which means that variations in acoustic environments are present, compared with more pristine common environments. Our goal was to provide a useful benchmark for future research, by assessing the extent to which results from previous research generalize to a larger population and in-the-wild collection methodology. We measured the most commonly suggested acoustic and linguistic features' association with anxiety and provided detailed correlation tables broken down by demographics.

Our findings are decidedly mixed. On the one hand, with our larger dataset, we find modest correlations between anxiety and several features of speech, including speaking duration and acoustic features such as MFCCs, LPCCS, Shimmer, Fundamental Frequency and first formant. However, other features shown in prior work to correlate with anxiety — were not significantly associated with anxiety in our study - including second and third formant, Jitter, and ZCR-zPSD. Although these null findings do not entirely rule out the potential of more sophisticated learning models for this task, we believe that researchers should be wary of inherent difficulties. Readers should also note that our data collection already sidesteps additional challenges that we expect to influence the detection of anxiety disorders from speech, such as variations in accents, dialects and spoken language. On the other hand, we did find statistically significant correlations for a subset of speech features from previous research. This suggests that there may be a fundamental pathway between anxiety and the production of speech, one that is robust enough to be used to make a prediction model. One of those models is described in the next Chapter.

Chapter 5

Screening for GAD from Acoustic and Linguistic Speech Features

In Chapter 4, acoustic and linguistic features of speech that have a significant correlation with GAD-7 were identified. In this Chapter, the performance of a model that predicts the presence of Generalized Anxiety Disorder (GAD) from acoustic and linguistic features of impromptu speech will be evaluated. This work was published in [98].

5.1 Overview

Prior work (described in Section 2.4) suggests that it is possible to detect anxiety disorders from speech. However, the largest sample size among these previous studies was N=239 participants, with an average number of participants of N=115, which limits the generalizability of the results. In addition, the number of participants might not be the only factor for generalizability. Aside from Di Matteo et al. [10] and Baird et al. [64], the prior studies explored are mostly limited to very specific demographics - McGinnis et al. [60] focused on children; Weeks et al. [56], Salekin et al. [63], and Rook et al. [65] focused on undergraduate students at a University or College.

We hypothesize that by recruiting a substantially larger cohort (N = 2,000) from a broader demographic than prior work, it is possible to achieve above random prediction accuracy in screening for GAD using acoustic and linguistic features that have been previously suggested.

The data used to create the model (described in Chapter 3) is the same data used in Chapter 4 to measure the correlation of acoustic and linguistic features with the GAD-7. Recall that the data came from 2,000 participants who were recruited and completed the Generalized Anxiety Disorder-7 item scale (GAD-7) and provided an impromptu speech sample in response to a modified version of the Trier Social Stress Test (TSST). In this chapter we make use of the linguistic and acoustic features that were found in Chapter 4 to be associated with anxiety disorders, along with demographic information. We build a model that predicts if participants fall above or below the screening threshold for GAD based on the GAD-7 scale threshold of 10. Separate models for each sex are also evaluated. We report the mean area under the receiver operating characteristic (AUROC) from a repeated 5-fold cross-validation to evaluate the performance of the models.

5.2 Methods

5.2.1 Inputs to the Classification Model

The inputs to our models are acoustic and linguistic features that were determined in a previous Chapter (Chapter 4) to have a statistically significant correlation with the GAD-7. These features were found to be correlated with the GAD-7 after controlling for demographic variables – age, sex, and personal income. These features are presented in Table 5.1 in three sub-tables: one for the all-sample, one for the female samples, and one for the male samples.

Aside from the acoustic and linguistic features, we also explored the use of demographic information such as age, sex, and personal income as input features to the model. We decided to use these demographics as features in the model since they were made available to us from the Prolific recruitment platform [67]. Should the model be used as a diagnostic screener in the future, it should also be possible to obtain these demographics, and so we feel it is reasonable to make use of them.

5.2.2 Construction and Evaluation of Classification Models

The specific goal in this chapter is to evaluate the performance of a binary classifier that predicts if a person's speech sample is in the 'anxious' or 'nonanxious' class based on features of speech. The binary classification label is determined by processing the GAD-7 scale (which ranges from 0-21 in value) into two classes of anxious (GAD-7 $\geq$ 10) and nonanxious (GAD-7 $<$ 10), where 10 is a well-established screening threshold [18] for GAD.

A logistic regression model is trained to make predictions between the two classes. The data for the model training and evaluation were processed in the following steps:

1. The input features were normalized so each feature would have a mean of zero and a standard deviation of 1.

a) All Samples (N=1744)

Feature	r	P
AllPunc	0.13	<.001
WC	-0.12	<.001
Speaking duration	-0.12	<.001
Period	0.12	<.001
assent	0.10	<.001
negemo	0.10	<.001
relativ	-0.09	<.001
motion	-0.08	<.001
Shimmer	0.08	<.001
swear	0.08	<.001
anger	0.08	<.001
mfcc_std_2	-0.08	.002
mfcc_std_3	-0.07	.002
focusfuture	-0.07	.003
mfcc_mean_2	-0.07	.004
adverb	-0.07	.004
time	-0.07	.004
function	-0.07	.005
negate	0.07	.006
prep	-0.06	.007
WPS	-0.06	.007
anx	0.06	.008
f0_std	0.06	.01
hear	0.06	.01
mfcc_std_5	-0.06	.01
death	0.06	.01
ipron	-0.06	.01
see	-0.06	.01
affect	0.06	.02
i	0.05	.02
family	0.05	.02
mfcc_std_4	-0.05	.03
sad	0.05	.03
ppron	0.05	.03
space	-0.05	.04
article	-0.05	.04
leisure	0.05	.04
friend	0.05	.047

b) Female Samples (N=862)

Feature	r	P
Period	0.16	<.001
AllPunc	0.14	<.001
WC	-0.13	<.001
speaking_duration	-0.11	<.001
adverb	-0.11	<.001
negemo	0.11	<.001
anger	0.11	.002
mfcc_std_3	-0.10	.002
motion	-0.10	.003
Shimmer	0.10	.004
assent	0.10	.004
see	-0.09	.006
relativ	-0.09	.006
lpcc_std_6	-0.09	.008
lpcc_std_4	-0.09	.008
mfcc_mean_2	-0.09	.01
intensity_mean	-0.09	.01
mfcc_mean_1	-0.09	.01
sad	0.08	.01
Dic	-0.08	.02
power	0.07	.03
WPS	-0.07	.03
lpcc_std_10	-0.07	.03
intensity_std	-0.07	.03
lpcc_std_12	-0.07	.04
death	0.07	.04
mfcc_mean_8	0.07	.04
percept	-0.07	.046
lpcc_mean_4	0.07	.049

c) Male Samples (N=882)

Feature	r	P
speaking_duration	-0.13	<.001
AllPunc	0.13	<.001
WC	-0.12	<.001
assent	0.11	.001
relativ	-0.10	.002
leisure	0.10	.002
hear	0.10	.003
swear	0.10	.004
time	-0.10	.004
Apostro	0.09	.005
mfcc_std_2	-0.09	.005
power	-0.09	.01
ppron	0.09	.01
Sixltr	-0.09	.01
mfcc_std_5	-0.09	.01
anx	0.08	.01
negate	0.08	.01
negemo	0.08	.01
article	-0.08	.01
mfcc_mean_5	-0.08	.02
Period	0.08	.02
prep	-0.08	.02
focusfuture	-0.08	.02
family	0.08	.02
f0_std	0.07	.04
mfcc_std_4	-0.07	.04
ipron	-0.07	.04
Shimmer	0.07	.04
affect	0.07	.04
mfcc_std_11	-0.07	.045
f1_mean	0.07	.047
motion	-0.07	.048

Table 5.1: Correlation of statistically significant acoustic and linguistic features with GAD-7

2. The data was under-sampled to equalize representation from both the anxious and nonanxious classes. This avoids the problem of class imbalance, which, if it happens, causes low predictive accuracy for the minority class (which is the anxious class in our case). Specifically, samples were randomly selected and removed from the majority class until the majority class has an equal number of samples as the minority class.

3. The data is notionally split into three separate sets, but we will also employ a higher-level cross validation approach described below. So there is a training dataset which is used to train the model parameters; a validation dataset which is used to select the best hyperparameters during training; and a test dataset which is used to evaluate the performance of the trained model. Here we will use the area under the receiver operating characteristic (AUROC) metric. The datasets are created within each sampling of the cross-validation scheme described below.

4. The cross-validation (CV) scheme employed a nested resampling with two-level nested cross-validations– one CV nested within another [99]. In the outer loop, the data is split into 20% test data and 80% training and validation data. In the inner loop the 80% training and validation data is further split into 20% validation data and 80% training data. The inner loop is repeated 5 times, each with different sampling to get a different 20/80 split. For each such split, the best hyperparameters are selected that maximize the accuracy on the validation data after training on the inner loop's training data. After selecting the best hyperparameters from the inner CV loop, training is once again done on the whole 80% of the outer loop training plus validation data and mean area under the receiver operating characteristic (AUROC) results are reported on the outer loop's test data. The outer loop will iterate five times, each time selecting a different 20% for test data, until all the samples have been left out and tested on. This whole process will be repeated 7 times each with different random under-sampling seed where in each of the 7 iterations, 5 AUROC are reported from the outer CV loop giving a total of 35 values. The mean and standard deviation of the AUROC values are used as the final metrics in this study to measure performance.

5.2.3 Feature Selection

In the construction of the model, a subset of features were selected from those listed in Table 5.1. The goal of feature selection is to avoid using duplicate information (where the same information is present in different features) and to maximize the prediction performance of our model.

To gain an initial sense of which features were closely related information, the correlation between the features presented in Table 5.1 were computed. We then used only one of each pair of the highly correlated (r > 0.8) features.

In model construction, it might not always be true that using all the available features maximizes the prediction accuracy; doing so may actually reduce accuracy due to overfitting [29]. Thus, to maximize the prediction performance of our model, we selected a subset of features using the following method: we began with the single feature that had the highest correlation with the GAD-7 and then measured the prediction performance of a trained model (on the validation data) using only that feature. Subsequent features are then added one-by-one in order of correlation, until all the features listed in Table 5.1 are either used or discarded. This method of feature selection is known as Sequential Forward Selection (SFS) [100].

5.2.4 Statistical Analysis

We compute the prediction performance of the model as the mean Area Under the Receiver Operating Characteristic curve (AUROC) of the 35 models produced by the five-fold cross-validation process described in Section 5.2.2. It will be compared to the mean AUROC of a model that makes a random prediction which would typically have a mean AUROC of 0.5, using a modified t-test developed by [101]. The modified t-test considers the fact that the individual AUROC values are not independent from each other, while in the original t-test, the samples are expected to be independent. In our case, since the AUROC generated from one model shares some training data (due to the multiple under-sampling and the 5-fold cross-validation) with another, the AUROC are not independent of each other. In our results, we have considered a statistically significant difference at a value significance level of .05.

5.3 Results

5.3.1 Classification Model Performance

In this section, the mean AUROC of a binary classification model that classifies between the anxious and nonanxious classes is presented. The subsections below summarize our main empirical results for different types of inputs to the classification models.

a) All Samples

Similar features	**r**	***P***
AllPunc & Period	0.93	<.001
i & ppron	0.81	<.001

b) Female Samples

Similar features	**r**	***P***
AllPunc & Period	0.93	<.001
Intensity_mean & intensity_std	0.93	<.001

c) Male Samples

Similar features	**r**	***P***
AllPunc & Period	0.93	<.001

Table 5.2: Features with high inter-correlation (similar features) with each other

a) All Samples

Feature	**r**	***P***
AllPunc	0.13	<.001
assent	0.10	<.001
relativ	-0.09	<.001
motion	-0.08	<.001
mfcc_std_2	-0.08	.002
mfcc_std_3	-0.07	.002
focusfuture	-0.07	.003
mfcc_std_5	-0.06	.01
death	0.06	.01
see	-0.06	.01
mfcc_std_4	-0.05	.045

b) Female Samples

Feature	**r**	***P***
Period	0.16	<.001
motion	-0.10	.003
see	-0.09	.006
Dic	-0.08	.02
power	0.07	.03
lpcc_std_10	-0.07	.03
death	0.07	.04

c) Male Samples

Feature	**r**	***P***
speaking_duration	-0.13	<.001
assent	0.11	.001
relativ	-0.10	.002
leisure	0.10	.002
time	-0.10	.004
ppron	0.09	.01
negemo	0.08	.01
article	-0.08	.01
mfcc_mean_5	-0.08	.01
focusfuture	-0.08	0.02
f1_mean	0.07	.047

Table 5.3: Subset of acoustic and linguistic features used after feature selection

Acoustic and Linguistic Feature Selection

The model is constructed using a subset of the acoustic and linguistic features as described above in Section 5.2.3. Table 5.2 presents the pairs of features with high correlation ($r > 0.8$). Only one feature of each pair was selected.

Section 5.2.3 describes the method of sequentially selecting the acoustic and linguistic features for use in the model, beginning with the feature with the highest correlation and incrementally adding the next-most correlated until no improvement is found. Table 5.3 shows the subset of features used for the three different samples of the data. Figure 5.1 gives the mean AUROC as a function of the number of selected features for each of the three datasets. The number of features that achieved the highest mean AUROC was 11 for the combined male-female sample, seven for the female samples and eleven for the male samples.

Table 5.4 gives the mean AUROC across the 35 data splits, as described in the methods section. It also gives the result of the t-test comparison of the best model with that of a random model.

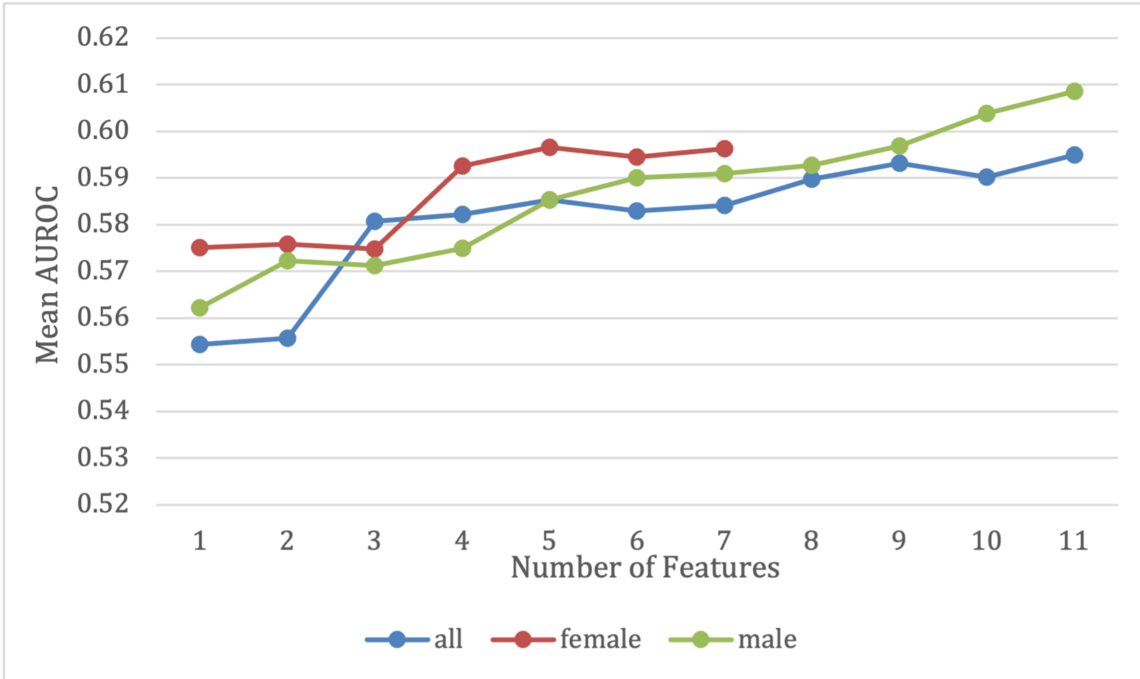


Figure 5.1: Mean AUROC as a function of the number of selected features

Dataset	Mean AUROC (SD)	t-test (df)	*P*
All samples (N=1744)	0.59 (0.02)	4.93 (34)	<.001
Female samples (N=862)	0.60 (0.04)	4.25 (34)	<.001
Male samples (N=882)	0.61 (0.06)	4.21 (34)	<.001

Table 5.4: Mean AUROC of a model trained using the subset of features

Using Participants' Demographics

This section presents the performance of the model when it is augmented with age, sex, and income demographic information. Table 5.5 gives the mean AUROC of a logistic regression model that uses both the demographic information and the acoustic and linguistic features. It also includes a modified t-test comparison with a random model. The AUROC curves for the model built on the all-samples that predicts GAD from acoustic features, linguistic features, and demographics are presented in Appendix M. One of the AUROC curves is shown in Figure 5.2

Table 5.6 shows the result of a t-test between the model with only acoustic/linguistic features and the model that also uses the demographic information. Table 5.7 separates out each of the demographic features and shows the mean AUROC of these models when using a single demographic at a time, together with the acoustic and

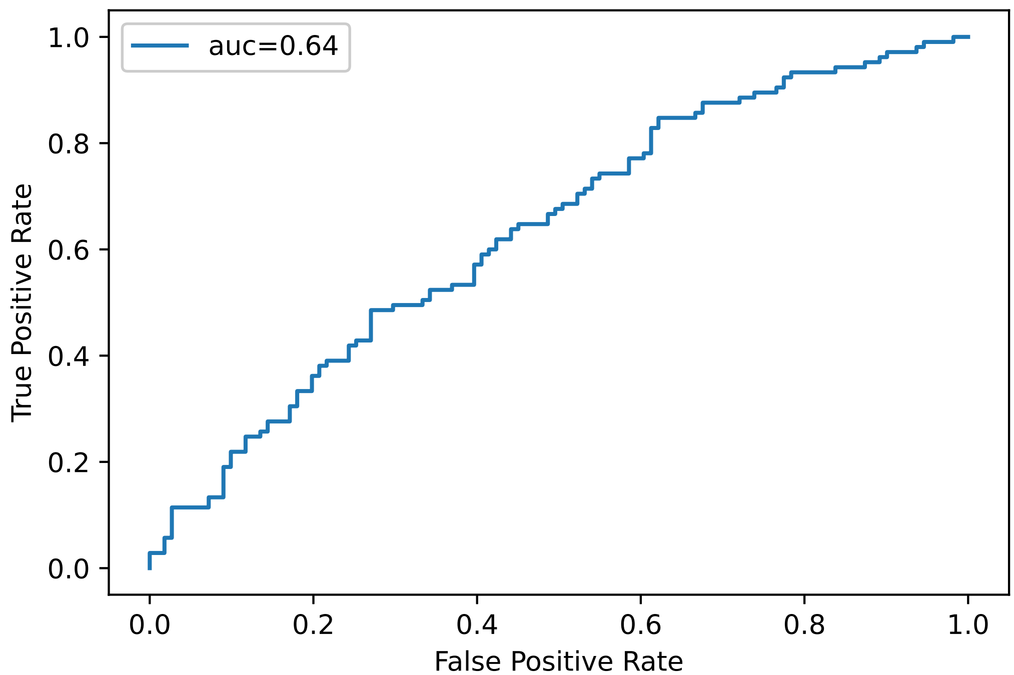


Figure 5.2: AUROC curve for the model built on the all-samples that predicts GAD from acoustic features, linguistic features, and demographics

Dataset	**Mean AUROC (SD)**	**t-test (df)**	***P***
All samples (N=1744)	0.64 (0.03)	6.21 (34)	<.001
Female samples (N=862)	0.66 (0.04)	5.89 (34)	<.001
Male samples (N=882)	0.62 (0.07)	4.36 (34)	<.001

Table 5.5: Mean AUROC of a model trained using demographic info (age, sex, income) in addition to the acoustic and linguistic features and comparison with a random model

linguistic features.

5.4 Discussion

The main objective of this work is to investigate the prediction performance of a model that screen for GAD from acoustic and linguistic features of impromptu speech. To do so, we have explored a logistic regression model, and in the following subsections, we discuss the findings presented in the Results section, as well as the limitations of this study.

Dataset	Mean AUROC (SD)		t-test (df)	*P*
	Acoustic and Linguistic features	Demographics, Acoustic and Linguistic features		
All samples (N=1744)	0.59 (0.02)	0.64 (0.03)	4.01 (34)	<.001
Female samples (N=862)	0.60 (0.04)	0.66 (0.04)	4.21 (34)	<.001
Male samples (N=882)	0.61 (0.06)	0.62 (0.07)	0.76 (34)	.45

Table 5.6: Comparison of a model trained using only acoustic/linguistic features with a model that also uses demographic information

Dataset	Mean AUROC (SD) of Model Using Acoustic and Linguistic Features AND Only		
	Age	Income	Sex
All samples (N=1744)	0.64 (0.03)	0.60 (0.02)	0.59 (0.02)
Female samples (N=862)	0.66 (0.04)	0.60 (0.05)	-
Male samples (N=882)	0.62 (0.07)	0.61 (0.06)	-

Table 5.7: Mean AUROC of the model when adding a single demographic to the acoustic and linguistic features

5.4.1 Acoustic and Linguistic Features as input

The logistic regression model uses the statistically significant acoustic and linguistic features selected by the feature selection method discussed in Section 5.2.3 and presented in Table 5.3. Although the correlation between the features used and the GAD-7 is very small (the highest being 0.13), the model built using these features was able to perform significantly better (with $P < .05$) than a random model. The mean AUROC results presented in Table 5.4 suggests that there is some signal to be detected from the combined effect of the acoustic and linguistic features of speech.

From these acoustic and linguistic signals, an automated screener that listens passively to speech has the potential to frequently monitor speech samples from the patient. Furthermore, the probability of correct prediction could be improved by using multiple measurements, under the assumption that each measurement, from different speech is relatively independent. This improvements we get from multiple measurements will be discussed in detail in Chapter 7.

5.4.2 Using Participant's Demographics

In a scenario where an anxiety screening/prediction model could be deployed, and an individual's demographic information can easily be collected, it makes sense to explore the predictive capability of this additional information. Table 5.5 shows that a model built using the demographic information as input in addition to the acoustic and linguistic features was still able to perform significantly better than a random model. Table 5.6 compares this model with a model that only has the acoustic/linguistic features as input. The result of the t-test show that the demographic information improved the mean AUROC of the models built on the all-sample and female samples but did not improve the model built on the male samples.

The impact of each of the demographics was also separately explored. Table 5.7 shows the mean AUROC of the model when each one of the demographics was used separately to augment the acoustic and linguistic features. It shows that the addition of age improved the prediction performance of the model, while the addition of either sex or income did not show a significant improvement. Also, the fact that the addition of age impacted the prediction performance for the model built on the female participants suggests that anxiety in females depends on age compared to anxiety in males.

Comparing our result with prior studies that worked on predicting GAD, there were studies that achieved above random prediction for GAD [64][65], and there were studies that did not [10]. Our models did perform significantly better than a random model, and we speculate that this might be attributed to the larger and demographically broad sample size that enabled our model to learn a large amount of information in predicting GAD. We also note that the one prior study that did not succeed at predicting GAD [10] did, in fact, succeed in predicting SAD. They believe that the symptoms of SAD might be more manifested in the participants' behaviour and, therefore, speech compared to GAD. Other studies focusing on SAD also had success in above random prediction [56][63].

5.4.3 Conclusion

In this study, we developed a model to predict the presence or absence of GAD based on speech. Our results show that it is possible to achieve the above random prediction accuracy from acoustic and linguistic features of speech. The prediction accuracy can also be further improved when adding basic demographic information. Even though we have investigated adding three different types of demographics (age, sex, and income), the most influential one or the demographic that showed improvement in prediction accuracy was age. So, the acoustic and linguistic features of speech together with basic demographic info may be used in a system to trigger early intervention, monitor treatment response, or detect relapses.

Chapter 6

Anxiety Prediction from Linguistic Patterns of Speech

In Chapter 5, we presented and discussed the prediction performance of a model that predicts GAD from acoustic and LIWC features. The LIWC features used for the prediction are features that are based on the count of individual words and are not aware of the context of the individual words. For example, if the word "happy" is seen in a transcript, it will be counted as a positive sentiment word whether or not it follows the word "not." In this chapter, we explore if a context-aware model would have a better prediction performance compared to a model that is based on the LIWC features. This work is currently published as a preprint [102] and is under peer review.

6.1 Overview

The previous chapter and most prior work on anxiety detection explored the count of single words (using the LIWC) to find an association with anxiety or to predict anxiety. However, there are some studies which found specific word categories to be associated with anxiety, while others found no such association. For example, both [9] and [57] found that the word categories for 'perceptual process' were associated with anxiety, whereas no other prior studies did so. Similarly, the first-person singular pronouns category was associated with anxiety in [57] but nowhere else. These inconsistencies may be explained if the context for the specific words were taken into account. In this chapter, we hypothesize that there is a greater predictive power in looking at the larger context of multiple words compared to the LIWC single-word-based method. This can be done using recent advances in Natural Language Processing [39], which has newly powerful methods of converting language into numerical quantities that represent meaning and learning features that are patterns of those meanings.

Transformer-based neural networks [39][103][40][41][42] are models that have been recently shown to have powerful predictive capabilities, based on multiple input words. Transformers detect linguistic patterns and can be separately trained to make specific predictions based on those patterns. The objective of this Chapter is to determine if a transformer-based language model can be used to screen for Generalized Anxiety Disorder from impromptu speech transcripts.

The dataset is the same one used in previous Chapters, but we only make use of the speech transcripts and not the audio or the acoustic features. Using the transcripts from each participant, a transformer-based neural-network model (pre-trained on large textual corpora) was fine-tuned on the speech transcripts and the GAD-7 to predict above or below a screening threshold of the GAD-7. We report the area under the receiver operating characteristic (AUROC) on test data and compare the results with a baseline logistic regression model using the Linguistic Inquiry and Word Count (LIWC) features as input. We will also use an Integrated Gradient [104] method to determine specific sequences of words that strongly affect the predictions and infer specific linguistic patterns that influence the predictions.

6.2 Methods

6.2.1 Construction and Evaluation of Baseline Classification Model

In this section, the inputs, structure, and the evaluation of a baseline model will be described. The inputs to this model are the linguistic features acquired using Linguistic Inquiry and Word Count (LIWC). As discussed in earlier chapters, LIWC is based on the count of words from a given transcript into different pre-set categories. The full set of categories for LIWC can be found in [28].

A transcript of the audio recordings was produced using a commercial speech-to-text system from Amazon Web Services [93] - the transcription accuracy on a written text had an average word error rate (WER) of 7% (SD 4.6%). In Chapter 4, we identified LIWC features that had a significant ($P < .05$) correlation with the GAD-7. These features are listed in Table 6.1. It is these features that are used as the input to the baseline prediction model.

A logistic regression model was trained to make predictions between the two classes. The construction and evaluation steps were as follows: First, the input features were normalized so each feature would have a mean of zero and a standard deviation of 1. Next, the data was under-sampled to equalize representation

Feature	r	P	Feature	r	P
AllPunc	0.13	<.001	WPS	-0.06	.007
Word Count	-0.12	<.001	anx	0.06	.008
Period	0.12	<.001	hear	0.06	0.01
assent	0.10	<.001	death	0.06	0.01
negemo	0.10	<.001	ipron	-0.06	0.01
relativ	-0.09	<.001	see	-0.06	0.01
motion	-0.08	<.001	affect	0.06	0.02
swear	0.08	<.001	i	0.05	0.02
anger	0.08	<.001	family	0.05	0.02
focusfuture	-0.07	.003	sad	0.05	0.03
adverb	-0.07	.004	ppron	0.05	0.03
time	-0.07	.004	space	-0.05	0.04
function	-0.07	.005	article	-0.05	0.04
negate	0.07	.006	leisure	0.05	0.04
prep	-0.06	.007	friend	0.05	0.047

Table 6.1: Correlation of significant LIWC linguistic features with the GAD-7

from both the anxious and nonanxious classes. This avoids the problem of class imbalance, which, if it happens, causes low predictive accuracy for the minority class (which is the anxious class in our case). To do so, samples were randomly selected and removed from the majority class until the majority class had an equal number of samples as the minority class.

The model construction and training step use three datasets: a training dataset (80% of the entire sub-sampled data); a validation dataset (20% of the training data) which is used to select the best hyperparameters during training; and a test dataset (20% of the entire sub-sampled dataset and separate from the training data). The test set is used to evaluate the performance of the trained model using the area under the receiver operating characteristic (AUROC) metric.

6.2.2 Construction and Evaluation of Transformer-Based Model

The advent and remarkable success of transformer-based neural networks for Natural Language Processing was discussed in Chapter 2. One property that distinguishes different transformer models is the amount of textual words/tokens that will fit into the context window that the model considers at one time. That window size is limited by the computational burden of the method of attention [39][105]. These windows range in size from size 512 tokens [40] up to 4096 [43].

The transcripts that were collected, described in Chapter 3, requires participants to speak for 5 minutes and range in size up to 1,190 tokens (recall that tokens are either words or parts of a word). For that reason, our model will require a transformer model that can process sequences of that length. We selected the Longformer transformer model [106] (obtained from the Huggingface model hub [107]) because it has a contextual window size of 4,096 tokens.

The Longformer model has approximately 149M total parameters and was pre-trained on a corpus of long documents such as the Books corpus [108] (0.5B tokens), English Wikipedia (2.1B tokens), Realnews dataset [109] (1.8B tokens), and Stories corpus [110] (2.1B tokens). We fine-tuned that model (as described in Chapter 2) to create a classifier for the anxiety classification task. Fine-tuning takes a pre-trained model, which consists of a transformer block followed by a 'language model' multi-layer perceptron (MLP) "head" that predicts a word subsequent to the context. That prediction head is removed and replaced with an untrained (and randomly initialized) MLP neural network called a 'classification head.' In our case, this need only be a one-output binary classification indicating anxious or nonanxious. That newly created model is then fine-tuned on the specific task of predicting above or below the screening threshold for GAD based on the GAD-7 scale.

When fine-tuning the model, we explore three methods: 1) keeping the base transformer block as it is and only fine-tuning the classification head (the MLP); 2) Lower learning rate for the base transformer block compared to the head; and 3) Equal learning rate for both the base model and the classification head. The third method gave us the best performance even though the first method had the fastest completion time since we were only updating the weights of the classification head. In the Results section below, we will present the results of the third method, which gave the best performance.

The input dataset was processed similarly to the datasets for the baseline model. Beginning with the set of transcripts from all participants, the data was first under-sampled to equalize representation from both the anxious and nonanxious classes. The model fine-tuning step also uses three datasets: a training dataset (80% of the full data); a validation dataset (20% of the training data); and a test dataset (20% of the full dataset) which is used to evaluate the performance of the trained model using AUROC metric. Note that we did not use the 5-fold cross-validation method we used in Chapter 5 due to the time it takes to fine-tune a single transformer-based model. Instead, we used the 20/80 test/training data split methodology, which is another standard in the machine learning community [44].

6.2.3 Model Interpretation

Deep neural networks [111], including the transformer network used in this work, do not lend themselves to an easy explanation of which features/factors were important for any specific prediction. This contrasts with the logistic regression model (the baseline) in which the weights on each feature are informative. One contribution of this work is to provide some interpretation of the results of this model – specifically, given a specific transcript, to suggest which words or group of words were the most influential in the model's prediction of anxious or nonanxious.

To achieve this model interpretation, we used the Integrated Gradient (IG) [104] method. IG computes a score associated with every input (word/token) to the model. The score is a function of the rate of change of the prediction with respect to that specific input. When the score of specific input is higher and positive it is an indication that the input had a larger influence towards producing a positive classification (which is the anxious class in our case). Similarly, a high negative score indicates influence towards the negative, nonanxious case. This score is referred to as the attribution score of the input token. We used a library called Transformer Interpret [112] to compute the attribution scores for each word in a given transcript.

Using the attribution score, we can report specific words/tokens that are influential in the prediction towards both anxious and nonanxious prediction results. From there, we explore the specific context of those words to look for patterns of language that were influential. Below we describe the specific method for selecting influential words and important patterns.

First, the attribution score of every word/token in all the transcripts from all participants is computed. In viewing the plot (Figure 6.1) of the distribution of the number of words with each score, the knee of the distribution appears around a threshold attribution score of 0.05, which provides a tractable number of words to explore. Those tokens with scores above the threshold of 0.05 are presented in the Results section.

To determine if there were patterns in the context surrounding the high-attribution words, we manually reviewed the surrounding context of each high-attribution word. The patterns we have observed from these contexts, together with the specific direction of the prediction (anxious/nonanxious), are presented in the Result section.

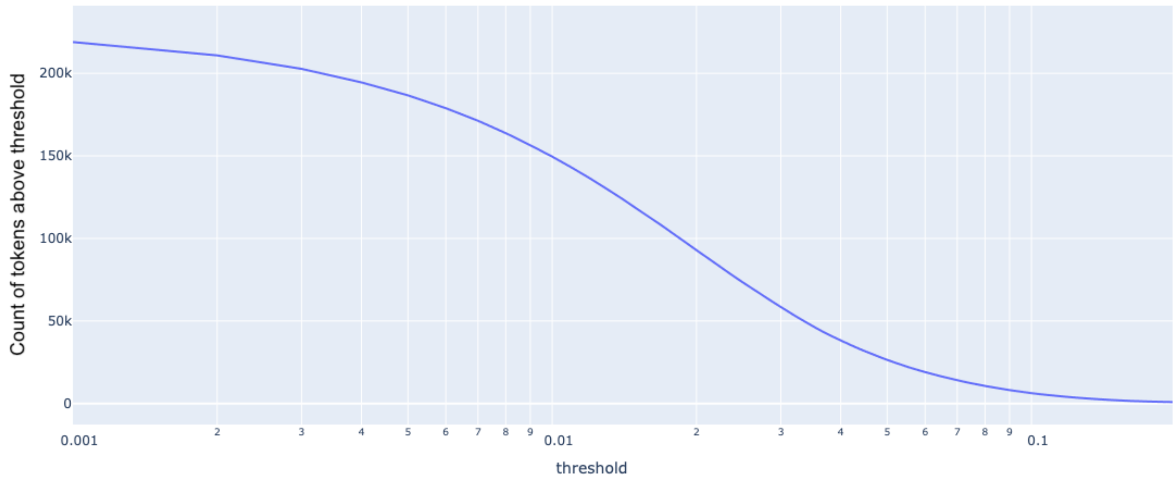


Figure 6.1: The number of tokens which have an attribution score greater than a threshold

6.3 Results

6.3.1 Classification Model Performance

In this section, the AUROC of the two binary classification models described above is presented. The first model is the logistic regression model that uses the LIWC Features as input. The LIWC features used were the ones shown to be significantly correlated with the GAD-7 in Chapter 4 and listed in Table 6.1. The second model is the fine-tuned transformer-based model. Table 6.2 presents the AUROC value obtained for these two models, while Figure 6.2 presents the AUROC curve of the fine-tuned transformer-based model.

Model	**AUROC**
Logistic regression with LIWC features	0.58
Transformer-based model	0.64

Table 6.2: AUROC value of classification models on held-out test dataset

6.3.2 Model Interpretation: Tokens Used to Predict both Anxious and Nonanxious

In Section 6.2.3, we described the Integrated Gradient method that is used to determine an attribution score for each word in a transcript. That score gives an indication of how strongly the word is implicated in the prediction towards anxious (if

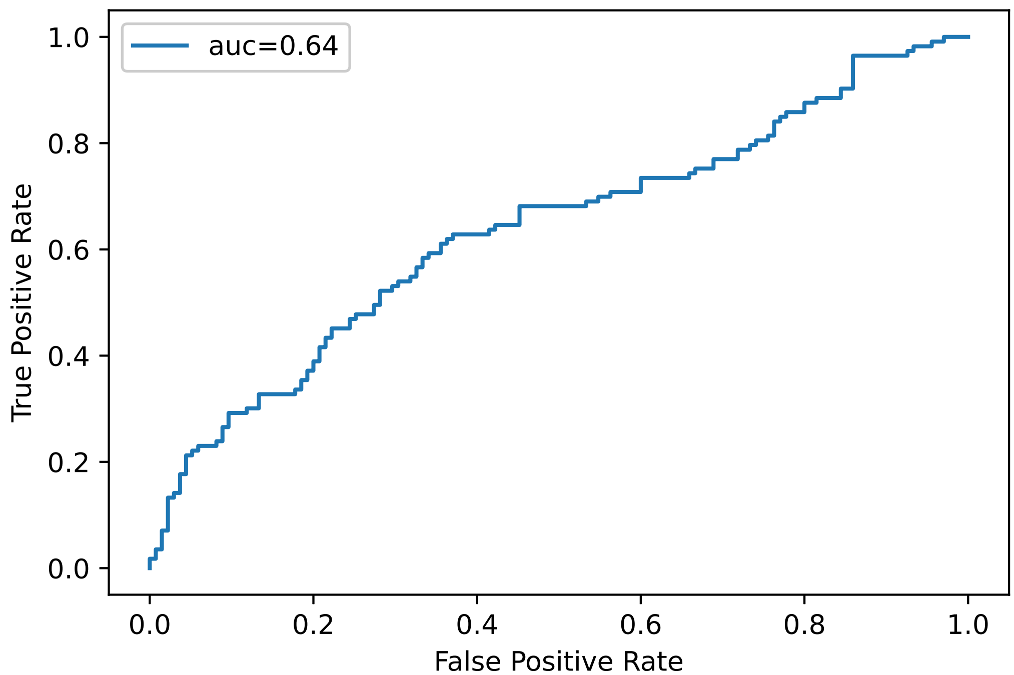


Figure 6.2: AUROC Curve of the fine-tuned transformer-based model

positive) or nonanxious (if negative). Table 6.3 presents the number of times (across all transcripts) that a specific token (listed in Column 1) had a high attribution score (absolute value above .05 as described above) based on the IG method. The tokens presented in Table 6.3 were selected because they have a high count in having both a high positive and high negative attribution score – i.e., at predicting both anxious and nonanxious. Note that tokens could be words, parts of a word and also special characters used by the speech-to-text system to indicate non-verbal events. For example, the speech-to-text system generates a "." to indicate a silent pause in speech.

Table 6.4 presents the patterns that were observed with examples taken from the actual transcripts of the recruited participants where the same token has influenced the prediction towards anxious in some cases and towards nonanxious in other cases. Column 1 lists those tokens and indicates which direction (anxious/nonanxious) the row in the table describes. Column 3 describes the pattern of the context that we inferred was relevant, using the qualitative analysis described in Section 6.2.3. Column 4 gives a specific example of that pattern, taken from the transcripts (but sometimes modified for clarity), and Column 5 gives the number of occurrences of that pattern across all transcripts.

[1][Silent pause]: A silent pause in a speech, as determined by the speech-to-text software.
[2][Filled pause]: A pause consisting of filler words such as "um", "mm", "uh", "hmm", or "mhm"

Token	**Number of times influencing towards anxious (%)**	**Number of times influencing towards nonanxious (%)**
I	3032 (88%)	427 (12%)
[Silent pause][1]	2933 (20%)	11557 (80%)
[Filled pause][2]	2039 (59%)	1395 (41%)
and	913 (57%)	682 (43%)

Table 6.3: Tokens with high attribution and high counts of prediction influence

6.4 Discussion

The goal of this chapter is to determine how well a transformer-based neural network model can predict GAD and compare it to an LIWC-based logistic-regression predictor. In this section, we discuss the implications of the findings presented in the previous section, as well as the limitations of the study.

6.4.1 Classification Model Performance

The logistic regression model with LIWC features is the baseline point of comparison. Table 6.2 shows that this model performed better than a random model (as it has an AUROC above 0.5). This indicates that the count and type of words used by individuals do give us some insight into their anxiety and is in line with prior work [9][57][58][59][65] that explored the association between LIWC features and anxiety.

Table 6.2 also gives the performance of the fine-tuned transformer model, which is larger than the baseline model by 10%. We believe that a model that takes context into account can achieve higher predictive performance. It suggests that transformer models, which search multi-word contexts to find patterns, can extract additional information to use in a prediction than is possible with single-word prediction. The results in Table 6.3 and Table 6.4 allow us to understand, in more detail, what the model based its predictions on, discussed below.

6.4.2 Model Interpretation

Here we discuss our attempt to provide an interpretation of the results from the transformer-based Longformer model. Table 6.3 shows the tokens with high attribution scores, as defined above, and a high count at influencing the prediction toward anxious and nonanxious. The first entries in Table 6.4 describe the effects of the singular pronoun "I." Depending on the context, the use of the word "I" in different

Token	**Prediction class**	**Definition of Pattern**	**Example of Pattern**	**Number of Occurrences in All Transcripts**
I				
	Anxious			
		"I" followed by a filled pause	I have um I worked very well	476
		"I" is the first word in a sentence but in the middle of the transcript.	I get on well with various different groups	1567
		Starting a sentence and pausing after just saying "I"	I [Silent pause]	208
		"I" used together with am, have	I am able to relate	1515
	Nonanxious			
		"I" used in a sentence to reference others.	I was able to remember all their names	47
		"I" is the very first word in the transcript	<speech starts>I think I would be perfect for this job	171
		"I" used to describe a positive thing about oneself	I am imaginative	77
[Silent pause]				
	Anxious			
		[Silent pause] used before or after a [Filled pause]	[Silent pause] um mm [Silent pause]	1740
		Starting a sentence and pausing within a short period of time.	my [Silent pause]	2057
	Nonanxious			
		Pauses during speech that are not accompanied by a [Filled pause] and produce a correct sentence.	bring a specific [Silent pause] area of expertise of functionality	11557
[Filled pause]				
	Anxious			
		Filled pause used together with a silent pause	[Silent pause] um mm [Silent pause]	1577
		Filled pause used in the beginning of a speech	<speech starts>hello um I just like to	23
	Nonanxious			
		Filled pause used in the middle of a sentence without a silent pause	many years playing music at parties um starting at the age of	480

Token	Prediction class	Definition of Pattern	Example of Pattern	Number of Occurrences in All Transcripts
and				
	Anxious			
		Finishing a sentence with ”and”	really think about it in detail and	519
		Using “and” more than once in a sentence	was tasked in doing that and and I did that successfully and that	187
		Starting a sentence and pausing after just saying “and”	and [Silent pause] sometimes things are	282
	Nonanxious			
		“and” used grammatically correctly in a sentence	eight people for twelve years and after that I managed an additional	572

Table 6.4: Cases in which the tokens in Table 6.3 influenced the prediction of both anxious and nonanxious

contexts influences both towards the prediction of anxious and nonanxious. By contrast, previous studies have shown increased use of “I” to be associated only towards the direction of anxiety [57]. A possible reason that “I” is associated with anxiety is that anxious individuals will try to divert their attention from anxiety-inducing events by focusing on themselves. This might result in the frequent use of “I” in their speech.

The present work, however, shows how the context around the word “I” matters – although its presence influenced the prediction toward anxiety for the majority (88%) of the cases, it also influenced the prediction towards nonanxious in 12% of the cases. One pattern around “I” that influenced towards the prediction of nonanxious is when it is used to reference others (e.g., “I was able to remember all their names”). This case is opposite to where anxious individuals tend to focus on themselves, and so a possible reason as to why focusing on others would influence the prediction of nonanxious. Another pattern of “I” that influences towards the prediction of nonanxious is when it is one of the very first words at the start of speech (i.e., at the very beginning). This may be because confident people might start their speech by introducing themselves or placing the focus on themselves before proceeding with whatever the subject matter of their speech is. Similarly, relating to confidence, there is a pattern where “I” is used to speak something positive about oneself that influences towards the prediction of nonanxious. These cases suggest that confidence is related to the state of being nonanxious.

Silent pauses ([Silent pause] in Tables 6.3 and 6.4) mainly had an influence towards the prediction of nonanxious, for 80% of the cases. This is in line with prior work [113], which indicates that anxiety is associated with a reduction in the number of silent pauses during speech. The authors suggest that pausing during speech represents a cognitive activity which is seen more in nonanxious individuals compared to anxious individuals.

However, there were also times when a silent pause influenced the prediction towards anxiety. The difference was the context: when a silent pause is used together with a filled pause, and also pausing after saying a single word. These cases hint towards the difficulty of producing complete sentences and instead using filler words in the middle of their speech or being unable to finish a sentence. This might be due to a higher level of anxiety.

The other two types of tokens presented in Table 6.3 ([Filled Pauses] and "and") had a high count in influencing toward the prediction towards both anxious and nonanxious. We believe they have a high count because they are commonly used tokens in a speech to text transcripts. A pattern that stands out around both the two ([Filled Pauses] and "and") types of tokens is the use of a grammatically correct language which is exhibited more so by the nonanxious participants. Prior work [114] suggests that anxiety causes disfluencies in speech and so could be a possible explanation for the anxious participants' use of grammatically incorrect language. Our results suggest that the model is picking up on this grammatical incorrectness.

6.4.3 Conclusion

In this study, we have presented results from a large-N study aiming to predict if participants who provide samples of speech fall below or above the screening threshold for GAD based on the GAD-7 scale. More specifically, we investigate the importance of multi-word context when predicting the presence or absence of anxiety. While prior studies have shown that the choice of individual words are a good predictor of mental health disorders, we have shown that choice of words, together with the context, is an even better predictor. Furthermore, transformer-based neural network models can be leveraged to find such linguistic patterns that help identify if a certain word, given the context, would predict anxious or not. So, the linguistic patterns of speech identified using Transformer models could probably be applied towards the screening of other types of mental health disorders.

Chapter 7

Full Multi-Modal Model

In Chapter 5 we presented a model that predicts anxiety from acoustic, simple linguistic (LIWC) and demographic features. In Chapter 6 we presented a transformer-based neural network model that focused only on the transcripts and the full context of the words. One of the main goals of this work is to produce and evaluate the performance of a model that predicts general anxiety disorder from speech. To that end we now describe the construction and evaluation of a model that predicts above or below the screening threshold of GAD by merging all the models from the previous chapters into one single model.

We hypothesize that by combining all the input features and models described in earlier chapters, it is possible to achieve a better prediction performance in screening for GAD.

7.1 Methods

7.1.1 Inputs to the Classification Model

In this chapter, we explore different streams of information as inputs to a final classification model that predicts above or below the screening threshold of GAD. The first inputs are the acoustic features that were determined to have a statistically significant correlation with the GAD-7. These features were identified and presented in Chapter 4. The second sets of features are the LIWC features, which will be referred to as the simple linguistic features which were also identified and selected in Chapter 4.

The third input to the classification model comes from the transformer-based model which by itself takes the transcript as an input. We have explored using both the binary and probabilistic (continuous) output of the transformer-based model as an input to the final model. The binary input gave us a better prediction performance compared to the continuous input so the results we present are based on the binary input. Figure 7.1 illustrates the structure of the full multi-modal model.

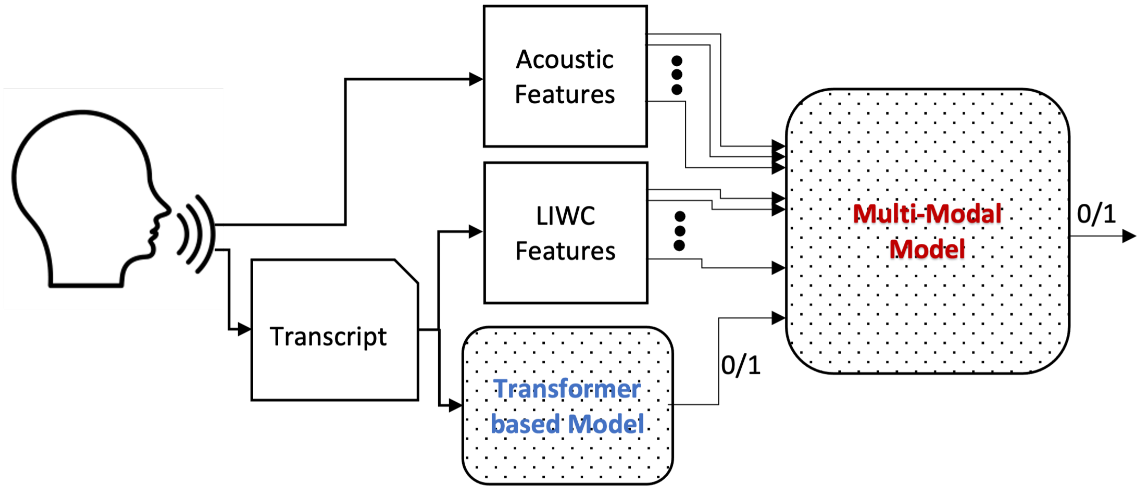


Figure 7.1: Multi Modal Model

We have also explored using demographic information such as age, sex and income as input features to the model. We will explore the prediction performance of the different types of inputs separately and also by combining all the inputs into one single model.

7.1.2 Construction and Evaluation of Classification Models

This section describes the construction, training and the evaluation of the prediction model that classifies if a sample is above or below the screening threshold of GAD. The prediction model is a logistic regression model that classifies between the two classes - anxious and nonanxious. The construction and evaluation of the prediction model steps are similar to the ones described in Chapter 6. First, all the input features were normalized so each feature would have a mean of zero and a standard deviation of 1. Next, the data was under-sampled to equalize representation from both the anxious and nonanxious classes. To do so, samples were randomly selected and removed from the majority class until the majority class had an equal number of samples as the minority class.

The model construction and training step uses three datasets: a training dataset (80% of the entire sub-sampled data); a validation dataset (20% of the training data) which is used to select the best hyperparameters during training; and a test dataset (20% of the entire sub-sampled dataset and separate from the training data). The test set is used to evaluate the performance of the trained model using the area under the receiver operating characteristic (AUROC) metric.

7.2 Results

7.2.1 Predicting Anxiety from Speech

This work has explored several methods for the prediction of above or below the GAD-7 threshold based on speech as a method for screening for GAD. Table 7.1 presents the AUROC of the models, which used the different kinds of features (separately) and the full multi-modal binary anxious/nonanxious model, listed as 'Combined' in the table.

Input	**AUROC**
Acoustic only	0.6
LIWC only	0.58
Transformer-based	0.64
Combined	0.67

Table 7.1: Prediction performance when using different types of speech features as input

7.2.2 Using Participant's Demographics

This section presents the model's performance when it is augmented with demographic information such as age, sex, and income. Table 7.2 presents and compares the prediction performance of two types of models: a model that uses the combined speech features as input and a model that uses the combined speech features together with the participants' demographic.

Input	**AUROC**
Combined Speech features	0.67
Combined Speech features plus demographics	0.68

Table 7.2: Comparison of a model trained using only the combined speech features with a model that also uses demographic information

7.3 Discussion

Previous chapters (Chapters 5 & 6) that discuss the prediction performance of models which takes different kind of features from speech have shown that it is possible to achieve a better prediction performance than a random model. Table 7.1, which presents the performance of the different types of inputs separately, also shows that each type of input had a better performance than a random model. This is an interesting result (to know we could achieve above random prediction with each type of input) because one might end up with a case where only one type of information from speech is available while others might not be. For example, if the only information we have from a patient is a written text, we will not have access to the acoustic features. Similarly, the only available information we get could be an out-of-order list of words for privacy and confidentiality reasons. This was the case in another study in our research group where we were allowed to keep only the out-of-order list of words since it was a passively collected speech sample, and we had to keep the privacy of the participant [9]. In that case, we can only make use of the LIWC features and not the context-aware transformer-based model. However, for the cases where we can make use of all the available input features, the results presented in Table 7.1 show that we were able to achieve a better prediction performance when all the available speech features were used.

In Chapter 5, we have shown that including demographic information such as age, sex, and income could enhance the prediction performance of a model. Similarly, in this chapter, we explore how much performance gain we could achieve when we add demographics together with the other speech features. Table 7.2 shows a slight performance gain when augmenting demographic information and, therefore, suggests that demographics should be considered when deploying an anxiety detection system.

7.3.1 Improvement from Multiple Predictions

Here, we will discuss how we could improve a model's prediction performance by taking multiple measurements. One possible method of taking multiple measurements is through passive collection of speech sample while an individual is going through their daily activities. Assuming that a model's prediction accuracy is better than random, it should be possible to increase the accuracy in a procedure in which multiple measurements are taken, under the assumption that each measurement from a different sample of speech is independent. This enhanced accuracy could be achieved by considering the model's native accuracy as follows: Let the accuracy of a correct prediction from a single measurement be a, and that we take N successive speech samples and make N successive corresponding predictions using our model.

Our decision procedure is that the majority measurement is correct - whichever result, anxious or non-anxious, happens in more than $N/2$ of the measurements. We are interested in the probability that this decision procedure produces a correct result.

It is possible to calculate the probability that n or more of the N measurements would have a correct prediction using the cumulative binomial distribution function (shown in Equation 7.1). Given the decision procedure of taking the correct result to be the majority result of the N trials, we set the value of n to be $\lceil N/2 \rceil$, which computes the total probability of more than $N/2$ correct answers. As long as the single prediction $a > 0.5$, then the computed probability A will be greater than a.

$$A = \sum_{i=n}^{N} \binom{N}{i} a^i (1-a)^{N-i} \tag{7.1}$$

Note that, however, if the accuracy of a single prediction a is random ($a = 0.5$), then there will be no improvement in the final probability. Figure 7.2 shows the values of A for different values of N and a while setting n to be $\lceil N/2 \rceil$. Observe that for a single prediction accuracy, $a = 0.65$, we get to an accuracy, A, of 0.9 at around 17 measurements. If we could take 2 to 3 measurements per day, we would get around 14 to 21 measurements per week which gets us to a correct decision probability of around 0.9 from a single prediction accuracy of 0.65.

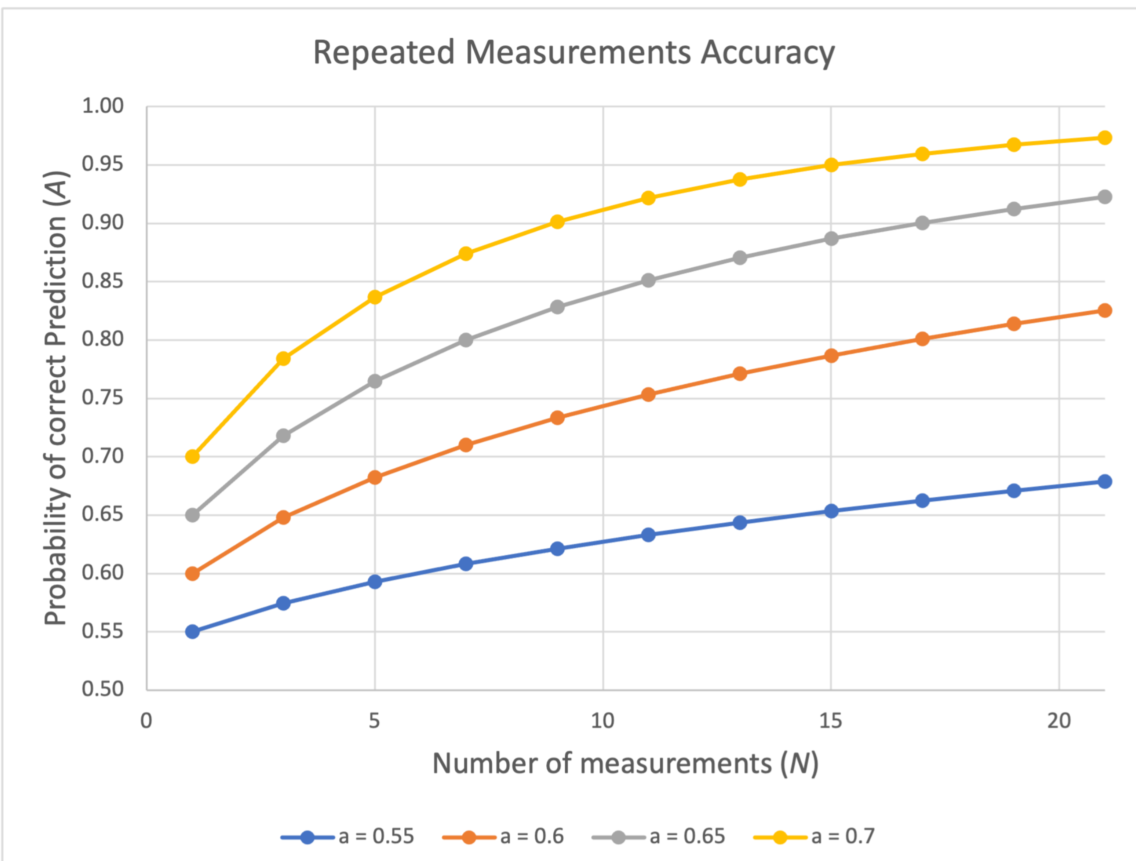


Figure 7.2: Accuracy improvement from a repeated measurement

This result does rely on the assumption that the accuracy of individual predictions are independent when in reality, they are not since the measurements are coming from the same person. However, the set of words coming from the person would be different, and a more spaced-out measurements (such as the 2 to 3 measurements per day mentioned above) may reduce the dependency between the samples. So, to summarize, it is possible to increase the accuracy of correct prediction by taking multiple measurements and taking the class (anxious or non-anxious) that has been predicted for the majority of the time as the final predicted value.

7.3.2 Conclusion

This chapter presented the performance gain that can be achieved when combining all the available information from speech that we explored in previous chapters. In earlier chapters, we mainly had two models that predict the above or below the screening threshold of the GAD-7 – the first is a model that predicts from acoustic and LIWC features, and the other is the transformer-based model. It made sense for our next step to combine these models into one. As we hypothesized, this combined multi-modal model had a better prediction performance compared to the separate models. We also discussed that if we can make multiple measurements passively, this multi-modal model has the possibility to be used as an automated screener for anxiety disorder.

Chapter 8

Conclusion

In this work, we were able to show that the different information obtained from a speech audio sample has significant associations with anxiety. Furthermore, these different information also have the ability to predict above or below the screening threshold of GAD based on the GAD-7 scale.

In the sections below, we will summarize our key contributions, identify the limitations of our work, and finally, our recommendation for future work.

8.1 Contributions

8.1.1 A Speech Sample Dataset Labeled with Anxiety Severity

In this study, we have designed and constructed an online web-based system that was able to collect speech samples labelled with the GAD-7. The largest sample size from the prior works explored (presented in Chapter 2) was 239. This limits the potential for generalizability to a larger population. In our study, we were able to recruit 2,000 participants - a significant contribution to the study of anxiety disorder.

We also aimed to obtain broader demographics than those in the prior study. Most prior studies focused on a certain type of demographics. For example, the study by McGinnis et al. [60] focused on children, and the studies by Week et al. [56], Salekin et al. [63], and Rook et al. [65], focused on undergraduate students. Both these types significantly limited the age range of the participants. The data presented in Table 3.3 show that the age range of our participants had a broader distribution. The same is true for personal income, which showed a range of economic statuses in our participant pool.

8.1.2 Validated Acoustic and Linguistic Features Suggested by Prior Work

One of our main objectives in this study was to validate the association of previously suggested acoustic and linguistic features of speech with anxiety severity. To summarize the results, word count and speaking duration were negatively correlated with anxiety scores ($r = -0.12; P < .001$), indicating that participants with higher anxiety scores spoke less. Several acoustic features were also significantly ($P < .05$) associated with anxiety, including the Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficient (LPCCs), Shimmer, Fundamental Frequency, and first formant. In contrast to previous literature, second and third formant, Jitter, and ZCR-zPSD acoustic features were not significantly associated with anxiety. Linguistic features, including negative emotion words, were also associated with anxiety ($r = 0.10; P < .001$). Additionally, some linguistic relationships were sex-dependent. For example, the count of words related to power was positively associated with anxiety in the female samples ($r = 0.07, P = .03$), while it was negatively associated in the male sample($r = -0.09, P = .01$).

8.1.3 Prediction of Anxiety Disorder From Acoustic and Linguistic Features

After validating the acoustic and linguistic features suggested to have an association with anxiety in prior work, we used these significantly correlated features to build and train a prediction model. A logistic regression model using only acoustic and linguistic speech features achieved a significantly greater prediction accuracy than a random model (mean AUROC = 0.57, SD = 0.03, P < .001). When separately assessing female samples, a mean AUROC of 0.55 (SD = 0.05, P = .01) was observed. The model constructed on the male samples achieved a mean AUROC of 0.57 (SD = 0.07, P = .002). The mean AUROC increased to 0.62 (SD = 0.03, P < .001) on the all-sample dataset when demographic information (age, sex, and income) was included indicating the importance of demographics when screening for anxiety disorders. The performance also increased for the female dataset to 0.62 (SD = 0.04, P < .001) when using demographic info (age and income). In both cases, age had the highest impact on the prediction performance gain. An increase in performance was not observed when adding demographic information for the model constructed on the male samples.

These results show that a logistic regression model using acoustic and linguistic speech features can achieve above-random accuracy for predicting GAD. Importantly, the addition of basic demographic variables further improves model performance, suggesting a role for speech and demographic information to be used as automated, objective screeners of GAD.

8.1.4 Prediction of Anxiety Disorder from Transcripts using Context-aware Transformer Models

Another contribution of this study is the prediction of anxiety disorders from multi-word contexts. This was achieved using a transformer-based models. These models are capable of making a prediction based on multiple input words, unlike a prediction model which has the LIWC features as input. Our objective was to determine if a transformer-based language model can be used to screen for Generalized Anxiety Disorder from impromptu speech transcripts.

To summarize the results, The baseline LIWC-based logistic regression model had an AUROC value of 0.58. The AUROC of the fine-tuned transformer model achieved a value of 0.64. Specific words that were often implicated in the predictions were also dependent on the context. For example, the first-person singular pronoun "I," influenced towards an anxious prediction 88% of the time while it influenced towards nonanxious 12% of the time, depending on the context. Silent pauses in speech, also often implicated in predictions, influenced the prediction of anxious 20% of the time and nonanxious 80% of the time.

The results show that there is evidence that a transformer-based neural network model has increased predictive power compared to the single-word-based LIWC model. We have also shown that specific words in a specific context – a linguistic pattern – is part of the reason for the better prediction. This suggests that such transformer-based models could play a useful role in anxiety screening systems.

8.2 Limitations

In this section, we will present and discuss some of the limitations of our work. Some of these limitations might open a door for future work.

The first limitation of this study is the use of self-report measures to assess GAD. Self-report measures, by nature, are subjective opinions that individuals have about themselves while filling out the questionnaires and may not completely capture clinical symptoms. In this study, we took these self-report questionnaires as the true label of the audio samples. However, we believe that this is a good first step that gave us encouraging preliminary results. A psychiatric diagnosis would be an improved label, but it is clearly much more expensive to acquire.

A second limitation of this study is the selection bias that might be introduced during the recruitment of the participants. As presented in Figure 3.1, only 48.7%

(2212/4542) of the participants who initially accepted the offer from Prolific to participate finished the study. We were not able to collect the GAD-7 scores of the participants who did not complete the study; therefore, we do not know their levels of anxiety. It is possible that these participants had higher levels of anxiety, which caused them to drop out of the study.

A third limitation concerns the differences in the recording devices and recording locations of the participants performing each task. Ideally, we would want every sample to be recorded using the same microphone in the same location with the same acoustics. This would reduce the potential bias introduced by different factors such as recording quality or background noise. At the same time, in a real-life scenario where an application to detect anxiety might be deployed, the recording equipment and the location will likely differ for everyone. Hence, this limitation could be unavoidable, and it might even be essential to take these types of differences into consideration.

A fourth limitation of this study arises from the data collection method used with respect to the scenarios of use that were described in the Introduction section. We suggested that the prediction of anxiety from speech could be applied to a passively collected speech data gathered while the patient is going through their daily activities. This could help in automated anxiety screening, treatment monitoring, and relapse detection. However, the data used in this study were actively collected when the participants spoke in front of a camera, and it may be substantially different from such passively collected speech.

A fifth limitation was the use of a web-based participant recruitment method. Individuals willing to work on a web-based participant recruitment platform may be limited to a particular type of demographics in a certain society. For example, we noted that in our recruitment pool, there was a higher percentage of anxious participants compared with the general population. In our study, we sought generalizability, and even though our participants were more diverse in terms of demographics compared with prior studies, it could be more generalizable if we recruited participants from sources other than a web-based recruitment.

Another limitation of this study is the use of a modified version of the Trier Social Stress Test (TSST). In the original TSST, participants are asked to describe why they should be hired for their dream job in front of a live panel of judges. However, in our study, we asked the recruited participants to describe why they should be hired for their dream job in front of a camera at their own location. This is a limitation towards achieving the full replication of the TSST as a stress induction task. Nonetheless, we had an internal check where we asked them how anxious they felt before and after the TSST task (more info can be found in Chaper 3), and we observed, on average, a 25%

increase in the participants level of anxiety. Furthermore, despite its limitations, this approach also had important benefits because it enabled us to recruit a large number of participants, which would otherwise have been extremely difficult for an in-person study.

A further limitation of this work is the accuracy of the speech-to-text (STT) transcription. In this work, we used Amazon's STT program [93], which had a good transcription accuracy with an average word error rate (WER) of 7% (SD 4.6%). The fact that the word error rate is not zero means that we get the wrong transcription for some words, and our model might make a wrong prediction based on these words. However, we speculate that since STT software is getting better each year, the WER would become closer and closer to zero, so the prediction of a model based on these transcripts would get better.

A final limitation is the subjective/qualitative nature of the pattern detection in Table 6.4, which forms the basis of the insights for the interpretation of the transformer model. As described in the Methods section of Chapter 6, the transcripts were analyzed manually, and instances were selected that we believe exhibit similar patterns across multiple contexts. These are our subjective opinion of what constitutes a similar pattern, and so other researchers might be able to find other patterns that we might have overlooked. In future studies, we aim to release our transcripts for other researchers to go through our transcripts as we did and see if any other interesting patterns could be detected.

8.3 Future Work

We believe that the work around the prediction of anxiety disorder from speech can be improved in many ways. In this section, we will discuss some of the possible future works.

A key next step is to produce an openly available dataset that other researchers can also use, based on our data. Note that we can not make the audio or the video openly available to keep the privacy and the identity of the participants in our study. Instead, we will extract as many features as possible that are currently used by the research community while making sure that the collected extracted features can not be used to reconstruct the original audio or video. We might also make the speech-to-text transcript openly available after removing all the identifiable information (such as nouns such as the name of a person or a place) of each participant. This will allow for other researchers to go through our transcripts as we did in Chapter 6 and see if any other interesting patterns could be detected.

In the limitation section above, we mentioned that this study makes use of an actively collected speech sample instead of a passively collected one. The prediction of anxiety from a passively collected speech sample has the advantage of monitoring the mental health state of a patient while they go through their daily activities. Future studies could investigate the models that we suggest, but using a passively collected speech instead of an actively collected speech sample.

In Chapter 7, we have discussed that the results from multiple measurements have the possibility to improve prediction accuracy. This is our theoretical speculation. Therefore, we recommend that future studies explore the collection of multiple speech samples sampled throughout the day or week and investigate the extent to which the prediction accuracy can be improved.

Another possible future work is to explore whether features of speech from task 1 (simple reading of a passage) exhibit correlations with the GAD-7. The results would inform us how much signal of anxiety can be detected without explicitly evoking stress and anxiety. Note that here, we can only make use of the acoustic features since the transcript would be the same for the different speech samples. Another method is to explore whether these features could be used as a control for the features of task 2 (the modified TSST task).

Finally, most prior studies in the prediction of mental health disorders from speech have shown that the choice of individual words are a good predictor of mental health state. In this study, we have shown that the choice of words, together with the context, is an even better predictor. Furthermore, transformer-based neural network models can be leveraged to find such linguistic patterns that help identify if a certain word, given the context, would predict anxious or not. So, we recommend that future studies explore the linguistic patterns of speech identified using Transformer models and apply them towards the screening of other types of mental health disorders.

Appendix A

Generalized Anxiety Disorder 7-item (GAD-7) scale

Generalized Anxiety Disorder 7-item (GAD-7) scale

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid as if something awful might happen	0	1	2	3

Total Score	___	=	Add Columns	___	+	___	+	___

Source: *Spitzer, Robert L., Kurt Kroenke, Janet BW Williams, and Bernd Löwe. "A brief measure for assessing generalized anxiety disorder: the GAD-7." Archives of internal medicine 166, no. 10 (2006): 1092-1097.*

Appendix B

Acoustic Features suggested to have an Association with Anxiety in Previous Studies

B.1 Mel Frequency Cepstral Coefficients (MFCC)

The Mel Frequency Cepstral Coefficients, or MFCC for short, are one of the most widely-used features extracted from speech for different tasks such as Speech-To-Text (STT) [19], Speaker Recognition [20], and also the prediction of different mental health illnesses [21].

There are multiple steps of processing to obtain the final MFCC values. Given a certain window of speech audio, the first step is to transform the audio signal from a time-domain representation to a frequency-domain representation using a Discrete Fourier Transform (DFT), producing a frequency spectrum. Then, the spectrum will be filtered using mel-scale, which mimics the response of the human ear. Then the log of the mel-scale signal is taken and then another DFT is applied. This last step produces a 1-D array representing the spectrum of the log of the spectrum and is known as the *Cepstrum*. The MFCCs are the amplitude of the first 13 values (usually, but can also use more depending on the task) of the Cepstrum.

B.2 Linear Prediction Cepstral Coefficient (LPCC)

While the MFCC come from the mel-scale filtering, LPCC are coefficient derived from the linear prediction cepstral representation of an audio signal. All the steps of the extraction of LPCC are similar to MFCC (described above) except for the filtering step. During the MFCC extraction, the frequency spectrum are filtered using mel-scale. However, during the extraction of LPCC, the frequency spectrum are filtered on a linear scale.

B.3 ZCR zPSD

ZCR zPSD is an abbreviated form for the Zero Crossing Rate (ZCR) of the z-score of the Power Spectral Density (PSD). The PSD is a measure of a signal's power content versus frequency. The z-score of the PSD (zPSD) measures how different the values are from the mean in terms of the standard deviation - i.e., if the z-score is 0 then the data point's score is identical to the mean and a z-score of 1 indicates a value that is one standard deviation from the mean. It could also be both positive and negative. The zero crossing rate (ZCR) is the rate at which the zPSD changes from positive to negative and vice versa.

B.4 Amount of speech

The amount of time a person is speaking and related metrics such as the percentage of silence. We will use two features to measure the amount of speech: the amount of time, in seconds, that speech was present, and the total number of words present in a speech-to-text transcript.

B.5 Articulation rate

The articulation rate indicates how fast a person is speaking. We use a software library called My Voice Analysis [22] to extract this feature from a speech audio.

B.6 Fundamental frequency

The fundamental frequency (F0) of speech is the rate at which the glottis vibrates, and is also known as the voice *pitch*. This glottis vibration creates a periodic wave for which a particular signal $s(t)$ in a single period t would be equal to the signal at $s(t+T)$ where T is the period of the wave. The fundamental frequency is the inverse of the smallest possible period T.

A review conducted in 2006 [23] indicates that there are more than 70 ways of extracting F0, which shows the task of extracting F0 is not an easy one. For this study, we will be using a popular and widely cited tool known as Praat [24]. By default, Praat calculates the F0 value for every 10 ms window producing 100 F0 values per second. This means that the F0 value will vary (as indeed pitch does vary during speech) over time. Therefore, descriptive statistics such as the mean F0 and the standard deviation of F0 can be used for further analysis of the fundamental frequency.

B.7 F1, F2, F3

These are the first, second and third formants [25] which are the spectral maximum from an acoustic resonance of the vocal tract [26]. They are frequency peaks in a spectrum which have a high degree of energy. Typically, the first formant is around 500Hz, the second is around 1500Hz and the third is around 2500Hz.

B.8 Jitter

In signal processing, jitter is the cycle-to-cycle F0 variation in the sound wave. i.e., given N F0 computations from certain length audio (for example, 100 F0 sampled per second using the default Praat settings), jitter would be the average absolute difference between successive F0 samples. Praat will also be used for the estimation of jitter.

$$Jitter = \frac{1}{N}\sum_{i=1}^{N-1}|F0_i - F0_{i+1}| \tag{B.1}$$

B.9 Shimmer

Shimmer is the cycle-to-cycle amplitude variation of the sound wave. In equation B.2 below A_i and A_{i+1} are amplitudes of consecutive periods.

$$Shimmer = \frac{1}{N}\sum_{i=1}^{N-1}|A_i - A_{i+1}| \tag{B.2}$$

B.10 Intensity

Intensity is often interpreted as the loudness of sound. The intensity for a given sound frame is calculated as the squared mean of the amplitude of a sound wave within that given frame.

Appendix C

Web Application Screenshot

GAD-7 Questionnaire Page

UNIVERSITY OF TORONTO | **Measuring Anxiety in Speech**

Questionaire

Please answer the questions below by selecting the response that best fits you.

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all	Several days	Over half the days	Nearly every day
1. Feeling nervous, anxious, or on edge	○	○	○	○
2. Not being able to stop or control worrying	○	○	○	○
3. Worrying too much about different things	○	○	○	○
4. Trouble relaxing	○	○	○	○
5. Being so restless that it's hard to sit still	○	○	○	○
6. Becoming easily annoyed or irritable	○	○	○	○
7. Feeling afraid as if something awful might happen	○	○	○	○

Next>>

Task 1 Page

UNIVERSITY OF TORONTO | **Measuring Anxiety in Speech**

For this task, read the passage below, called "My Grandfather." When you are ready to read the passage out loud, click the Record button and begin talking. Click stop after you are finished. Note that clicking Stop will submit your video recording.

My Grandfather:
You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an ancient, black frock coat, usually minus several buttons.

A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon a small organ.

Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

Record Stop

Task 2 Page

UNIVERSITY OF TORONTO | **Measuring Anxiety in Speech**

For this task, imagine that you are interviewing for a job you really want.

Before starting, please do the following:

1. **Decide what that "dream" job is for you.**
2. **Think about how you would convince an interviewer to hire you**

When ready, push the record button. You have 5 minutes to explain why you're the right person for the job. Be as convincing as possible. Assume that you are speaking to a person who is interviewing you for that job.

Your video will be viewed by researchers studying your behaviour and language.

Recording time countdown: 5m 0s

Record Stop

You'll have to allow your browser to use the microphone and the camera

Post-Task Self-Reported Anxiety Measure Page

UNIVERSITY OF TORONTO | **Measuring Anxiety in Speech**

Level of Anxiety

Please tell us how anxious you felt during the video recording. Anxiety is a feeling of being worried, nervous or stressed.

Not anxious at all | Somewhat anxious | Very anxious | Extremely anxious

Next>>

Appendix D

My Grandfather Passage

My Grandfather:
You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an ancient, black frock coat, usually minus several buttons.

A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon a small organ.

Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

Appendix E

Speech Encouragement Statements

Statements used to encourage speech when silence is detected:

- You still have some time left. Please continue!
- What are your personal strengths?
- What are your major shortcomings?
- Do you have enemies? Why?
- What do you think about teamwork?
- What do your boss/family/colleagues think about you? Why?

Appendix F

Excluded Data Analysis

Excluded data analysis

The following Table shows significant correlations obtained from the data that have not been included in our study based on the data inclusion steps described in Section "Recruitment and Data Inclusion."

N = 256		
Feature	**r**	***P***
home	0.20	0.04
Sixltr	-0.19	0.04
anx	0.19	0.049

Appendix G

Ethics Protocol

UNIVERSITY OF TORONTO

OFFICE OF THE VICE-PRESIDENT, RESEARCH AND INNOVATION

Human Participant Ethics Protocol Submission

CONFIDENTIAL

0 - Identification

RIS Human Protocol Number
37584

Protocol Title
Measuring Anxiety in Speech

Protocol Type
Investigator Submission

Applicant Information

Applicant Name
Prof Jonathan Rose

Rank / Position	**Department / Faculty**
Professor	Dept of Electrical & Computer Eng - Faculty of App

Business Telephone	**Extension**
416-978-6992	

Email Address
jonathan.rose@ece.utoronto.ca

Amendments Details

Please describe the proposed study amendments or modifications. (Amend the body of the protocol as required) :

The previous version of the study only contains one standardized questionnaire - the GAD-7, which is a validated, standard way to screen for Generalized Anxiety Disorder. Some of our results suggest that there is co-morbidity with depression, and so we also want to screen for that using the well-known PHQ-8 (Patient Health Questionnaire, 8 item questionnaire), a standardized questionnaire that measures the level of depression. It will be administered at the end of the protocol. We have attached a PDF of the PHQ-8 questionnaire in this protocol. The PHQ-8 explicitly omits the 9th Question of the PHQ-9, which asks about self harm. In addition, if the subject clicks anything other than a 'not at all' on Question 2, which is the level of "Feeling down, depressed or hopeless", the website will provide a list of online mental health resources.

Will the proposed amendment change the overall purpose of the study? if Yes, a new protocol maybe requested by the REB. ○ Yes ◉ No

Will the proposed amendment affect the vulnerability of the participant group or the research risk? ○ Yes ◉ No

What follow-up action do you recommend for study participants who are already enrolled in the study. Select all that apply.

- Inform study Participants: ☐
- Revise consent / assent forms (attach forms in section 9): ☐
- Other- Please Describe: ☐

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

No action required: ☒

Collaborators/Co-Investigators

Name	Department	Email	Phone	Designation	Alt Contact
Bazen Teferra	Electrical and Computer Engineering	bazen.teferra@mail.utoronto.ca	416-738-5590	Co-Investigator	
Ludovic Rheault	Dept of Political Science	ludovic.rheault@utoronto.ca	4169785018	Co-Investigator & Alt	X
Bill Simpson	Winterlight Labs	bill@winterlightlabs.com	905-515-2820	Collaborator	

Projected Project Dates

Estimated Start Date
1-May-19

Estimated End Date
31-Dec-21

2 - Location

Location of the Research: ☒ University of Toronto ☒ Other Locations

Other Location Details

Type	Name	Location	Country	Contact	Email	Description
Other	Winterlight Labs	Toronto	Canada	Bill Simpson	bill@winterlightlabs.com	

Administrative Approval/Consent

Administrative Approval/Consent Needed: ○ Yes ◉ No

Community Based Particatory Research Project? ○ Yes ◉ No

Other Ethic Boards Approval(s)

Another Institution or Site involved? ◉ Yes ○ No

Approval of another Instittutional REB required? ○ Yes ◉ No

3 - Agreements and Reviews

Funding

Project Funded? ◉ Yes ○ No

External Funds Administered by U of T

App No.	Fund No.	Sponsor/Program	Status	Fund End Date	Peer Reviewed
198282	508827	Social Sciences & Humanities	Awarded	2022-02-28	

Internal U of T Funding

Source	Status	Peer Reviewed
XSeed	Awarded	X

Agreements

Funding/non-funding Agreement in Place? ◉ Yes ○ No

Document Title	Document Date

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Notice of Award	2018-06-24

Any Team Member Declared Conflict of Interest? (•) Yes ( ) No

Reviews

[x] This research has gone under scholarly review by thesis committee, departmental review committee, peer review committee, or some other equivalent

Type of Review : -e.g.: departmental research committee, supervisor, CIHR, SSHRC, OHTN, etc.

SSHRC (the grant *was* peer-reviewed)

[x] This review was specific to this protocol

[ ] The review was part of a larger grant

[ ] This research will go under scholarly review prior to funding

[ ] This review will not go under a scholarly review

4 - Potential Conflicts

Conflict of Interest

Will researchers, research team members, or immediate family members receive any personal benefit? (•) Yes ( ) No

Description of benefits

A new collaborator, Dr. Bill Simpson from Winterlight Labs, works for a Winterlight and so the knowledge and information gained from this work could benefit Winterlight and therefore Dr. Simpson. Indeed, the goals of Winterlight are very similar to the goals of this research project, and that is the reason for the collaboration.

Restrictions on Information

Are there any restrictions regarding access to, or disclosure of information (during or after closure)? ( ) Yes (•) No

Researcher Relationships

Are there any pre-existing relationships between the researchers and the researched? ( ) Yes (•) No

Collaborative Decision Making

Is this a community based project - i.e.: a collaboration between the university and a community group? ( ) Yes (•) No

5 - Project Details

Summary

Rationale

Describe the purpose and scholarly rationale for the project

In modern Psychiatric and Psychology practice, diagnosis is made through regular human interaction and conversation between practitioner and patient, in-person [1]. Our long-term goal is to aid the process of diagnosis by providing information that comes from other sources [4], including speech monitored during periods outside of clinical interaction. This gives rise to the goal of detecting relevant clinical features from that speech. One feature we would like to begin with is the detection of anxiety within a human speech. We chose anxiety as the first feature to look at because it appears in several forms of the disorder [2,3] and is associated with other mental health disorders. To train a computer model to detect anxiety in speech, we need to acquire human speech data along with 'labels' of the level of anxiety present in that speech. As a first step in creating a model, we measured anxiety in a speech from individuals without explicitly stimulating anxiety – instead, made a neutral request for speech and have participants provide us with a few minutes of audio and video of the participant speaking. Now, in the present amendment, we intend to measure anxiety in speech while following the well-known Trier social stress Task (TSST) [5]. The purpose of TSST is to induce moderate stress in participants. Studies [6,7] show a difference in response between participants with high anxiety and low anxiety after exposure to moderate stress. For this reason, we intend to measure the difference between a baseline/neutral speech to a question based on the TSST, hypothesizing that it will be different for participants with high anxiety levels compared to participants with low anxiety levels.
We have three methods of 'labelling' the level of anxiety in the speech: 1) We will ask them to rate their own level of anxiety. 2) We ask the participants to fill out a standard anxiety questionnaire (the GAD-7 and the PHQ-8) that gives us a sense of their base level of anxiety and depression. 3) Using human raters, we will view the video to assess their apparent level of anxiety. Our goal will be to see how well it is possible to predict these measures – the self-rated current state of anxiety, the more general anxiety, and the human labelled anxiety – based on the speech that the subjects produce. We hypothesize that it will be possible to

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

predict the self-rated anxiety level with some reasonable correlation.

[1] Denman, C. (2011). The place of psychotherapy in modern psychiatric practice. Advances in Psychiatric Treatment, 17(4), 243-249. doi:10.1192/apt.bp.109.007807
[2] J. Herbert, L. Brandsma, and L. Fischer. "Assessment of social anxiety and its clinical expressions," Chapter 3 in Social Anxiety, pages 45 – 94, S. Hofmann and P. DiBartolo, eds. Academic Press, San Diego, third edition, 2014.
[3] M. R. Liebowitz, "Social phobia," in Modern Problems in Pharmacopsychiatry, Vol. 22, pp. 141–173, 1987.
[4] D. Ben-Zeev, E. Scherer, R. Wang, H. Xie, and A. Campbell, "Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health, Psychiatric Rehabilitation Journal, Sept 2015, Vol. 38, No. 3, pp. 218-226.

Methods

Describe formal/informal procedures to be used

Subjects will be recruited from an online recruitment platform called Prolific (https://prolific.ac). This platform provides subjects for this kind of experiment who do so in return for payment. The subjects access the entire study from their computer at their own location. Once a subject accepts an offer and has consented to participate in the study, they will be asked to do the following: 1) Fill out the GAD-7 questionnaire, which is a validated survey [1]. It is attached in the document GAD7. This provides a measure of Generalized Anxiety over the previous two weeks and will be retained as data in the study – providing variables GAD1 through GAD7. 2) Next, after going through an audio and video check, the subject will be given a statement/story to read out loud, which is not designed to evoke any particular emotion. A set of example questions is provided in the attachment StimulusStatementQuestions. Note that we will experiment with these questions. The subject will be asked to record roughly 2-3 minutes about the statement/story by speaking verbally, into the microphone of their computer, which is also data that will be recorded in the study (called the Speech Data). They will also be asked to enable their camera to allow for video recording of their face while speaking. 3) The subject will be following a modified version of the Trier Social Stress Task (TSST) [2] with these steps: The subject will be told to imagine that they are in the role of a job applicant invited for a personal interview with a hiring manager. They are also asked to imagine that they are interviewing for a job that they really want (their 'dream' job). They are given a minimum time to prepare their thoughts – deciding what their 'dream' job is, and how they would convince an interviewer that they are the right person for that job. The subjects are also told that the video will be viewed by researchers studying their behaviour and language. Note that the Trier Social Stress Test (TSST) is used to invoke moderate anxiety in participants. In the last decade, the original TSST has been administered over 4,000 times in sessions worldwide and is now considered the standard protocol for the experimental induction of moderate stress [3]. While participants would normally deliver their speech in front of a panel of live people, we are following our previously approved protocol and are having the participants record their responses from their home computers, in video/audio. (Note that we are also only implementing Part I of the TSST, which normally also includes a second task involving mental arithmetic). Because the TSST is a public speaking task, it is designed to invoke similar stress levels to those experienced by people in their everyday lives when they speak in meetings, teach in front of students, deliver presentations, and so on. The full description of the TSST protocol that will be used is attached (TSST.pdf). 4) Fill out the PHQ-8 questionnaire, which is a validated survey [4]. It is attached in the document PHQ8. This provides a measure of Depression over the previous two weeks and will be retained as data in the study – providing variables PHQ1 through PHQ8. 5) Finally, the subject will be asked to rate their own level of anxiety, after each task (call this variable SRA, question attached in document PostAudioAnxietyLevel), which is also data to be recorded in the study.

Once the data is collected for the subjects, we will attempt to train a model that takes the Speech Data as input and can predict all or part of the GAD variables, or the SRA variable, or a human-labelled anxiety level. Our initial plan is to pre-process the speech data in several ways – directly using the raw audio, using Fourier transforms to extract frequency components, using the well-known cepstral coefficients of the audio – and to model these with different kinds of models (such as a support vector machine, or a neural network) for prediction. With sufficient data, we will separate the data into training, validation and testing sets, as is the usual practice.

[1] Spitzer RL, Kroenke K, Williams JB, Löwe B, "A brief measure for assessing generalized anxiety disorder: the GAD-7.", Arch Intern Med. 2006;166:1092–1097. doi: 10.1001/archinte.166.10.1092.
[2] Kirschbaum, Clemens, Karl-Martin Pirke, and Dirk H. Hellhammer. "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting." Neuropsychobiology 28.1-2 (1993): 76-81
[3] Kudielka, Brigitte M., Dirk H. Hellhammer, and Clemens Kirschbaum. "Ten Years of Research with the Trier Social Stress Test--Revisited." (2007).
[4] Kroenke, Kurt, et al. "The PHQ-8 as a measure of current depression in the general population." Journal of affective disorders 114.1-3 (2009): 163-173. APA

Copies of questionnaires, interview guided and/or other instruments used

Document Title	Document Date
GAD 7 Survey	2019-03-17
Stimulus Statements and Questions	2019-04-17
Post audio anxiety level questionnaire	2020-09-25
The Version of the Trier Social Stress Task Used	2020-09-30
PHQ-8 questionnaire	2021-04-08

Clinical Trials

Is this a clinical trial? ◯ Yes ◉ No

6 - Participants and Data

Participants and/or Data

What is the anticipated sample size of number of participants in the study? 2,121

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Describe the participants to be recruited, or the individuals about whom personally identifiable information will be collected. List the inclusion and exclusion criteria. Where the research involves extraction or collection personally identifiable information, please describe where the information will be obtained, what it will include, and how permission to access said information is being sought.

Participants will be recruited through an anonymous online recruitment platform.
Inclusion criteria:
Ages 18-65
Fluent in English
Participants must have completed at least 10 previous studies on Prolific
Participants must have completed 95% of their previous Prolific studies in a satisfactory manner (i.e., their participation was completed and approved).
Exclusion Criteria: None.
During the research, video recordings will be collected from participants which will include their face, which is personally identifiable information. The consent form clearly asks for permission to record the video.
Justification for Sample Size
We initially feel that a smaller number of participants may be needed, near 50. However, experience in machine learning suggests that having the ability to acquire larger datasets is beneficial, and so, due to the ease of online recruitment and the availability of funds we have set the sample size to 2000.

Is there any group or individual-level vulnerability related to the research that needs to be mitigated (for example, difficulty understanding consent, history of exploitation by researchers, or power differential between the researcher and the potential participant)? ○ Yes ◉ No

Recruitment

Is there recruitment of participant? ◉ Yes ○ No

Recruitment details including how, from where, and by whom

Participants will be recruited through an online participant recruiting platform called Prolific (https://www.prolific.co). Prolific is a company and website, which allows researchers (academic and market research) to recruit respondents for studies and tests for a fee. It is similar to the well-known Amazon Mechanical Turk (https://www.mturk.com) but is more specifically designed for scientific experiments. Prolific maintains a list of registered participants and for each participant a variety of demographic data such as age, gender, primary language spoken, etc. When a study is offered to the participants in Prolific, the authors of the study specify inclusion criteria that limit which potential participants are allowed to enroll.
The researchers have registered a "Research Account" and will post an advertisement for a study on the Prolific website. Individuals with a "Participant Account" on the Prolific website who meet the eligibility requirements for our study will see the study as one of the many possible studies in which they can participate.

Is participant observation used? ○ Yes ◉ No

Will translation materials be used/required? ○ Yes ◉ No

Attach copies of all recruitment posters, flyers, letters, email text, or telephone scripts

Document Title	Document Date
Not Applicable	

Compensation

Will the participants receive compensation? ◉ Yes ○ No

Type of Compensation

☒ Financial

☐ In-kind

☐ Other

Compensation Justification Details

Participants will be compensate with £2 for participation (payment is in pounds sterling as prolific is a UK company). This amounts to a rate of approximately $CAD 15 per hour of time in which the participant will participating in the study.

Is there a withdrawal clause in the research procedure? ○ Yes ◉ No

7 - Investigator Experience

Investigator Experience with this type of research

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Please provide a brief description of the previous experience for this type of research by the applicant, the research team, and any persons who will have direct contact with the applicants. If there is no previous experience, how will the applicant and research team be prepared?

The team has recent expertise that is relevant to the study of anxiety using mobile recording technology.
The principal research supervisor, Professor Rose, has been an active researcher in the field of Electrical and Computer engineering for over 30 years. His main field of research was not in this area, but in the architecture and Computer-Aided Design of Field-Programmable Gate Arrays. For that work he has been elected to be a Fellow of the IEEE, the ACM, the Royal Society of Canada, and a foreign member of the American National Academy of Engineering. Over the past seven years he has engaged in research relating to the use of mobile computing in interdisciplinary projects including medical applications, and more recently (the past five years) he has focused on technology to support mental health. In the field of Smoking Cessation, he has been engaged in the development (with Dr. Peter Selby's group from the CAMH Nicotine Dependence Clinic) of a mobile application called My Change Plan. This Android-based app has been released on the Google Play Store and on the iOS App Store. In addition, he had a separate project with Dr. Selby that came to successful fruition, on the topic of building and testing an online chatbot to help smokers move towards the decision to quit. That project had an ethics-approved protocol for recruiting people on the same Prolific platform as this project. For the past three years, he has also worked with Dr. Martin Katzman, from the START mood disorder clinic, on a research project that is trying to measure mental health state based on data collected from a smartphone. That project provides the initial motivation for the present project.
Dr. Ludovic Rheault is an Assistant Professor in the Department of Political Science at the University of Toronto (St. George Campus), Canada. He was previously a postdoctoral fellow at the University of California, Riverside during the academic year 2013-2014. He obtained his Ph.D. in 2013 from the University of Montreal in Canada, and teaches primarily in the areas of political methodology and Canadian politics. His primary research interest is the application and development of methods from the fields of natural language processing and machine learning to address social science questions.
Dr. Bill Simpson is an Adjunct Clinical Professor at McMaster University and Director of Clinical Operations at Winterlight Labs. He obtained his Bachelor's degree in Psychology (2007) and his PhD in Neuroscience from McMaster University (2016). His doctoral research centred around the relationship between circadian rhythms, inflammatory biomarkers and the development of postpartum depression in at-risk women. As a Research Associate between his Bachelor's and PhD, he worked on large clinical trials and observational cohort studies for Social Anxiety, Obsessive Compulsive Disorder and Generalized Anxiety Disorder. His work has focused extensively on how digital technology can improve the assessment, measurement and treatment of psychiatric conditions. Since 2011, he has been heavily involved in digital health startups and in his current position at Winterlight Labs he oversees the use of Winterlight's speech and language analysis platform for measuring the severity of neurodegenerative and psychiatric disease in clinical trials. He is a member of the Society for Biological Psychiatry, advisory board member for multiple Master's level programs in digital health and technology, the author of 18 peer-reviewed publications.

Are community members collecting and/or analyzing data? (•) Yes ( ) No

Please describe the community members research team status (eg. employees, volunteers, or participants). What training will they received?

We will be collaborating with Dr. Bill Simpson of Winterlight Labs to gain experience in analyzing the data that we will be collecting. None of the researchers will receive training.

8 - Possible Risks and Benefits

Possible Risks

Potential Risk Details:

Physical Risks	( ) Yes	(•) No
Psychological/emotional Risks	( ) Yes	(•) No
Social Risk	( ) Yes	(•) No
Legal Risk	( ) Yes	(•) No

Potential Benefits

Benefit Description

This research, in its long-term form, has the potential to develop an objective metric for detecting the level of anxiety on an individual. Being able to detect and measure the level of anxiety can benefit the community of people suffering from mental health disorders, specifically the ones related to anxiety.

9 - Consent

Consent Process Details

Prospective subjects who choose to read about the study on the Prolific study listing page will be taken to a study description web page, hosted by Prolific. This study description contains a brief, one-paragraph description of the study followed by the consent form itself. Please note that we understand that the UofT guidelines for consent form are that it include the U of T logo/letterhead on the consent form However, since the Prolific website only allows us to enter in text, without images, the study description and consent guide as hosted by Prolific cannot be made to display the U of T logo/letterhead. Instead we make it clear that the project is being done by the University of Toronto in the text. The consent document is attached.

Uploaded letter/consent form(s)

Document Title	Document Date

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Document Title	Document Date
Consent Form	2021-05-02

Is there additional documentation regarding consent such as screening materials, introductory letters etc.: ◯ Yes ◉ No

Uploaded letter/consent form(s)

Will any information collected in the screening process - prior to full informed consent to participate in the study - be retained for those who are later excluded or refuse to participate in the study? ◯ Yes ◉ No

Is the research taking place within a community or organization which requires formal consent be sought prior to the involvement of the individual participants ◯ Yes ◉ No

Are any participants not capable (e.g.: children) of giving competent consent? ◯ Yes ◉ No

10 - Debriefing and Dissemination

DeBrief

Will deception or intentional non disclosure be used? ◯ Yes ◉ No

Will a written debrief be used? ◯ Yes ◉ No

Do participants/communities have the right to withdraw their data following the debrief? ◯ Yes ◉ No

Information Feed Back Details following completion of a participants participation in the project

n/a

Procedural details which allow participants to withdraw from the project

The participants can withdraw from the study at any time prior to completion of their participation. If they withdraw in this time period, their data will be automatically deleted. Once their participation in the study is complete, they will have a 24 hour grace period to request the removal of their data. This can be done through the Prolific website and maintains the participants anonymity.

☐ Not Applicable

What happens to a participants data and any known consequences related to the removal of said participant

For anyone that does not complete the study, their partial data will be deleted.

☐ Not Applicable

List reasons why a participant can not withdraw from the project (either at all or after a certain period of time)

Once the participation is complete, participants will have a 24 hour grace period to request the removal of their data. If they didn't request for the removal of their data during this period, we will not allow the participant to retroactively have their data deleted. This is made clear on the consent form.

☐ Not Applicable

11 - Confidentiality and Privacy

Confidentiality

Is the data confidential? ◉ Yes ◯ No

Will the confidentiality of the participants and/or informants be protected? ◉ Yes ◯ No

List confidentiality protection procedures

There will be no identifiable information collected from Prolific (our participant recruitment platform), but a video recording will be made of each participant, during the participation. The video will contain their face, which is personally identifiable. The video recording will be stored in an encrypted storage which will be accessible only to the researchers involved in this study and be shared with Winterlight Labs. Privacy will be protected as described in the data sharing agreement with Winterlight.

Are there any limitations on the protection of participant confidentiality? ◯ Yes ◉ No

Is participant anonymity/confidentiality not applicable to this research project? ◯ Yes ◉ No

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Data Protection

Describe how the data (including written records, video/audio recordings, artifacts and questionnaires) will be protected during the conduct of the research and subsequent dissemination of results

All data will be encrypted in transit from the participant's computers to the our servers, and all data on backend servers will be encrypted at rest, accessible only to research staff. Only aggregate data will be disseminated.

Explain for how long, where and what format (identifiable, de-identified) data will be retained. Provide details of their destruction and/or continued storage. Provide a justification if you intend to store identifiable data for an indefinite length of time. If regulatory requirements for data retention exists, please explain.

Data will be retained indefinitely to enable future analyses of the data. A key research question of interest to the investigators is how different analysis techniques may provide different results. Retaining the data indefinitely enables the data to be used in future studies which may attempt to apply different analysis techniques.

Will the data be shared with other researchers or users? (●) Yes ( ) No

Please describe how and where the data will be stored and any restrictions that will be made regarding access.How will participant consent be obtained? If data is to be made open access, please describe how and where they will be maintained.

Data will only be shared with Winterlight Labs. De-identified data (aggregate numerical values obtained from Winterlight's software) from this population may be shared with government agencies, researchers, and other third parties. De-identified data may also be shared so Winterlight can obtain market approval for new products resulting from this study. Winterlight will not share the raw audio recordings or transcripts with any third party (as detailed in the data transfer agreement), though these data will be held internally by Winterlight for continued development, research and commercial use.

12 - Level of Risk and Research Ethics Board

Level of Risk for the Project

Group Vulnerability: Low

Research Risk: Low

Risk Level: 1

Explanation/Justification

Explanation/Justification detail for the group vulnerabilty and research risk listed above

The group being recruited on the Prolific platform is a very general group, with no particular risk profile. One of the questions we are going to ask to stimulate audio speech is intentionally neutral, and hence the risk is low. The other question follows the Trier Social Stress Task (TSST), which induces moderate stress but is not any more than that experienced everyday lives, and hence the risk is low.

Research Ethics Board

REB Associated with this project: Health Sciences

13 - Application Documents Summary

Uploaded Documents

Document Title	Document Date
Draft Data Sharing Agreement	2020-07-31
Response To Reviewers Comments on September21_20	2020-09-30
Response to Reviewer	2021-05-09
Notice of Award	2018-06-24
GAD 7 Survey	2019-03-17
Stimulus Statements and Questions	2019-04-17

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Document Title	Document Date
Post audio anxiety level questionnaire	2020-09-25
The Version of the Trier Social Stress Task Used	2020-09-30
PHQ-8 questionnaire	2021-04-08
Consent Form	2021-05-02

14 - Applicant Undertaking

I confirm that I am aware of, understand, and will comply with all relevant laws governing the collection and use of personal identifiable information is research. I understand that for research involving extraction or collection of personally identifiable information, provincial, federal, and/or international laws may apply and that any apparent mishandling of said personally identifiable information, must be reported to the office of research ethics.

As the Principal Investigator of the project, I confirm that I will ensure that all procedures performed in accordance with all relevant university, provincial, national, and/or international policies and regulations that govern research with human participants. I understand that if there is any significant deviation in the project as originally approved, I must submit an amendment to the Research Ethics Board for approval prior to implementing any change.

☒ I have read and agree to the above conditions

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

UNIVERSITY OF TORONTO

OFFICE OF THE VICE-PRESIDENT, RESEARCH AND INNOVATION

RIS Protocol Number: 37584

Approval Date: 10-May-21

PI Name: Prof Jonathan Rose

Division Name:

Dear Prof Jonathan Rose:

Re: Your research protocol application entitled, "Measuring Anxiety in Speech"

The Health Sciences REB has conducted a Delegated review of your application and has granted approval to the attached protocol for the period 2021-05-10 to 2022-03-14.

This approval covers the ethical acceptability of the human research activity; please ensure that all other approvals required to conduct your research are obtained prior to commencing the activity.

Please be reminded of the following points:

- An **Amendment** must be submitted to the REB for any proposed changes to the approved protocol. The amended protocol must be reviewed and approved by the REB prior to implementation of the changes.
- An annual **Renewal** must be submitted for ongoing research. Renewals should be submitted between 15 and 30 days prior to the current expiry date.
- A **Protocol Deviation Report** (PDR) should be submitted when there is any departure from the REB-approved ethics review application form that has occurred without prior approval from the REB (e.g., changes to the study procedures, consent process, data protection measures). The submission of this form does not necessarily indicate wrong-doing; however follow-up procedures may be required.
- An **Adverse Events Report (AER)** must be submitted when adverse or unanticipated events occur to participants in the course of the research process.
- A **Protocol Completion Report** (PCR) is required when research using the protocol has been completed.
- If your research is funded by a third party, please contact the assigned Research Funding Officer in Research Services to ensure that your funds are released.

Best wishes for the successful completion of your research.

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Appendix H

Participants Recruitment Consent Form

Measuring Anxiety in Speech and Video

In this study, you will be asked to read a statement and answer questions using **an active webcam and microphone**. You will complete several tasks, each of which will require you to record your voice and face for about 2-5 minutes.

First, you will be invited to fill out a short questionnaire that asks you about your mood in the past two weeks. Next, you will be asked to answer questions that require you to speak out loud into a microphone with an active webcam recording your face as you speak. Finally, you will be asked a few questions about how you felt during the recording session.

The remainder of this description contains a Consent Guide with more details about the study. Please read the Consent Guide and only participate in this study if you agree to record your voice and face using a webcam.

<u>Consent Guide</u>

INFORMED CONSENT

You are invited to participate in a research study conducted by a group of researchers from the University of Toronto and Winterlight Labs.

This guide explains the purpose of the research study, provides information about the procedures involved, possible risks and benefits, and the rights of participants.

Please read this guide carefully. Your participation in this study is entirely voluntary – you do not have to take part. You should only agree to the study if you agree to the terms of the study described in this guide. If you do not agree with any of the terms, please do not enroll in the study.

INTRODUCTION AND GOAL

The goal of this study is to see if it is possible to develop an automated method for detecting anxiety in speech. Our goal is to find better ways to diagnose issues surrounding anxiety and to detect if treatments are working. The first step towards this goal is to gather data in the form of speech, video and questionnaire-based measures of anxiety.

WHAT WILL HAPPEN DURING THIS STUDY?

As a participant, you will be asked to do the following:

1) Fill out a questionnaire about your mood and feelings over the previous two weeks.
2) Read a neutral statement – i.e. a passage of text that is not designed to evoke any particular emotion. You will be asked to enable the microphone and camera on your computer so that your voice and face can be recorded.
3) Give a speech to convince an interviewer that you are the right person for your ‘dream’ job. You will be expected to speak for 5 minutes, and recorded as above.

4) Rate your own level of anxiety.

All the above procedures are administered using a web application. All data collected will be recorded and stored securely on our server.

Your recordings will not be analyzed in real-time. The researchers will not have access to your name or phone number and, therefore, will not be able to provide assistance if you report a moment of crisis or thoughts/intent of self-harm while completing the speech assessment.

WHAT ARE THE RESPONSIBILITIES OF STUDY PARTICIPANTS?
Your computer must have a microphone and a webcam that can record your voice and the video of your face. You will be asked to grant permission for the web browser to start recording.

You must be willing to respond with honesty to the questionnaire provided in the study.

STUDY TIMELINE AND ACTIVITIES
The study begins with a short questionnaire that should take about 1 minute. Next, you will be asked to read a short statement/story out loud using an active microphone and webcam, which will also take about 1 or 2 minutes. Then, we ask you to think about what your 'dream' job would be and to provide reasons why you should be hired for that job. You will have time to prepare and five minutes to speak. Finally, you will be asked to rate your own level of anxiety, twice each after each task, which will take less than 1 minute.

In total, this study should take between 9 and 10 minutes.

WHAT ARE THE RISKS OR HARM OF PARTICIPATING IN THIS STUDY?
This study requires you to engage in a public speaking task, which some participants find stressful. However, this task is designed to invoke similar stress levels to those experienced by people in their everyday lives when speaking in meetings, teach in front of students, deliver presentations, and so on. There are no other risks or harm to you as a result of your participation in this study.

WHAT ARE THE BENEFITS OF PARTICIPATING IN THIS STUDY?
There is no direct benefit to you, the participant, for participating in this study. However, your participation in this research will contribute to the development of technologies that can help people with anxiety monitor and improve their condition.

WHAT ARE THE COSTS OF PARTICIPATING IN THIS STUDY?
There are no costs to you as a result of your participation in this study.

HOW WILL MY PRIVACY BE RESPECTED?

All data recorded during the study will be stored on an encrypted server. The only information associated with your identity is your voice and video recording. The video files will be stored securely and accessible only to the University of Toronto researchers involved in this study.

The audio data from the video files will be shared with the company Winterlight Labs, for the purposes of analysis and improvement of methods of detection of anxiety. Your speech recordings (and not the video) will be sent to Winterlight's Canadian servers over an encrypted internet connection. Staff at Winterlight will have access to your recordings to transcribe them as well as check them for quality issues, including background noise or microphone problems.

The information collected in this study will be used to understand the relationship between speech and anxiety. Winterlight will use the anonymous speech data you provide to develop digital biomarkers for psychiatric disease commercially. These biomarkers and your data will be used by Winterlight for research, commercial and development purposes and will be retained by Winterlight indefinitely.

Once received by Winterlight, a team of human transcriptionists will transcribe your data. Winterlight will then use its software to analyze both the audio recording and the transcript generated from your recording. This analysis will produce a set of numerical values that describe many aspects of your speech. It is not possible to reconstruct either the audio recording or the transcript of your sample from these numerical values alone.

De-identified data (aggregate numerical values obtained from Winterlight's software) from this population may be shared with government agencies, researchers, and other third parties. De-identified data may also be shared so that Winterlight can obtain market approval for new products resulting from this study. Winterlight will not share the raw audio recordings or transcripts with any third party (as detailed in the data transfer agreement). However, these data will be held internally by Winterlight for continued development, research and commercial use.

The results of this research study may be presented at meetings or in publications, but your identity will not be disclosed.

A complete and up to date version of Winterlight's privacy and security policy is available on their website: winterlightlabs.com/docs/WLL_Privacy_Notice.pdf

HOW WILL MY INFORMATION BE KEPT CONFIDENTIAL?
The investigators and researchers will keep the information they see or receive about you confidential, to the extent permitted by applicable laws. As some electronic data is stored on computers on American soil, your data could be accessed by the US government as per the PATRIOT act.

Only the audio portion of the video recording will be shared with Winterlight Labs. This audio recording will be stored on an encrypted server, located in Canada, and only designated members of the study team (from Winterlight and the University of Toronto) will have access to the recording.

The data collected during this trial will be stored indefinitely, with no set timeline for deletion. **If you withdraw from the study before finishing it, your data will be deleted.**

Note that once your participation in the study is complete, you can request the removal of your data within 24 hours of completion. To do so, send a message through the Prolific messaging system as follows: go to the submissions tab of your Prolific webpage, select this study ("Measuring Anxiety from Speech and Video") and click on "Contact researcher" to send the message

Data collected during this study will be used to conduct research, and the findings may be published in a scientific journal. Any use of the study's data for a publication will preserve the anonymity of participants, and none of the video recordings we collect will ever be included in a research publication.

By signing this consent form, you consent to the collection, access, use and disclosure of your information as described above, and to conduct this research study.

You may request a summary of any research publication that arises as a result of this study. Please contact Professor Jonathan Rose at jonathan.rose@ece.utoronto.ca if you wish to be given access to such publication if they are made in the future.

DO THE INVESTIGATORS HAVE ANY CONFLICT OF INTEREST?
There are no conflicts of interest to declare related to this study.

WHAT ARE MY RIGHTS AS A STUDY PARTICIPANT?
You do not waive your legal rights by participating in this study. You have the right to ask questions and receive answers throughout this study.

This study has been reviewed by the University of Toronto Research Oversight and Compliance Office. You may also contact the University of Toronto Research Oversight and Compliance Office – Human Research Ethics Program if you have questions about your rights as a participant:

- By e-mail: ethics.review@utoronto.ca
- By telephone: +1 416-946-3273

QUALITY ASSURANCE
The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the University of Toronto Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team.

THIS CONCLUDES THE CONSENT GUIDE. PLEASE ONLY PARTICIPATE IN THIS STUDY IF YOU AGREE WITH ALL THE TERMS STATED ABOVE.

Appendix I

Correlation Between Demographics and Acoustic and Linguistic features.

Correlation between Demographics and acoustic/Linguistic features

Significant correlations (*P*<.05) are reported.

Age vs Acoustic/Linguistic features

Acoustic		
Feature	**r**	***P***
localabsoluteJitter	0.27	<.001
mfcc_mean_6	0.13	<.001
mfcc_std_6	-0.13	<.001
mfcc_mean_11	0.13	<.001
f0_median	-0.12	<.001
lpcc_std_3	0.12	<.001
f0_mean	-0.11	<.001
lpcc_std_2	-0.11	<.001
mfcc_mean_8	0.10	<.001
intensity_std	0.10	<.001
mfcc_std_3	-0.09	<.001
lpcc_std_10	0.09	<.001
lpcc_std_8	-0.08	0.001
f1_mean	-0.07	0.004
lpcc_mean_4	-0.07	0.005
lpcc_std_13	0.07	0.006
mfcc_mean_10	0.06	0.009
f3_stdev	-0.06	0.01
intensity_mean	0.06	0.01
mfcc_std_5	0.06	0.01
mfcc_mean_3	0.06	0.01
f2_stdev	-0.06	0.01
f0_min	-0.06	0.02
lpcc_std_9	-0.06	0.02
mfcc_std_12	0.06	0.02
lpcc_std_11	0.06	0.02
mfcc_std_11	0.05	0.02
f2_mean	-0.05	0.04
speaking_duration	-0.05	0.04
lpcc_std_5	0.05	0.04
f1_stdev	0.05	0.046
mfcc_std_8	-0.05	0.047

Linguistic		
Feature	**r**	***P***
compare	-0.18	<.001
insight	-0.16	<.001
conj	-0.15	<.001
see	0.14	<.001
cause	-0.14	<.001
sexual	-0.14	<.001
motion	0.13	<.001
relativ	0.13	<.001
discrep	0.13	<.001
article	0.13	<.001
they	0.12	<.001
achieve	-0.12	<.001
adverb	-0.12	<.001
focuspast	0.12	<.001
i	-0.11	<.001
work	-0.10	<.001
adj	-0.10	<.001
sad	0.10	<.001
nonflu	-0.10	<.001
number	0.10	<.001
informal	-0.10	<.001
home	0.10	<.001
social	0.10	<.001
risk	0.10	<.001
negate	0.09	<.001
verb	0.09	<.001
space	0.09	<.001
cogproc	-0.09	<.001
filler	0.09	<.001
leisure	0.08	<.001
ipron	-0.08	<.001
ingest	0.08	<.001
time	0.08	<.001
WC	-0.08	<.001
Sixltr	-0.08	0.001

percept	0.08	0.001
drives	-0.07	0.002
auxverb	0.07	0.002
quant	0.07	0.002
pronoun	-0.07	0.005
Period	0.07	0.005
you	0.06	0.006
negemo	0.06	0.006
family	0.06	0.009
power	-0.06	0.01
money	0.06	0.01
we	0.06	0.02
male	0.05	0.03
relig	0.05	0.03

Income vs Acoustic/Linguistic features

Acoustic		
Feature	**r**	***P***
f0_median	-0.22	<.001
f0_mean	-0.22	<.001
mfcc_std_5	0.20	<.001
mfcc_mean_2	0.14	<.001
localabsoluteJitter	0.13	<.001
mfcc_std_11	0.13	<.001
lpcc_std_7	0.12	<.001
speaking_duration	0.10	<.001
mfcc_std_2	0.09	<.001
mfcc_mean_1	0.08	<.001
f3_mean	-0.08	<.001
lpcc_std_3	0.08	0.001
f2_mean	-0.07	0.002
mfcc_mean_6	0.07	0.002
mfcc_std_4	0.07	0.002
f4_stdev	-0.07	0.003
mfcc_mean_4	0.07	0.004
lpcc_mean_4	-0.07	0.004
f2_stdev	-0.07	0.004
f4_mean	-0.06	0.01
f3_stdev	-0.06	0.02
f0_max	-0.06	0.02

Linguistic		
Feature	**r**	***P***
space	0.13	<.001
i	-0.13	<.001
relativ	0.13	<.001
we	0.12	<.001
WPS	0.11	<.001
negemo	-0.11	<.001
Period	-0.11	<.001
affect	-0.11	<.001
AllPunc	-0.11	<.001
article	0.11	<.001
anger	-0.10	<.001
anx	-0.10	<.001
motion	0.10	<.001
WC	0.09	<.001
ppron	-0.09	<.001
family	-0.08	<.001
female	-0.07	0.002
money	0.07	0.003
feel	-0.07	0.004
posemo	-0.07	0.005
hear	-0.06	0.007
negate	-0.06	0.009

f1_mean	-0.05	0.03
lpcc_std_12	0.05	0.03
mfcc_mean_12	0.05	0.03
lpcc_mean_5	-0.05	0.04
mfcc_mean_8	0.05	0.04

assent	-0.06	0.01
you	0.06	0.02
conj	-0.06	0.02
pronoun	-0.05	0.03
certain	-0.05	0.04
leisure	0.05	0.04
Apostro	-0.05	0.04
relig	-0.05	0.04

Appendix J

Significant Feature Inter-correlations of the All-sample Set

Correlations color coding
> 0.3
> 0.6
> 0.8

	AllPunc	WC	speaking_duration	Period	assent
AllPunc					
WC	-0.70				
speaking_duration	-0.69	0.78			
Period	0.93	-0.72	-0.72		
assent	0.47	-0.44	-0.49	0.43	
negemo	0.29	-0.26	-0.36	0.25	0.26
relativ	-0.26	0.20	0.20	-0.20	-0.18
motion	-0.20	0.19	0.12	-0.16	-0.10
Shimmer	0.42	-0.52	-0.71	0.47	0.33
swear	0.12	-0.06	-0.13	0.10	0.10
anger	0.30	-0.29	-0.37	0.25	0.26
mfcc_std_2	-0.26	0.31	0.38	-0.25	-0.31
mfcc_std_3	-0.53	0.49	0.67	-0.57	-0.30
focusfuture	-0.11	0.17	0.04	-0.09	-0.08
mfcc_mean_2	-0.29	0.45	0.42	-0.31	-0.21
adverb	-0.07	0.20	0.14	-0.14	0.04
time	-0.07	0.09	0.03	-0.05	-0.03
function	-0.24	0.33	0.14	-0.21	-0.25
negate	0.32	-0.17	-0.38	0.24	0.28
prep	-0.39	0.24	0.32	-0.30	-0.39
WPS	-0.61	0.53	0.53	-0.66	-0.23
anx	0.16	-0.16	-0.21	0.13	0.13
f0_std	0.05	-0.04	-0.08	0.01	0.16
hear	0.10	-0.06	-0.09	0.06	0.08
mfcc_std_5	-0.39	0.47	0.50	-0.39	-0.27
death	0.02	-0.01	-0.01	0.03	-0.02
ipron	-0.16	0.31	0.17	-0.19	-0.15
see	-0.08	0.10	0.07	-0.06	-0.07
affect	0.32	-0.32	-0.38	0.30	0.34
i	0.30	-0.29	-0.33	0.24	0.20
family	0.12	-0.14	-0.17	0.13	0.09
mfcc_std_4	-0.49	0.47	0.60	-0.50	-0.34
sad	0.12	-0.12	-0.18	0.10	0.13
ppron	0.25	-0.16	-0.30	0.18	0.19
space	-0.24	0.15	0.22	-0.18	-0.21
article	-0.22	0.17	0.19	-0.12	-0.22
leisure	0.05	-0.09	-0.11	0.07	0.07
friend	0.16	-0.18	-0.21	0.14	0.19

	negemo	relativ	motion	Shimmer	swear	anger
relativ	-0.18					
motion	-0.13	0.49				
Shimmer	0.27	-0.15	-0.09			
swear	0.20	-0.04	-0.01	0.09		
anger	0.70	-0.20	-0.14	0.27	0.22	
mfcc_std_2	-0.18	0.16	0.11	-0.24	-0.04	-0.17
mfcc_std_3	-0.24	0.09	0.07	-0.47	-0.11	-0.24
focusfuture	-0.04	0.12	0.24	-0.02	0.04	-0.04
mfcc_mean_2	-0.19	0.13	0.08	-0.34	-0.03	-0.20
adverb	-0.04	0.00	-0.01	-0.12	0.02	-0.03
time	-0.01	0.63	0.14	-0.05	0.01	-0.04
function	-0.03	-0.12	0.04	-0.09	-0.03	-0.06
negate	0.51	-0.16	-0.02	0.26	0.15	0.48
prep	-0.31	0.35	0.12	-0.18	-0.14	-0.33
WPS	-0.18	0.12	0.08	-0.30	-0.06	-0.16
anx	0.56	-0.10	-0.10	0.14	0.02	0.22
f0_std	0.12	-0.03	0.01	0.02	0.06	0.13
hear	0.05	-0.08	-0.06	0.05	0.01	0.06
mfcc_std_5	-0.17	0.21	0.14	-0.34	0.00	-0.17
death	0.12	0.01	-0.02	0.05	0.03	0.14
ipron	-0.03	-0.11	0.06	-0.12	-0.02	-0.08
see	-0.04	0.07	0.10	-0.03	0.04	-0.03
affect	0.46	-0.36	-0.14	0.25	0.08	0.37
i	0.22	-0.28	-0.19	0.22	0.08	0.26
family	0.17	-0.11	-0.06	0.16	0.12	0.17
mfcc_std_4	-0.23	0.12	0.07	-0.41	-0.07	-0.21
sad	0.51	-0.04	-0.03	0.13	0.10	0.19
ppron	0.23	-0.30	-0.09	0.20	0.09	0.24
space	-0.21	0.78	0.21	-0.15	-0.06	-0.21
article	-0.22	0.28	0.09	-0.12	-0.04	-0.21
leisure	0.02	-0.02	0.05	0.10	0.08	0.11
friend	0.14	-0.08	-0.02	0.17	0.03	0.19

	mfcc_std_2	mfcc_std_3	focusfuture	mfcc_mean_	adverb	time
mfcc_std_3	0.42					
focusfuture	0.05	0.00				
mfcc_mean_2	0.15	0.11	0.04			
adverb	0.03	0.14	-0.01	0.03		
time	0.02	0.00	0.12	0.04	0.09	
function	0.06	0.10	0.23	0.05	0.31	-0.07
negate	-0.18	-0.27	0.09	-0.16	0.08	0.04

	mfcc_std_2	mfcc_std_3	focusfuture	mfcc_mean_	adverb	time
prep	0.20	0.19	-0.02	0.11	-0.13	-0.02
WPS	0.11	0.42	0.01	0.22	0.12	0.06
anx	-0.10	-0.13	-0.02	-0.13	-0.02	0.03
f0_std	-0.04	0.02	-0.03	0.01	0.04	-0.01
hear	-0.02	-0.08	-0.01	-0.06	0.07	0.03
mfcc_std_5	0.51	0.46	0.09	0.27	0.03	0.09
death	-0.02	-0.03	0.00	0.01	-0.03	0.00
ipron	0.09	0.10	0.14	0.12	0.11	-0.07
see	0.02	0.05	0.11	0.04	0.00	0.05
affect	-0.24	-0.20	-0.08	-0.21	0.02	-0.16
i	-0.17	-0.17	-0.05	-0.20	-0.01	-0.01
family	-0.12	-0.07	-0.02	-0.09	-0.02	-0.01
mfcc_std_4	0.48	0.74	0.01	0.15	0.06	0.02
sad	-0.10	-0.12	-0.03	-0.07	-0.02	0.02
ppron	-0.18	-0.18	0.08	-0.17	-0.02	-0.02
space	0.18	0.11	-0.01	0.13	-0.06	0.11
article	0.15	0.05	0.02	0.15	-0.18	0.07
leisure	-0.01	-0.07	-0.07	-0.05	0.02	-0.03
friend	-0.09	-0.13	-0.04	-0.10	-0.05	-0.02

	function	negate	prep	WPS	anx	f0_std
negate	0.21					
prep	0.01	-0.43				
WPS	0.10	-0.17	0.17			
anx	0.01	0.29	-0.15	-0.11		
f0_std	-0.04	0.12	-0.10	-0.01	0.07	
hear	0.04	0.16	-0.13	-0.06	0.10	0.03
mfcc_std_5	0.06	-0.15	0.16	0.29	-0.15	0.05
death	-0.05	0.01	-0.02	-0.04	0.00	-0.01
ipron	0.57	0.12	-0.08	0.11	-0.03	0.04
see	0.09	0.02	-0.01	0.03	-0.01	-0.02
affect	-0.04	0.26	-0.31	-0.19	0.29	0.11
i	0.21	0.26	-0.37	-0.17	0.18	0.02
family	-0.02	0.12	-0.18	-0.07	0.10	0.03
mfcc_std_4	0.11	-0.24	0.21	0.37	-0.16	0.03
sad	-0.03	0.22	-0.14	-0.07	0.15	0.05
ppron	0.36	0.33	-0.45	-0.13	0.17	0.03
space	-0.14	-0.26	0.49	0.11	-0.14	-0.04
article	-0.06	-0.30	0.25	0.07	-0.16	-0.09
leisure	-0.09	0.07	-0.09	-0.07	0.00	0.02
friend	-0.10	0.14	-0.16	-0.06	0.08	0.05

	hear	mfcc_std_5	death	ipron	see	affect
mfcc_std_5	-0.07					
death	0.00	0.00				
ipron	0.00	0.13	-0.04			
see	0.02	0.06	0.05	0.07		
affect	0.07	-0.27	-0.01	-0.10	-0.03	
i	0.14	-0.26	-0.05	-0.16	-0.08	0.35
family	0.01	-0.12	0.05	-0.08	0.02	0.17
mfcc_std_4	-0.09	0.64	-0.03	0.13	0.03	-0.24
sad	-0.01	-0.08	0.06	-0.02	-0.02	0.21
ppron	0.15	-0.21	-0.03	-0.04	0.04	0.33
space	-0.12	0.18	0.02	-0.13	0.04	-0.35
article	-0.11	0.17	0.05	-0.11	0.04	-0.29
leisure	0.25	0.00	0.04	-0.15	0.11	0.14
friend	0.07	-0.13	0.00	-0.14	-0.05	0.22

	i	family	mfcc_std_4	sad	ppron	space
family	0.13					
mfcc_std_4	-0.20	-0.09				
sad	0.07	0.09	-0.11			
ppron	0.81	0.17	-0.20	0.11		
space	-0.32	-0.14	0.14	-0.07	-0.39	
article	-0.38	-0.11	0.08	-0.07	-0.43	0.32
leisure	0.01	0.14	-0.07	0.02	-0.01	-0.02
friend	0.18	0.20	-0.12	0.08	0.19	-0.11

	article	leisure	friend
leisure	0.07		
friend	-0.13	0.10	

Appendix K

Significant Feature Inter-correlations of the Female-samples

	Period	AllPunc	WC	speaking_duration	adverb
Period					
AllPunc	0.93				
WC	-0.73	-0.70			
speaking_duration	-0.71	-0.69	0.80		
adverb	-0.15	-0.07	0.21	0.14	
negemo	0.26	0.29	-0.27	-0.38	-0.08
anger	0.27	0.31	-0.30	-0.38	-0.02
mfcc_std_3	-0.58	-0.56	0.55	0.71	0.14
motion	-0.18	-0.20	0.17	0.10	0.00
Shimmer	0.45	0.42	-0.55	-0.73	-0.11
assent	0.41	0.43	-0.43	-0.49	0.00
see	-0.12	-0.13	0.13	0.11	0.01
relativ	-0.21	-0.25	0.16	0.19	0.02
lpcc_std_6	-0.45	-0.41	0.38	0.45	0.05
lpcc_std_4	-0.47	-0.45	0.41	0.46	0.10
mfcc_mean_2	-0.36	-0.33	0.47	0.44	0.06
intensity_mean	-0.48	-0.46	0.36	0.47	0.05
mfcc_mean_1	-0.63	-0.58	0.50	0.61	0.05
sad	0.10	0.13	-0.14	-0.21	-0.05
Dic	0.06	0.13	-0.01	-0.17	0.17
power	0.00	-0.06	0.01	0.07	-0.17
WPS	-0.62	-0.59	0.50	0.51	0.12
lpcc_std_10	-0.39	-0.39	0.37	0.40	0.00
intensity_std	-0.33	-0.31	0.19	0.23	0.02
lpcc_std_12	-0.41	-0.38	0.43	0.43	0.08
death	0.07	0.05	-0.02	-0.05	-0.04
mfcc_mean_8	0.29	0.27	-0.36	-0.38	-0.10
percept	-0.03	-0.03	0.00	0.02	-0.03
lpcc_mean_4	0.40	0.40	-0.34	-0.45	-0.05

Correlations color coding	
> 0.3	
> 0.6	
> 0.8	

	negemo	anger	mfcc_std_3	motion	Shimmer
anger	0.70				
mfcc_std_3	-0.27	-0.25			
motion	-0.12	-0.12	0.11		
Shimmer	0.30	0.29	-0.49	-0.09	
assent	0.23	0.26	-0.34	-0.08	0.33
see	-0.07	-0.05	0.10	0.07	-0.03
relativ	-0.19	-0.20	0.16	0.48	-0.16
lpcc_std_6	-0.18	-0.18	0.42	0.06	-0.34

	negemo	anger	mfcc_std_3	motion	Shimmer
lpcc_std_4	-0.17	-0.15	0.76	0.10	-0.29
mfcc_mean_2	-0.21	-0.21	0.19	0.07	-0.40
intensity_mean	-0.13	-0.14	0.42	0.09	-0.42
mfcc_mean_1	-0.19	-0.19	0.51	0.09	-0.49
sad	0.53	0.21	-0.14	-0.02	0.18
Dic	0.14	0.15	-0.10	0.04	0.09
power	-0.08	-0.09	0.02	-0.03	-0.04
WPS	-0.18	-0.15	0.44	0.08	-0.29
lpcc_std_10	-0.14	-0.15	0.37	0.09	-0.27
intensity_std	-0.04	-0.05	0.24	0.07	-0.26
lpcc_std_12	-0.16	-0.14	0.42	0.07	-0.34
death	0.16	0.16	-0.07	0.00	0.11
mfcc_mean_8	0.10	0.10	-0.39	-0.01	0.26
percept	-0.01	-0.04	0.03	-0.02	-0.02
lpcc_mean_4	0.16	0.14	-0.59	-0.07	0.31

	assent	see	relativ	lpcc_std_6	lpcc_std_4
see	-0.08				
relativ	-0.14	0.03			
lpcc_std_6	-0.25	0.09	0.10		
lpcc_std_4	-0.27	0.08	0.13	0.47	
mfcc_mean_2	-0.22	0.05	0.10	0.19	0.20
intensity_mean	-0.22	0.08	0.13	0.54	0.57
mfcc_mean_1	-0.29	0.08	0.15	0.62	0.63
sad	0.13	0.00	-0.03	-0.08	-0.07
Dic	0.08	0.01	-0.20	-0.07	-0.05
power	-0.03	-0.13	0.08	0.00	0.00
WPS	-0.22	0.06	0.13	0.29	0.36
lpcc_std_10	-0.28	0.07	0.10	0.59	0.60
intensity_std	-0.11	0.06	0.09	0.48	0.47
lpcc_std_12	-0.31	0.10	0.10	0.42	0.64
death	0.02	0.03	-0.03	-0.03	-0.05
mfcc_mean_8	0.22	-0.07	-0.10	-0.02	-0.36
percept	-0.06	0.51	-0.10	0.02	0.01
lpcc_mean_4	0.24	-0.08	-0.11	-0.50	-0.75

	mfcc_mean_	intensity_me	mfcc_mean_1	sad	Dic
intensity_mean	0.31				
mfcc_mean_1	0.28	0.77			
sad	-0.12	-0.05	-0.07		

	mfcc_mean_	intensity_me	mfcc_mean_	sad	Dic
Dic	-0.07	-0.10	-0.10	0.03	
power	0.03	0.01	0.00	-0.07	-0.04
WPS	0.25	0.38	0.43	-0.08	-0.06
lpcc_std_10	0.35	0.59	0.57	-0.03	-0.08
intensity_std	0.21	0.93	0.68	0.01	-0.06
lpcc_std_12	0.36	0.64	0.63	-0.06	-0.09
death	-0.02	-0.02	-0.02	0.09	-0.09
mfcc_mean_8	-0.10	-0.30	-0.28	0.06	0.04
percept	0.00	0.03	0.02	0.00	0.11
lpcc_mean_4	-0.12	-0.52	-0.55	0.08	0.11

	power	WPS	lpcc_std_10	intensity_stc	lpcc_std_12
WPS	0.04				
lpcc_std_10	0.02	0.23			
intensity_std	0.00	0.23	0.54		
lpcc_std_12	-0.01	0.27	0.57	0.57	
death	-0.02	-0.05	0.02	-0.01	-0.03
mfcc_mean_8	-0.03	-0.21	-0.07	-0.21	-0.52
percept	-0.14	0.03	-0.02	0.01	0.01
lpcc_mean_4	-0.02	-0.29	-0.60	-0.40	-0.63

	death	mfcc_mean_	percept	lpcc_mean_4
mfcc_mean_8	0.05			
percept	-0.02	-0.04		
lpcc_mean_4	0.06	0.33	-0.02	

Appendix L

Significant Feature Inter-correlations of the Male-samples

	speaking_duration	AllPunc	WC	assent	relativ
speaking_duration					
AllPunc	-0.69				
WC	0.78	-0.70			
assent	-0.50	0.50	-0.45		
relativ	0.24	-0.30	0.22	-0.22	
leisure	-0.09	0.03	-0.08	0.07	-0.07
hear	-0.08	0.08	-0.04	0.09	-0.07
swear	-0.15	0.13	-0.07	0.10	-0.05
time	0.07	-0.12	0.12	-0.07	0.64
Apostro	-0.26	0.34	-0.03	0.22	-0.18
mfcc_std_2	0.40	-0.30	0.30	-0.34	0.12
power	0.00	-0.08	-0.02	-0.09	0.13
ppron	-0.29	0.26	-0.13	0.18	-0.30
Sixltr	0.13	-0.17	-0.13	-0.18	0.14
mfcc_std_5	0.62	-0.49	0.51	-0.32	0.16
anx	-0.19	0.18	-0.16	0.11	-0.07
negate	-0.37	0.34	-0.19	0.33	-0.21
negemo	-0.35	0.29	-0.25	0.28	-0.16
article	0.21	-0.23	0.15	-0.23	0.25
mfcc_mean_5	0.08	-0.03	0.07	-0.02	0.06
Period	-0.72	0.93	-0.74	0.45	-0.23
prep	0.30	-0.41	0.22	-0.44	0.37
focusfuture	0.04	-0.08	0.16	-0.06	0.11
family	-0.23	0.18	-0.17	0.13	-0.09
f0_std	-0.10	0.05	-0.05	0.16	-0.03
mfcc_std_4	0.58	-0.47	0.44	-0.33	0.13
ipron	0.18	-0.16	0.32	-0.21	-0.07
Shimmer	-0.69	0.42	-0.52	0.34	-0.18
affect	-0.41	0.36	-0.34	0.40	-0.33
mfcc_std_11	0.55	-0.38	0.40	-0.25	0.14
f1_mean	-0.02	0.00	-0.01	0.05	-0.06
motion	0.16	-0.21	0.19	-0.11	0.49

Correlations color coding	
> 0.3	(green)
> 0.6	(yellow)
> 0.8	(red)

	leisure	hear	swear	time	Apostro
hear	0.23				
swear	0.11	0.05			
time	-0.04	0.04	0.01		
Apostro	0.02	0.14	0.17	0.00	
mfcc_std_2	-0.03	-0.03	-0.07	-0.02	-0.10

	leisure	hear	swear	time	Apostro
power	-0.13	-0.15	-0.07	0.02	-0.19
ppron	-0.03	0.13	0.13	-0.04	0.44
Sixltr	-0.12	-0.20	-0.13	-0.04	-0.44
mfcc_std_5	-0.04	-0.08	-0.05	0.04	-0.14
anx	0.03	0.07	0.07	0.07	0.20
negate	0.08	0.15	0.20	0.02	0.55
negemo	0.03	0.03	0.26	0.05	0.32
article	0.04	-0.10	-0.06	0.06	-0.29
mfcc_mean_5	0.05	-0.02	-0.05	0.06	-0.01
Period	0.05	0.05	0.10	-0.10	0.11
prep	-0.09	-0.09	-0.19	0.04	-0.35
focusfuture	-0.11	-0.03	0.04	0.12	0.11
family	0.18	0.00	0.19	-0.02	0.10
f0_std	0.01	0.05	0.08	0.00	0.10
mfcc_std_4	-0.04	-0.08	-0.10	0.01	-0.19
ipron	-0.15	-0.03	-0.04	-0.05	0.19
Shimmer	0.08	0.06	0.11	-0.08	0.16
affect	0.18	0.05	0.14	-0.14	0.25
mfcc_std_11	-0.01	-0.06	-0.05	0.04	-0.15
f1_mean	0.01	0.01	0.05	0.00	0.07
motion	-0.04	-0.07	-0.03	0.15	-0.01

	mfcc_std_2	power	ppron	Sixltr	mfcc_std_5
power	0.03				
ppron	-0.12	-0.13			
Sixltr	0.09	0.29	-0.45		
mfcc_std_5	0.49	-0.02	-0.20	0.08	
anx	-0.06	0.00	0.14	-0.12	-0.12
negate	-0.19	-0.15	0.31	-0.35	-0.21
negemo	-0.20	-0.02	0.22	-0.13	-0.19
article	0.09	0.16	-0.43	0.18	0.10
mfcc_mean_5	-0.16	-0.01	-0.07	-0.01	-0.19
Period	-0.30	-0.01	0.17	-0.05	-0.52
prep	0.19	0.16	-0.42	0.31	0.18
focusfuture	0.03	-0.03	0.11	-0.14	0.04
family	-0.08	0.02	0.17	-0.13	-0.10
f0_std	-0.01	-0.01	0.01	-0.04	0.13
mfcc_std_4	0.51	-0.02	-0.20	0.10	0.76
ipron	0.10	-0.15	-0.05	-0.23	0.15
Shimmer	-0.28	-0.01	0.18	-0.08	-0.44
affect	-0.24	0.02	0.27	-0.14	-0.28

	mfcc_std_2	power	ppron	Sixltr	mfcc_std_5
mfcc_std_11	0.31	0.02	-0.20	0.08	0.65
f1_mean	0.02	0.00	0.01	-0.03	0.14
motion	0.10	-0.02	-0.09	-0.04	0.11

	anx	negate	negemo	article	mfcc_mean_5
negate	0.30				
negemo	0.57	0.51			
article	-0.15	-0.33	-0.22		
mfcc_mean_5	-0.03	-0.04	-0.04	0.03	
Period	0.15	0.26	0.25	-0.13	-0.01
prep	-0.14	-0.45	-0.35	0.26	0.03
focusfuture	-0.03	0.07	-0.02	-0.03	0.01
family	0.10	0.15	0.18	-0.11	-0.05
f0_std	0.10	0.15	0.16	-0.06	-0.22
mfcc_std_4	-0.12	-0.25	-0.22	0.12	-0.21
ipron	-0.05	0.09	-0.08	-0.11	0.03
Shimmer	0.14	0.26	0.25	-0.14	-0.04
affect	0.30	0.29	0.50	-0.29	0.00
mfcc_std_11	-0.09	-0.18	-0.19	0.13	-0.05
f1_mean	0.02	0.07	0.05	-0.04	-0.32
motion	-0.10	-0.05	-0.13	0.08	0.05

	Period	prep	focusfuture	family	f0_std
prep	-0.30				
focusfuture	-0.07	0.00			
family	0.19	-0.18	-0.08		
f0_std	0.02	-0.11	-0.04	0.04	
mfcc_std_4	-0.48	0.21	0.01	-0.10	0.04
ipron	-0.19	-0.06	0.14	-0.10	0.03
Shimmer	0.47	-0.16	-0.01	0.17	0.03
affect	0.34	-0.32	-0.06	0.20	0.16
mfcc_std_11	-0.42	0.13	0.02	-0.08	0.17
f1_mean	-0.07	-0.08	-0.01	0.06	0.26
motion	-0.18	0.10	0.23	-0.05	0.01

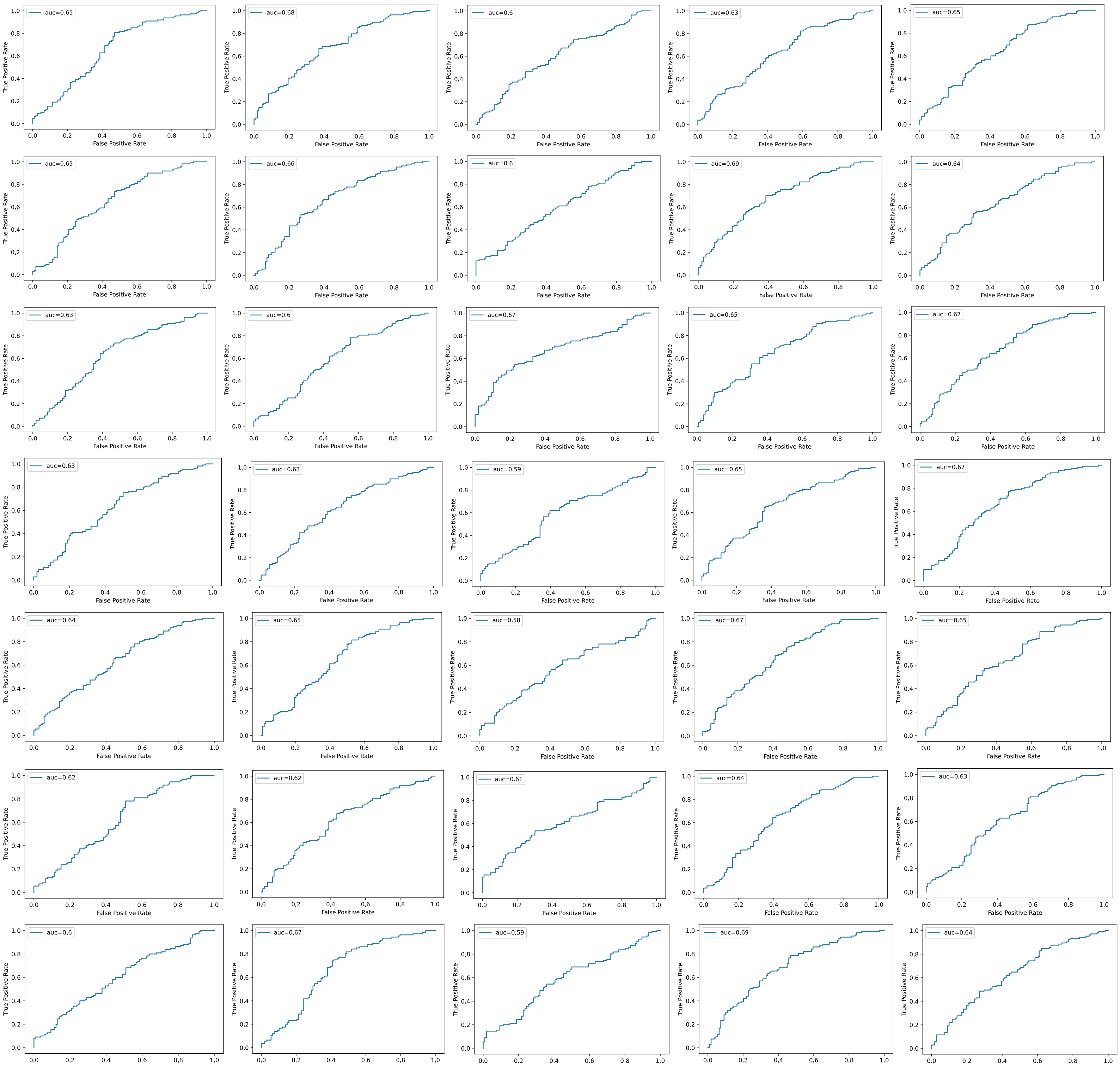
	mfcc_std_4	ipron	Shimmer	affect	mfcc_std_11
ipron	0.14				
Shimmer	-0.38	-0.13			
affect	-0.26	-0.11	0.24		

	mfcc_std_4	ipron	Shimmer	affect	mfcc_std_11
mfcc_std_11	0.58	0.09	-0.41	-0.22	
f1_mean	0.11	0.01	0.00	0.02	0.12
motion	0.07	0.08	-0.12	-0.12	0.11

	f1_mean	motion
motion	-0.04	

Appendix M

AUROC Curves of a Prediction Model Built using Acoustic Features, Linguistic Features and Demographics


auc=0.65
auc=0.68
auc=0.6
auc=0.63
auc=0.65
auc=0.65
auc=0.66
auc=0.6
auc=0.69
auc=0.64
auc=0.63
auc=0.6
auc=0.67
auc=0.65
auc=0.67
auc=0.63
auc=0.63
auc=0.59
auc=0.65
auc=0.67
auc=0.64
auc=0.65
auc=0.58
auc=0.67
auc=0.65
auc=0.62
auc=0.62
auc=0.61
auc=0.64
auc=0.63
auc=0.6
auc=0.67
auc=0.59
auc=0.69
auc=0.64
True Positive Rate
False Positive Rate

Bibliography

[1] M. H. C. of Canada, "Making the case for investing in mental health in canada," *London (ON): Mental Health Commission*, 2013.

[2] D. Mechanic, "Removing barriers to care among persons with psychiatric symptoms," *Health Affairs*, vol. 21, no. 3, pp. 137–147, 2002.

[3] "Shortage of psychiatrists contributing to crisis in access to mental health treatment - ontario psychiatric association - OPA." (), [Online]. Available: https://eopa.ca/news-updates/shortage-of-psychiatrists-contributing-to-crisis-in-access-to-mental-health-treatment (visited on 12/15/2022).

[4] timcharlet. "Cost vs reward of psychiatry school," Doctorly.org. (), [Online]. Available: https://doctorly.org/cost-vs-reward-of-a-psychiatry-degree/ (visited on 11/17/2022).

[5] T. A. Brown, P. A. Di Nardo, C. L. Lehman, and L. A. Campbell, "Reliability of DSM-IV anxiety and mood disorders: Implications for the classification of emotional disorders," *Journal of Abnormal Psychology*, vol. 110, no. 1, pp. 49–58, Feb. 2001, ISSN: 0021-843X. DOI: 10.1037//0021-843x.110.1.49.

[6] S. Canada, *Canadian community health survey-annual component (cchs)*, 2013.

[7] D. Di Matteo, A. Fine, K. Fotinos, J. Rose, and M. Katzman, "Patient willingness to consent to mobile phone data collection for mental health apps: Structured questionnaire," *JMIR mental health*, vol. 5, no. 3, e56, Aug. 29, 2018, ISSN: 2368-7959. DOI: 10.2196/mental.9539.

[8] D. Di Matteo, K. Fotinos, S. Lokuge, J. Yu, T. Sternat, M. A. Katzman, and J. Rose, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: Exploratory study," *JMIR formative research*, vol. 4, no. 8, e18751, Aug. 13, 2020, ISSN: 2561-326X. DOI: 10.2196/18751.

[9] D. Di Matteo, W. Wang, K. Fotinos, S. Lokuge, J. Yu, T. Sternat, M. A. Katzman, and J. Rose, "Smartphone-detected ambient speech and self-reported measures of anxiety and depression: Exploratory observational study," *JMIR formative research*, vol. 5, no. 1, e22723, Jan. 29, 2021, ISSN: 2561-326X. DOI: 10.2196/22723.

[10] D. Di Matteo, K. Fotinos, S. Lokuge, G. Mason, T. Sternat, M. A. Katzman, and J. Rose, "Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: Cross-sectional study," *Journal of Medical Internet Research*, vol. 23, no. 8, e28918, Aug. 13, 2021, ISSN: 1438-8871. DOI: 10.2196/28918.

[11] D. Di Matteo, "Inference of anxiety and depression from smartphone-collected data," Ph.D. dissertation, University of Toronto, 2021.

[12] F. Durbano, Ed., *New Insights into Anxiety Disorders*, InTech, Mar. 20, 2013, ISBN: 978-953-51-1053-8. DOI: 10.5772/46003. [Online]. Available: http://www.intechopen.com/books/new-insights-into-anxiety-disorders (visited on 11/22/2022).

[13] M. Guha, "Encyclopedia of psychopharmacology," *Reference Reviews*, pp. 699–702, 2015.

[14] H.-U. Wittchen, "Generalized anxiety disorder: Prevalence, burden, and cost to society," *Depression and anxiety*, vol. 16, no. 4, pp. 162–171, 2002.

[15] M. Ansseau, B. Fischler, M. Dierick, A. Albert, S. Leyman, and A. Mignon, "Socioeconomic correlates of generalized anxiety disorder and major depression in primary care: The GADIS II study (generalized anxiety and depression impact survey II)," *Depression and Anxiety*, vol. 25, no. 6, pp. 506–513, 2008.

[16] D. J. Katzelnick, K. A. Kobak, T. DeLeire, H. J. Henk, J. H. Greist, J. R. T. Davidson, F. R. Schneier, M. B. Stein, and C. P. Helstad, "Impact of generalized social anxiety disorder in managed care," *American Journal of Psychiatry*, vol. 158, no. 12, pp. 1999–2007, 2001.

[17] A. P. Association and A. P. Association, Eds., *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed, Washington, D.C: American Psychiatric Association, 2013, 947 pp., ISBN: 978-0-89042-554-1 978-0-89042-555-8.

[18] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 22, 2006, ISSN: 0003-9926. DOI: 10.1001/archinte.166.10.1092.

[19] B. J. Mohan, "Speech recognition using mfcc and dtw," in *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, IEEE, 2014, pp. 1–4.

[20] V. Tiwari, "Mfcc and its applications in speaker recognition," *International Journal on Emerging Technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[21] E. Rejaibi, A. Komaty, F. Mériaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *arXiv preprint arXiv:1909.07208*, 2019.

[22] S. Sabahi, *My-voice-analysis*, Publication Title: GitHub repository, 2019. [Online]. Available: https://github.com/Shahabks/my-voice-analysis.

[23] R. M. Roark, "Frequency and voice: Perspectives in the time domain," *Journal of Voice*, vol. 20, no. 3, pp. 325–354, 2006.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[25] D. Aalto, J. Malinen, and M. Vainio, "Formants," in *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Jun. 25, 2018, ISBN: 978-0-19-938465-5. DOI: 10.1093/acrefore/9780199384655.013.419. [Online]. Available: http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-419 (visited on 01/25/2022).

[26] I. R. Titze, R. J. Baken, K. W. Bozeman, *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, May 2015, ISSN: 0001-4966. DOI: 10.1121/1.4919349. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4919349 (visited on 11/26/2022).

[27] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language. use: Our words, our selves," *Annual Review of Psychology*, vol. 54, pp. 547–577, 2003, ISSN: 0066-4308. DOI: 10.1146/annurev.psych.54.101601.145041.

[28] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," The University of Texas at Austin, 2015.

[29] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, Aug. 6, 2019, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1903070116. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1903070116 (visited on 03/25/2022).

[30] L. Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, p. 12, 1991.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00994018. [Online]. Available: http://link.springer.com/10.1007/BF00994018 (visited on 01/25/2022).

[32] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0042519 (visited on 11/27/2022).

[33] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks for Perception*, Elsevier, 1992, pp. 65–93.

[34] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000, ISBN: 81-317-1672-4.

[35] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Interspeech 2014*, ISCA, Sep. 14, 2014, pp. 1053–1057. DOI: 10.21437/Interspeech.2014-273. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2014/bengio14_interspeech.html (visited on 07/25/2022).

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, Sep. 6, 2013. arXiv: 1301.3781[cs]. [Online]. Available: http://arxiv.org/abs/1301.3781 (visited on 07/25/2022).

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: http://aclweb.org/anthology/D14-1162 (visited on 07/25/2022).

[38] Y. Li and T. Yang, "Word embedding for understanding natural language: A survey," in *Guide to Big Data Applications*, S. Srinivasan, Ed., vol. 26, Series Title: Studies in Big Data, Cham: Springer International Publishing, 2018, pp. 83–104, ISBN: 978-3-319-53816-7 978-3-319-53817-4. DOI: 10.1007/978-3-319-53817-4_4. [Online]. Available: http://link.springer.com/10.1007/978-3-319-53817-4_4 (visited on 07/25/2022).

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, Number: arXiv:1706.03762, Dec. 5, 2017. arXiv: 1706.03762[cs]. [Online]. Available: http://arxiv.org/abs/1706.03762 (visited on 06/08/2022).

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, May 24, 2019. arXiv: 1810.04805[cs]. [Online]. Available: http://arxiv.org/abs/1810.04805 (visited on 07/25/2022).

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[42] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020, Publisher: Springer, ISSN: 1572-8641.

[43] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer: The long-document transformer*, Number: arXiv:2004.05150, Dec. 2, 2020. arXiv: 2004.05150[cs]. [Online]. Available: http://arxiv.org/abs/2004.05150 (visited on 06/08/2022).

[44] T. M. Mitchell, *Machine learning*, Nachdr., ser. McGraw-Hill series in Computer Science. New York: McGraw-Hill, 2013, 414 pp., ISBN: 978-0-07-115467-3 978-0-07-042807-2.

[45] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, *Structured attention networks*, Feb. 16, 2017. arXiv: 1702.00887[cs]. [Online]. Available: http://arxiv.org/abs/1702.00887 (visited on 07/25/2022).

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[47] I. Pavlidis, J. Levine, and P. Baukol, "Thermal imaging for anxiety detection," in *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No.PR00640)*, Hilton Head, SC, USA: IEEE Comput. Soc, 2000, pp. 104–109, ISBN: 978-0-7695-0640-1. DOI: 10.1109/CVBVS.2000.855255. [Online]. Available: http://ieeexplore.ieee.org/document/855255/ (visited on 02/17/2022).

[48] A. Frick, M. Gingnell, A. F. Marquand, K. Howner, H. Fischer, M. Kristiansson, S. C. R. Williams, M. Fredrikson, and T. Furmark, "Classifying social anxiety disorder using multivoxel pattern analyses of brain function and structure," *Behavioural Brain Research*, vol. 259, pp. 330–335, Feb. 1, 2014, ISSN: 1872-7549. DOI: 10.1016/j.bbr.2013.11.003.

[49] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 261–268, Jul. 2011, ISSN: 17468094. DOI: 10.1016/j.bspc.2010.12.001. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809410000911 (visited on 02/17/2022).

[50] M. Chatterjee, G. Stratou, S. Scherer, and L.-P. Morency, "Context-based signal descriptors of heart-rate variability for anxiety assessment," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: IEEE, May 2014, pp. 3631–3635, ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854278. [Online]. Available: http://ieeexplore.ieee.org/document/6854278/ (visited on 02/17/2022).

[51] O. Doehrmann, S. S. Ghosh, F. E. Polli, *et al.*, "Predicting treatment response in social anxiety disorder from functional magnetic resonance imaging," *JAMA psychiatry*, vol. 70, no. 1, pp. 87–97, Jan. 2013, ISSN: 2168-6238. DOI: 10.1001/2013.jamapsychiatry.5.

[52] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," *Biomedical Signal Processing and Control*, vol. 31, pp. 89–101, Jan. 2017, ISSN: 17468094. DOI: 10.1016/j.bspc.2016.06.020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809416300805 (visited on 02/16/2022).

[53] M. Boukhechba, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes, "Predicting social anxiety from global positioning system traces of college students: Feasibility study," *JMIR mental health*, vol. 5, no. 3, e10101, Jul. 4, 2018, ISSN: 2368-7959. DOI: 10.2196/10101.

[54] T. Özseven, M. Düğenci, A. Doruk, and H. İ. Kahraman, "Voice traces of anxiety: Acoustic parameters affected by anxiety disorder," *Archives of Acoustics*, pp. 625–636, 2018, Publisher: Committee on Acoustics PAS, PAS Institute of Fundamental Technological Research, Polish Acoustical Society. DOI: 10.24425/AOA.2018.125156. [Online]. Available: https://journals.pan.pl/dlibra/publication/125156/edition/109196/content (visited on 01/25/2022).

[55] A. T. Beck and R. Steer, "Beck anxiety inventory (BAI)," *Überblick über Reliabilitäts-und Validitätsbefunde von klinischen und außerklinischen Selbst-und Fremdbeurteilungsverfahren*, vol. 7, 1988.

[56] J. W. Weeks, C.-Y. Lee, A. R. Reilly, A. N. Howell, C. France, J. M. Kowalsky, and A. Bush, ""the sound of fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder," *Journal of Anxiety Disorders*, vol. 26, no. 8, pp. 811–822, Dec. 2012, ISSN: 1873-7897. DOI: 10.1016/j.janxdis.2012.07.005.

[57] B. Anderson, P. R. Goldin, K. Kurita, and J. J. Gross, "Self-representation in social anxiety disorder: Linguistic analysis of autobiographical narratives," *Behaviour Research and Therapy*, vol. 46, no. 10, pp. 1119–1125, Oct. 2008, ISSN: 1873-622X. DOI: 10.1016/j.brat.2008.07.001.

[58] S. G. Hofmann, P. M. Moore, C. Gutner, and J. W. Weeks, "Linguistic correlates of social anxiety disorder," *Cognition & Emotion*, vol. 26, no. 4, pp. 720–726, Jun. 2012, ISSN: 0269-9931, 1464-0600. DOI: 10.1080/02699931.2011.602048. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02699931.2011.602048 (visited on 07/12/2022).

[59] A. R. Sonnenschein, S. G. Hofmann, T. Ziegelmayer, and W. Lutz, "Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy," *Cognitive Behaviour Therapy*, vol. 47, no. 4, pp. 315–327, Jul. 2018, ISSN: 1651-2316. DOI: 10.1080/16506073.2017.1419505.

[60] E. W. McGinnis, S. P. Anderau, J. Hruschak, R. D. Gurchiek, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum, M. Muzik, and R. S. McGinnis, "Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2294–2301, Nov. 2019, ISSN: 2168-2208. DOI: 10.1109/JBHI.2019.2913590.

[61] 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," *Psychosomatic Medicine*, vol. 59, no. 4, pp. 419–426, Aug. 1997, ISSN: 0033-3174. DOI: 10.1097/00006842-199707000-00012.

[62] J. W. Weeks, A. Srivastav, A. N. Howell, and A. R. Menatti, ""speaking more than words": Classifying men with social anxiety disorder via vocal acoustic analyses of diagnostic interviews," *Journal of Psychopathology and Behavioral Assessment*, vol. 38, no. 1, pp. 30–41, Mar. 2016, ISSN: 0882-2689, 1573-3505. DOI: 10.1007/s10862-015-9495-9. [Online]. Available: http://link.springer.com/10.1007/s10862-015-9495-9 (visited on 02/22/2022).

[63] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, p. 81, Jun. 2018, ISSN: 2474-9567. DOI: 10.1145/3214284.

[64] A. Baird, N. Cummins, S. Schnieder, J. Krajewski, and B. W. Schuller, "An evaluation of the effect of anxiety on speech — computational prediction of anxiety from sustained vowels," in *Interspeech 2020*, ISCA, Oct. 25, 2020, pp. 4951–4955. DOI: 10.21437/Interspeech.2020-1801. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/baird20_interspeech.html (visited on 09/16/2022).

[65] L. Rook, M. C. Mazza, I. Lefter, and F. Brazier, "Toward linguistic recognition of generalized anxiety disorder," *Frontiers in Digital Health*, vol. 4, p. 779 039, 2022, ISSN: 2673-253X. DOI: 10.3389/fdgth.2022.779039.

[66] C. S. Carver and T. L. White, "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales.," *Journal of Personality and Social Psychology*, vol. 67, no. 2, pp. 319–333, Aug. 1994, ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.67.2.319. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.67.2.319 (visited on 09/22/2022).

[67] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, Mar. 2018, ISSN: 22146350. DOI: 10.1016/j.jbef.2017.12.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214635017300989 (visited on 01/25/2022).

[68] J. Reilly and J. L. Fisher, "Sherlock holmes and the strange case of the missing attribution: A historical note on "the grandfather passage"," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 1, pp. 84–88, Feb. 2012, ISSN: 1092-4388, 1558-9102. DOI: 10.1044/1092-4388(2011/11-0158). [Online]. Available: http://pubs.asha.org/doi/10.1044/1092-4388%5C%282011/11-0158%5C%29 (visited on 01/25/2022).

[69] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1, pp. 76–81, 1993, ISSN: 0302-282X. DOI: 10.1159/000119004.

[70] G. Gerra, A. Zaimovic, U. Zambelli, M. Timpano, N. Reali, S. Bernasconi, and F. Brambilla, "Neuroendocrine responses to psychological stress in adolescents with anxiety disorder," *Neuropsychobiology*, vol. 42, no. 2, pp. 82–92, 2000, ISSN: 0302-282X. DOI: 10.1159/000026677.

[71] D. Jezova, A. Makatsori, R. Duncko, F. Moncek, and M. Jakubek, "High trait anxiety in healthy subjects is associated with low neuroendocrine activity during psychosocial stress," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 28, no. 8, pp. 1331–1336, Dec. 2004, ISSN: 0278-5846. DOI: 10.1016/j.pnpbp.2004.08.005.

[72] N. S. Endler and N. L. Kocovski, "State and trait anxiety revisited," *Journal of Anxiety Disorders*, vol. 15, no. 3, pp. 231–245, Jun. 2001, ISSN: 0887-6185. DOI: 10.1016/s0887-6185(01)00060-3.

[73] *Google web designer*. [Online]. Available: https://webdesigner.withgoogle.com/.

[74] *Firebase*. [Online]. Available: https://firebase.google.com/.

[75] A. Natal, G. Shires, and P. Jägenstedt. "Web speech API." (), [Online]. Available: https://wicg.github.io/speech-api/ (visited on 11/28/2022).

[76] A. Kaehler and G. R. Bradski, *Learning OpenCV: computer vision with the OpenCV library*, First edition, Second release. Sebastopol, CA: O'Reilly Media, 2017, 990 pp., OCLC: ocn968936315, ISBN: 978-1-4919-3799-0.

[77] prolific.ac, *Online participant recruitment for surveys and market research*. 2014. [Online]. Available: https://www.prolific.co/.

[78] Public Health Canada, *Mental health - anxiety disorders - canada.ca*, Backup Publisher: Canada.ca, 2009. [Online]. Available: https://www.canada.ca/en/health-canada/services/healthy-living/your-health/diseases/mental-health-anxiety-disorders.html.

[79] O. G. Cameron and E. M. Hill, "Women and anxiety," *Psychiatric Clinics of North America*, vol. 12, no. 1, pp. 175–186, Mar. 1989, ISSN: 0193953X. DOI: 10.1016/S0193-953X(18)30459-3. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0193953X18304593 (visited on 01/25/2022).

[80] S. M. A. Dijkstra-Kersten, K. E. M. Biesheuvel-Leliefeld, J. C. van der Wouden, B. W. J. H. Penninx, and H. W. J. van Marwijk, "Associations of financial strain and income with depressive and anxiety disorders," *Journal of Epidemiology and Community Health*, vol. 69, no. 7, pp. 660–665, Jul. 2015, ISSN: 1470-2738. DOI: 10.1136/jech-2014-205088.

[81] C. Krasucki, R. Howard, and A. Mann, "The relationship between anxiety disorders and age," *International Journal of Geriatric Psychiatry*, vol. 13, no. 2, pp. 79–99, Feb. 1998, ISSN: 0885-6230. DOI: 10.1002/(sici)1099-1166(199802)13:2<79::aid-gps739>3.0.co;2-g.

[82] B. G. Teferra, S. Borwein, D. D. DeSouza, W. Simpson, L. Rheault, and J. Rose, "Acoustic and linguistic features of impromptu speech and their association with anxiety: Validation study," *JMIR Mental Health*, vol. 9, no. 7, e36828, Jul. 8, 2022, ISSN: 2368-7959. DOI: 10.2196/36828. [Online]. Available: https://mental.jmir.org/2022/7/e36828 (visited on 07/11/2022).

[83] R. Kliper, S. Portuguese, and D. Weinshall, "Prosodic analysis of speech and the underlying mental state," in *Pervasive Computing Paradigms for Mental Health*, S. Serino, A. Matic, D. Giakoumis, G. Lopez, and P. Cipresso, Eds., vol. 604, Series Title: Communications in Computer and Information Science, Cham: Springer International Publishing, 2016, pp. 52–62, ISBN: 978-3-319-32269-8 978-3-319-32270-4. DOI: 10.1007/978-3-319-32270-4_6. [Online]. Available: http://link.springer.com/10.1007/978-3-319-32270-4_6 (visited on 01/25/2022).

[84] T. Wortwein, L.-P. Morency, and S. Scherer, "Automatic assessment and analysis of public speaking anxiety: A virtual audience case study," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China: IEEE, Sep. 2015, pp. 187–193, ISBN: 978-1-4799-9953-8. DOI: 10.1109/ACII.2015.7344570. [Online]. Available: http://ieeexplore.ieee.org/document/7344570/ (visited on 01/25/2022).

[85] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.1979.4766909. [Online]. Available: http://ieeexplore.ieee.org/document/4766909/ (visited on 01/25/2022).

[86] P. Laukka, C. Linnman, F. Åhs, *et al.*, "In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech," *Journal of Nonverbal Behavior*, vol. 32, no. 4, pp. 195–214, Dec. 2008, ISSN: 0191-5886, 1573-3653. DOI: 10.1007/s10919-008-0055-9. [Online]. Available: http://link.springer.com/10.1007/s10919-008-0055-9 (visited on 01/25/2022).

[87] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," *PloS One*, vol. 16, no. 4, e0248842, 2021, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0248842.

[88] M. A. Hagenaars and A. van Minnen, "The effect of fear on paralinguistic aspects of speech in patients with panic disorder with agoraphobia," *Journal of Anxiety Disorders*, vol. 19, no. 5, pp. 521–537, 2005, ISSN: 0887-6185. DOI: 10.1016/j.janxdis.2004.04.008.

[89] V. Silber-Varod, H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, "Do social anxiety individuals hesitate more? the prosodic profile of hesitation disfluencies in social anxiety disorder individuals," in *Speech Prosody 2016*, ISCA, 2016, pp. 1211–1215. DOI: 10.21437/SpeechProsody.2016-249. [Online]. Available: https://www.isca-speech.org/archive/speechprosody_2016/silbervarod16_speechprosody.html (visited on 01/25/2022).

[90] B. F. Fuller, Y. Horii, and D. A. Conner, "Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety," *Research in Nursing & Health*, vol. 15, no. 5, pp. 379–389, 1992, ISSN: 0160-6891. DOI: 10.1002/nur.4770150507.

[91] R. Lenain, J. Weston, A. Shivkumar, and E. Fristed, "Surfboard: Audio feature extraction for modern machine learning," in *Interspeech 2020*, ISCA, 2020, pp. 2917–2921. DOI: 10.21437/Interspeech.2020-2879. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/lenain20_interspeech.html (visited on 01/25/2022).

[92] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, Austin, Texas, 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003. [Online]. Available: https://conference.scipy.org/proceedings/scipy2015/brian_mcfee.html (visited on 01/25/2022).

[93] S. Hashemipour and M. Ali, "Amazon web services (AWS) – an overview of the on-demand cloud computing platform," in *Emerging Technologies in Computing*, M. H. Miraz, P. S. Excell, A. Ware, S. Soomro, and M. Ali, Eds., vol. 332, Series Title: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Cham: Springer International Publishing, 2020, pp. 40–47, ISBN: 978-3-030-60035-8 978-3-030-60036-5. DOI: 10.1007/978-3-030-60036-5_3. [Online]. Available: http://link.springer.com/10.1007/978-3-030-60036-5_3 (visited on 04/01/2022).

[94] T. Pellegrini, A. Hämäläinen, P. B. d. Mareüil, M. Tjalve, I. Trancoso, S. Candeias, M. S. Dias, and D. Braga, "A corpus-based study of elderly and young speakers of european portuguese: Acoustic correlates and their impact on speech recognition performance," in *Interspeech 2013*, ISCA, Aug. 25, 2013, pp. 852–856. DOI: 10.21437/Interspeech.2013-241. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2013/pellegrini13_interspeech.html (visited on 04/11/2022).

[95] K. Farrow, G. Grolleau, and N. Mzoughi, "What in the word! the scope for the effect of word choice on economic behavior: What in the word!" *Kyklos*, vol. 71, no. 4, pp. 557–580, Nov. 2018, ISSN: 00235962. DOI: 10.1111/kykl.12186. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/kykl.12186 (visited on 04/11/2022).

[96] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, Dec. 2004, ISSN: 1369-1473, 1467-842X. DOI: 10.1111/j.1467-842X.2004.00360.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-842X.2004.00360.x (visited on 01/25/2022).

[97] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz, "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 44, pp. 11 203–11 208, Oct. 30, 2018, ISSN: 1091-6490. DOI: 10.1073/pnas.1802331115.

[98] B. G. Teferra, S. Borwein, D. D. DeSouza, and J. Rose, "Screening for generalized anxiety disorder from acoustic and linguistic features of impromptu speech: Prediction model evaluation study," *JMIR formative research*, vol. 6, no. 10, e39998, Oct. 28, 2022, ISSN: 2561-326X. DOI: 10.2196/39998.

[99] B. Bischl, O. Mersmann, H. Trautmann, and C. Weihs, "Resampling methods for meta-model validation with recommendations for evolutionary computation," *Evolutionary Computation*, vol. 20, no. 2, pp. 249–275, 2012, ISSN: 1530-9304. DOI: 10.1162/EVCO_a_00069.

[100] A. Marcano-Cedeno, J. Quintanilla-Dominguez, M. G. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, Glendale, AZ, USA: IEEE, Nov. 2010, pp. 2845–2850, ISBN: 978-1-4244-5225-5. DOI: 10.1109/IECON.2010.5675075. [Online]. Available: http://ieeexplore.ieee.org/document/5675075/ (visited on 12/16/2022).

[101] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Advances in Knowledge Discovery and Data Mining*, H. Dai, R. Srikant, and C. Zhang, Eds., red. by T. Kanade, J. Kittler, J. M. Kleinberg, *et al.*, vol. 3056, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 3–12, ISBN: 978-3-540-22064-0 978-3-540-24775-3. DOI: 10.1007/978-3-540-24775-3_3. [Online]. Available: http://link.springer.com/10.1007/978-3-540-24775-3_3 (visited on 02/25/2022).

[102] B. G. Teferra and J. Rose, "Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: Model evaluation study (preprint)," JMIR Mental Health, preprint, Nov. 15, 2022. DOI: 10.2196/preprints.44325. [Online]. Available: http://preprints.jmir.org/preprint/44325 (visited on 11/17/2022).

[103] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (visited on 06/06/2022).

[104] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, Number: arXiv:1703.01365, Jun. 12, 2017. arXiv: 1703.01365[cs]. [Online]. Available: http://arxiv.org/abs/1703.01365 (visited on 06/08/2022).

[105] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, May 19, 2016. arXiv: 1409.0473[cs,stat]. [Online]. Available: http://arxiv.org/abs/1409.0473 (visited on 12/01/2022).

[106] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer*, Dec. 2, 2020. [Online]. Available: https://huggingface.co/docs/transformers/v4.20.1/en/model_doc/longformer#longformer.

[107] T. Wolf, L. Debut, V. Sanh, *et al.*, *HuggingFace's transformers: State-of-the-art natural language processing*, Number: arXiv:1910.03771, Jul. 13, 2020. arXiv: 1910.03771[cs]. [Online]. Available: http://arxiv.org/abs/1910.03771 (visited on 07/19/2022).

[108] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, Dec. 2015, pp. 19–27, ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.11. [Online]. Available: http://ieeexplore.ieee.org/document/7410368/ (visited on 12/02/2022).

[109] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, *Defending against neural fake news*, Dec. 11, 2020. arXiv: 1905.12616[cs]. [Online]. Available: http://arxiv.org/abs/1905.12616 (visited on 12/02/2022).

[110] T. H. Trinh and Q. V. Le, *A simple method for commonsense reasoning*, Sep. 26, 2019. arXiv: 1806.02847[cs]. [Online]. Available: http://arxiv.org/abs/1806.02847 (visited on 12/02/2022).

[111] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 28, 2015, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539. [Online]. Available: http://www.nature.com/articles/nature14539 (visited on 07/11/2022).

[112] C. Pierse, *Transformers interpret*, version 0.5.2, Feb. 14, 2021. [Online]. Available: https://github.com/cdpierse/transformers-interpret.

[113] A. W. Siegman, "The meaning of silent pauses in the initial interview," *The Journal of Nervous and Mental Disease*, vol. 166, no. 9, pp. 642–654, Sep. 1978, ISSN: 0022-3018. DOI: 10.1097/00005053-197809000-00004.

[114] G. F. Mahl, "Disturbances and silences in the patient's speech in psychotherapy," *Journal of Abnormal Psychology*, vol. 53, no. 1, pp. 1–15, Jul. 1956, ISSN: 0021-843X. DOI: 10.1037/h0047552.