



**Analysing Changes in the Acoustic Features of the Human Voice to
Detect Depression amongst Biological Females in Higher Education**

Joel Cooke

A thesis submitted to the University of Huddersfield in partial
fulfilment of the requirements for the degree of
Masters by Research

November 2023

Contents

- a. Statement of Copyright
- b. Ethics Declaration
- c. Acknowledgements

- 1. Abstract
- 2. Introduction
 - a. Topic and Context
 - b. Focus and Scope
 - c. Relevance and Importance
 - d. Questions and Objectives
 - e. Overview of Thesis Structure
- 3. Literature Review
 - a. Introduction to Literature Review
 - b. Current Depression Diagnosis
 - c. Acoustic Features and Depression
 - d. Detecting Depression Severity
 - e. The Impact of Age on the Voice
 - f. The Impact of Educational Level on the Voice
 - g. The Impact of Other Variables on the Voice
 - h. Future Directions
 - i. Summary of Literature Review
- 4. Methodology
 - a. Background
 - b. Participants
 - c. Vocal Tasks
 - d. Acoustic Features
 - e. Acoustic Feature Extraction
 - f. Data Analysis
- 5. Results
- 6. Discussion
- 7. Limitations
- 8. Conclusion

- i. References
- ii. Appendices

a) Statement of Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and s/he has given The University of Huddersfield the right to use such copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/ or Reproductions.

b) Ethics Declaration

All data gathered from participants was done so with their full consent. Data was deleted after analysis unless the participant specified their voice data could be used in the future. Ethical guidelines were adhered to throughout.

c) Acknowledgements

I would like to thank my supervisors, Professor Monty Adkins, and Professor Chia-Yuan Lin, for their tremendous support and guidance throughout the project. Also, my friends and family, for showing a continued interest in the progress of this thesis. This project is dedicated to them.

1) **Abstract**

Depression significantly affects a large percentage of the population, with young adult females being one of the most at-risk demographics. Concurrently, there is a growing demand on healthcare, and with sufficient resources often unavailable to diagnose depression, new diagnostic methods are needed that are both cost-effective and accurate. The presence of depression is seen to significantly affect certain acoustic features of the human voice. Acoustic features have been found to exhibit subtle changes beyond the perception of the human auditory system when an individual has depression. With advances in speech processing, these subtle changes can be observed by machines. By measuring these changes, the human voice can be analysed to identify acoustic features that show a correlation with depression. The implementation of voice diagnosis would both reduce the burden on healthcare and ensure those with depression are diagnosed in a timely fashion, allowing them quicker access to treatment.

The research project presents an analysis of voice data from 17 biological females between the ages of 20-26 years old in higher education as a means to detect depression. Eight participants were considered healthy with no history of depression, whilst the other nine currently had depression. Participants performed two vocal tasks consisting of extending sounds for a period of time and reading back a passage of speech. Six acoustic features were then measured from the voice data to determine whether these features can be utilised as diagnostic indicators of depression. The main finding of this study demonstrated one of the acoustic features measured demonstrates significant differences when comparing depressed and healthy individuals.

2) **Introduction**

a. Topic and Context

Researchers estimate human speech to have begun anywhere between 50,000 and 2 million years ago (Balter, 2015). "Like DNA and fingerprints, every human's voice is unique. It carries more information than we realize (or can hear)" (Singh, n.d). The human voice is extremely complex, depending upon interactions between our muscles and respiratory system (Services, 2019) and "the coordination of around 100 muscles with significant temporal precision" (Almaghrabi et al., 2023). The voice is said to "respond to any factor that effects the body or mind, and thereby the muscular, nervous, endocrine and other systems in the body" (Singh, 2019). With the advent and progression of technologies such as Digital Signal Processing (DSP) and Machine Learning (ML) in recent times, we are now able to analyse minute changes in the human voice in an objective and precise manner, to extract numerical values which can then be examined for the presence of medical conditions, (Almaghrabi et al., 2023).

Depression is a psychological disorder, diagnosed by a mental health professional. It "varies from: normal transient low mood in daily life itself, to clinical syndrome, with severe and significant duration and associated signs and symptoms, markedly different from normality" (Rondon Bernard, 2018). Sufferers of depression can experience mild, moderate, or severe symptoms. "Globally, an estimated 5% of adults suffer from depression. More women are affected by depression than men" (World Health Organization, 2023). Based on information gathered in October 2022, in Great Britain, "around 1 in 6 (16%) adults experienced moderate to severe depressive symptoms", compared with 10% of adults suffering from depression before the Covid-19 (co19) pandemic (Cullum Attwell, 2022). There are a variety of symptoms that signify a person might be suffering from depression, some include sadness, loss of pleasure and low self-esteem (Rondon Bernard, 2018). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM), symptoms of depression must be persistent for at least two weeks, impairing social, occupational, or other important areas of one's life. Symptoms must also not be caused by another medical condition or substance abuse (American Psychiatric Association, 2023).

The link between depression and the manifestation of changes in the human voice is drawn as "everything that could influence your physiology and your mind could impact your voice, as speech is a complex biomechanical process" (Singh et al., 2016). The workings of the human vocal system mean that the vocal folds "involuntarily and instantaneously respond to simply any factor that effects the body and mind, and thereby the muscular, nervous, endocrine and other systems in the body" (Singh, 2019).

In one sense, the idea of using the human voice to diagnose health conditions is not new. "In the early 1900s, Swiss psychologist Eugen Bleuler and his then-assistant Carl Jung pioneered the use of word association, one of the first observational empirical tests used in psychoanalysis. A delayed response

time or jarring word associations could indicate psychological conflicts and help toward a diagnosis” (Bourel, 2019). In Traditional Chinese Medicine (TCM), listening to the patient is one of the four main diagnostic methods (Wang & Dong, 2017). The Ayurvedic healthcare system also places major importance on the voice. In Ayurveda “the characteristics of the voice indicate the physiology and the pathology of the body” (Kalyani et al., 2021). In the eyes of western medicine however, the ways in which the voice is utilized and analysed in TCM and Ayurveda could be seen as somewhat subjective, as western medicine “relies heavily on information acquired through laboratory and other testing methods” (Wang & Dong, 2017). On the contrary, western medicine still employs the subjective analysis of the voice in certain situations, as doctors today perform Auditory Perceptual Analysis (APA) to assess the patient’s voice quality. Human-judged properties of the voice are considered as qualitative features of the voice, and are related to the nasality, breathiness, and roughness of the voice (Singh, 2019). The use of APA will be dependent upon the Doctors skills, experience, and perception, and therefore could mean there is subjectivity in the diagnostic process when performing APA.

Analysing the voices of those around us is an innate quality humans possess. For example, when a loved one’s voice changes due to sickness or sadness, it is the tone, pitch, undulation, and speech of enunciation that we immediately perceive as a deviation from the normal tone of the person. Despite this innate unsullied perception of the voice that could be argued that we inherently possess, we still are not able to articulate logically what precisely has changed in the persons voice, or what is wrong with them, as this is likely based on our intuition, and is often done on a subconscious level. Subjective analysis does not provide the same accuracy and reliability as objective measurements when attempting to use the voice to diagnose a patient for several reasons. Subjective observations contain variability, bias, and an inability to reproduce results which demonstrate a clear understanding of the path that was followed to reach the diagnosis. Based on the understanding that western medicine today acquires most diagnostic information through laboratory tests (Wang & Dong, 2017), this would render subjective analysis of the human voice largely inapt when diagnosing patients. Objective observations, however, are clear, reproducible, and contain hard scientific evidence of the path taken to achieve the diagnosis.

When the human voice is analysed using technologies such as DSP and ML, it is possible to detect the smallest of changes through micro-articulation, the “measurement of movements, dimensions and positions of the articulators in the human vocal tract during the process of speech production” (Singh et al., 2016). It is important to measure these movements when detecting depression, as depression is known to cause neurophysiological changes, which in turn affect the movement of the vocal folds (Almaghrabi et al., 2023). These changes in the production of speech “can be robustly calculated using computerized processing of the speech waveform” (Almaghrabi et al., 2023). The precision with which these technologies can analyse sound is much greater than the perception of the human ear. The human auditory system can perceive changes over $1/20^{\text{th}}$ of a second, whilst machines are able to perceive

changes far less than this, being able to detect micro-features in the voice. Micro-features are high resolution fine detail features that manifest between $1/20^{\text{th}}$ and $1/40^{\text{th}}$ of a second, and with our ability to detect changes stopping at less than $1/20^{\text{th}}$ of a second, this makes machines a vital part of the analysis process when measuring changes in micro-features. Macro-features on the other hand, can often be perceived by the human auditory system, however they are often connected to voice pattern matching (Singh, 2018), rather than detecting the presence of health conditions.

Medical conditions often “produce involuntary, often human imperceptible changes in the acoustic parameters of the human voice that you can correlate to those conditions” (Services, 2019). Having the ability to analyse the human voice on a granular level with the help of machines to detect micro-features, enables a plethora of acoustic features to be measured (with the assistance of speech analysis software). These acoustic features are considered as quantitative measurements, thus meaning data analysis (that shows a clear path on how results were reached) can be conducted to uncover differences in patterns between a healthy group of individuals, and a group of individuals suffering from a health condition (such as depression). Measuring a consistent change in an acoustic feature when comparing a sample and control group, with a great enough number of participants to provide statistical significance, can give a conviction that the specific acoustic feature has a correlation with a particular health condition.

Acoustic features can generally be classified into two main categories, temporal domain features and spectral domain features. Knowing which category each acoustic feature falls into will dictate what DSP tools are needed to measure these features. Temporal domain features show how the audio signal changes over time. They are measurable in the time domain (Solutions, 2020), therefore a waveform graph would be the most suitable tool to analyse these features. A waveform graph has information related to the duration, amplitude, and periodicity of the sound signal. This graph does not have any information related to the frequency content of the sound. Spectral domain features refer to the ‘spectrum’ of frequencies in the sound signal, and thus measurable in the frequency domain, using a DSP tool such as a spectrum graph. “A frequency domain graph will show how much of a signal lies within each given band over a range of frequencies (Singh, 2019). A spectrum sound graph displays frequency content information in detail, including information on the harmonics and formants of the sound signal. This graph is devoid of information related to the time domain. A spectrogram, however, displays information related to both the time domain and frequency domain, so can be utilised when analysing both types of acoustic features.

There are a range of acoustic features that can be measured from the human voice, including Jitter, Shimmer, Harmonic to Noise Ratio (HNR) and Speech Rate (SR). The acoustic features measured and analysed in this study are Fundamental Frequency (F0), Formant Frequency 1 (F1), Formant Frequency 2 (F2), Voicing Onset Time (VOT), Pitch Variability (PV) and Pause Length (PL). The acoustic features

VOT and PV have been chosen due to the limited academic research on their relationship with the detection of depression. F0 has often been referenced in the literature but has not been analysed based on sex. Analysing female voices separately will help to uncover whether previous findings in the literature related to F0 and depression are independent or dependent on the sex of an individual. F1, F2 and PL have shown potential to be used as indicators of depression, with previous studies finding correlations. However, a focus on whether these correlations apply to the age demographic in this study has not been investigated. Below is a list defining each of the relevant acoustic features to this study:

- Fundamental Frequency (F0): a temporal domain feature that is related to the frequency at which the vocal cords vibrate (Davenport et al., 1998). The F0 can vary depending on the length, thickness, and tension of a person's vocal cords. It can also vary based on biological sex, state of mind, time of day, lifestyle, and professional use of the voice (Teixeira et al., 2013). The average F0 range for adult males is between 150-300Hz. For adult females between 150-300Hz, and for children between 200-500Hz (Davenport et al., 1998).
- Formant Frequencies are the resonances of the vocal tract beyond the F0 and is where sound energy is distributed around a particular frequency (the resonant frequency). In order for these resonances to form, there needs to be no major obstruction in the vocal tract. Plosive sounds such as /p/ and /b/ cause obstructions due to the build-up of pressure, therefore stopping any clear resonances from forming. Vowels cause clear resonances due to the mouth being largely open when they are voiced. This allows for clear formant frequencies to be generated. The mouth does not always need to be open for formants to be formed. However, Nasal sounds can also generate formants as they do not fully obstruct the airflow, rather allowing air to flow through the nasal cavity.
- F1 is the first resonant frequency of the vocal tract and represents the most prominent frequency band in the trachea (Teixeira et al., 2013). F1 can be related to the position of the tongue in the mouth. The lower the tongue is when a sound is voiced, the higher the F1. F2 is the second resonant frequency and represents the most prominent frequency band in the mouth (Teixeira et al., 2013). Whilst F1 is related to the height of the tongue in the mouth, F2 is related to how far forward or back the tongue is when voicing sounds.
- Voicing Onset Time (VOT) is related to the time interval in between a consonant and voicing. VOT is measured after an unvoiced plosive sound (e.g., /t/, /p/ /k/) when it is followed by a voiced sound, such as the vowel /a/ or /u/. When we emit a plosive sound, the vocal tract is closed at a certain point, such as the lips for the letter /p/. Pressure builds up behind this location and is then released. During this short time between the pressure build up and release, the vocal folds do not

vibrate. The time between when the vocal folds are idle and the time when the vocal folds begin to vibrate again is what is measured as the VOT (Singh et al., 2016). This is usually measured in milliseconds (ms).

- Pause Length (PL) is the duration of a pause between words or phrases in a passage of speech. It is often measured in milliseconds (ms). The average PL naturally varies depending on if the speaker is reading back a pre-prepared passage of speech or speaking spontaneously in conversation with another. “Speakers in monologue discourse, for instance, typically vary the length and position of pauses on the basis of the information structure” (Gustafson-Capkova & Megyesi, 2001).
- Pitch Variability (PV) can “refer to one’s ability to vary fundamental frequency (F0) within or between syllables when speaking” (Wong et al., 2021). However, the PV measurement in this study is related to how spread out and by how much the range of frequencies vary by within a specific sound signal.

To measure acoustic features of the voice, there must be a sound signal to extract these measurements from. Participants of any voice study will complete Vocal Tasks (VT) that are designed to suit the aims of the research conducted and the acoustic features being analysed. The three main ways to record the speakers voice which are most referenced in the academic literature are through extended vowel sounds, read speech and dialogue/spontaneous speech. These types of VTs are listed and defined below:

1. Extended vowel sounds do not contain any language content; therefore, they are able to go beyond the constructs of language. As a result, extended vowel sounds have been used in studies where speakers have varying native languages (Omiya et al., 2019). Extended vowel sounds /a, /u and /e have been regularly used in previous studies (Patel et al., 2011; Omiya et al., 2019). All sounds cannot physically be extended, such as /t, /p and /k, as they are plosive sounds which cause obstructions in the vocal tract, stopping air from the lungs escaping, thus building up pressure behind the obstruction. All vowels, however, can be extended, as they do not produce any blockages. As well as vowels, there are also some consonants which can be extended, such as /m, which is classified as a nasal sound, as it stops air escaping through the mouth, rather allowing air to escape through the nasal cavity. Extended periodic sounds, as opposed to aperiodic sounds, allow for a greater number of acoustic features to be measured from the sound signal, as without the vibration of the vocal folds, there is no F0, and therefore no Formant Frequencies present. VOT and many other acoustic features also require the vibration of the vocal folds to be measured.

2. Read speech involves the speaker reading back pre-written passages of text. The benefits of read speech are that it maintains consistency across speakers, as each speaker is reading back the same words, creating a standardised uniform approach. The speech passages used in vocal studies can be chosen for their wide spectrum of frequencies to allow for sufficient acoustic features to be measured. However, there is evidence that read speech has been shown to suppress the expression of emotions when compared to spontaneous speech (Omiya et al., 2018).
3. Dialogue/spontaneous speech is a way of capturing how the speaker would normally sound talking to a friend for example. Everyone has their own vocal cues and ways of expressing themselves, and as a result, free form speech naturally uncovers parts of the speaker's personality. With this, the literal content of the speech segment can be studied, which is useful for helping to diagnose conditions such as schizophrenia (Oxford, n.d). However, free form speech can present many more uncertainties, as the speech content varies depending on the speaker (Shinohara et al., 2016). This means there is less control (especially when analysing large sample populations), as each speaker will naturally vary the length of their phrases. Evidence has suggested spontaneous speech to possess greater dynamic range (Jacewicz et al., 2009), meaning features such as pitch, pauses and the loudness of the voice could be impacted.

b. Focus and Scope

This thesis focuses on the diagnosis of depression amongst biological females between the ages of 18 and 26 in higher education. This includes both those students studying at an undergraduate and postgraduate level. Young adults are the age bracket most at risk to suffering from depression, in particular women, with 43% of women aged between 16 and 29 experiencing depressive symptoms, compared with 26% of men in the same age bracket (Cullum Attwell, 2022). This is a time when young adults face a lot of pressure in the face of finding jobs and careers to embark upon, attending university and trying to achieve high grades, becoming independent and living alone, learning how to be financially stable and also starting families. The age bracket selected is a time of transition which can put a lot of pressure on an individual. In a study (Manescu et al., 2020) which investigated depression amongst young women between the ages of 19 and 35, there were increased reports of suicidal ideation in women under 25 years old, highlighting the severity or risk to the age bracket chosen in this thesis.

A focus on biological females has also been chosen due to the male and female voices possessing varying acoustic features. This is because the vocal folds of women are smaller and lighter, meaning they vibrate at a faster rate. Smaller vocal folds mean a smaller resonating chamber, which results in higher formant frequencies (Lab, 2020). Analysing females separately could help reduce the change of introducing confounding results or masking any patterns which could be more difficult to observe if

biological male and females were analysed together. The focus is also on those in higher education, as “it is generally acknowledged that there is a great difference of education level between depressed and healthy people” (Wang et al., 2019), which has led studies to use education level as a co-variate when analysing differences in acoustic features of depressed and healthy individuals, to ensure that it does not impact results (Wang et al., 2019). A narrow focus is chosen to ensure the study was unique and focused on a specific angle which is not yet researched in the academic literature. A narrow focus also ensures the study is manageable and has clear outcomes that are measurable.

This thesis is unique and contributes to current knowledge within the field of detecting depression via the voice as it focuses on a specific age group and demographic that has not previously been studied in isolation. With evidence of the differences between male and female voices causing varying results when detecting depression, it is important to analyse the two sexes in isolation in order to uncover any gender specific patterns that may be present. Also, with the changes in acoustic features at various points in an individual's life, ensuring that this variable is controlled for enables confounding results to be minimised. The majority of previous studies within the field have analysed participants across a broad range of ages, with limited research on young adults conducted. Having a narrow focus also allows results from previous studies to be tested, to investigate whether their findings would be valid when analysing the demographic in question.

c. Relevance and Importance

There is a considerable amount of evidence to demonstrate that depression can be detected via the analysis of the human voice, due to differences in the acoustic features between those suffering from depression and healthy individuals. In general, it is observed that psychomotor retardation in the voice manifests when one suffers from depression (France, et al., 2000). This would suggest a slowing down of the voice and a delayed responses to speech. A variety of specific acoustic features of the human voice have been studied to find differences in the voice of individuals suffering from depression and healthy individuals. Acoustic features that have indicated a person has depression in past literature includes loudness (Wang et al., 2019), reduced vocal variability, monotonous speech prosody (Cannizzaro et al., 2004), increased formant frequencies, increased F1 bandwidths, decreased higher formant bandwidths (France, et al., 2000), higher jitter values, higher shimmer values and higher harmonic to noise ratio values (Silva et al., 2021). Several other acoustic features have shown correlations with the presence of depression. These are discussed in the literature review.

The detection of medical conditions via the voice could assist healthcare professionals during the diagnostic process, by empowering them with more patient information to make an accurate and well-informed diagnosis. With experts claiming 10-15% of diagnosed medical conditions to be incorrect

(Graber, 2013), it is crucial healthcare professionals utilize modern digital tools and advancements in the field that provide them with more data on the patient, to diagnose and treat the disease better (Kwo, 2021). The World Health Organization (WHO) have stated medical voice diagnosis' largest application to be within telemedicine, which would increase access to care and medical information (WHO, 2010). Modern mobile phones have the capability to transmit frequencies from as low as 50Hz to as high as 14kHz (Cox et al., 2009), and with human speech falling between 125Hz and 8kHz (Bigpi, 2020), this means voice diagnosis can be carried out remotely. Monitoring health remotely is vital, as currently "the lack of resources and adequately trained practitioners are critical barriers to effective depression diagnosis" (Almaghrabi et al., 2023).

With the advancement of technologies to analyse sound signals, and an increasing scope to run large-scale data analysis models, the field of voice diagnosis is growing at a rate of 14.5% each year, with the market value expected to reach \$2.5 billion by the end of 2023 (Future, 2021). An increasing amount of the population now has access to smartphones, allowing for the voice to be recorded with much more convenience. "When it is mature, voice analysis would be to health sciences what x-ray is for medicine" (Singh, 2018).

d. Questions and Objectives

As discussed, depression affects a substantial number of people worldwide. Further to this, young adults are most at risk of depression, specifically women. This comes at a time when there are a lack of resources and professionals available to diagnose depression. Concurrently, new advancements in technologies have displayed potential links to show the voice can be used as a diagnostic indicator and can be observed and analysed remotely. This study seeks to ascertain whether depression can be detected via the voice of biological females between the ages of 18 and 26 in higher education through analysis of the voice. This research will be conducted through remote voice studies of extended vowel sounds and readback speech, analysing if there are changes in the acoustic features F0, F1, F2, VOT, PV and PL between those suffering from depression, and those that do not, and have never suffered from depression. Each hypothesis was chosen based on the results in previous studies of using the voice to detect depression.

Hypothesis (H1): The F0 of individuals with depression will be higher than that of healthy individuals as well exhibiting a lower F0 variance across both vocal tasks.

Hypothesis (H2): F1 frequencies will be higher amongst individuals with depression.

Hypothesis (H3): Participants with depression with exhibit more variance of F2 frequencies.

Hypothesis (H4): The VOT of those with depression will be longer than healthy individuals due to an increased inertia in the vocal folds predicted upon the presence of depression.

Hypothesis (H5): The PV will be greater amongst healthy individuals.

Hypothesis (H6): PL will be longer in those individuals with depression.

Null Hypothesis (H0): The presence of depression will cause no changes in all the acoustic features in the voice of biological females in higher education between the ages of 18 and 26.

e. Overview of Thesis Structure

This thesis will first discuss the findings in the past literature that are relevant to this study, followed by the methodology used, to explain how the study was executed and data was gathered. The results of the voice studies conducted will then be presented, with a discussion of how these results link to previous findings in the literature and how these results can be built upon or improved further. Limitations of the study will then be listed, and finally conclusions will be drawn to determine whether the initial H1 was met, or whether H0 was true.

3) **Literature Review**

a. Introduction to Literature Review

In recent years, there has been an increase of published literature in the field of voice diagnosis. Many other health conditions as well as depression have shown signs that the voice can be used as a diagnostic indicator, from Co19 (Laguarda et al., 2020), Brain Cancer, Multiple Sclerosis (Huckvale & Beke, 2017), Suicidal Ideation (Shinohara et al., 2018) to Alzheimer's Disease (Laguarda et al., 2020). The increase in published papers over recent years has been a result of advancements in technologies that enable the voice to be analysed with higher levels of accuracy. As demand on healthcare grows, there is a growing interest in the field of voice diagnosis due to its potential to assist in the diagnostic process. Having access to accurate remote diagnostic tools are vital to reduce this burden (Hamid et al., 2020). A shortage of GPs has also increased this demand. In England, a decrease in qualified GPs was observed, from 0.52 GPs for every 1000 patients in September 2015, to 0.46 GPs in March 2021 (Dyson 2021). Using the voice is advantageous to meet this demand as it can be observed remotely and is non-invasive, meaning medical instruments are not required to be used on the body. Conventional screening systems that use invasive tools need to adhere to stringent regulations, whereas voice screening does not. (Coherent Market Insights, 2019), making it a much simpler screening tool to work with.

b. Current Depression Diagnosis

A statistic shows that "only 47.3% of mental health cases are detected accurately" (Kesari, 2021). When this is compared to diagnosing depression via the voice, some models have exhibited over an 80% accuracy (Kesari, 2021). "At present, the main ways to evaluate mental health are consultation with a physician or other experts, and self-administered questionnaires such as the General Health Questionnaire 30 (GHQ30) BeckDepression Inventory (BDI) and others" (Nakamura et al. 2015). These questionnaires are also capable of identifying the severity of depression that an individual has, from mild, to moderate, through to severe. Diagnostic methods such as these present benefits when there is a lack of trained professionals to diagnose depression, however, they risk introducing reporting bias, as a patient could unintentionally understate or overstate their condition (Nakamura et al., 2015). Alternative ways to detect depression are also available, through extracting saliva, blood, or conducting electrocardiograms and electroencephalograms (Higuchi et al., 2018). These are invasive however, which means expertise to conduct these screening methods are needed, making it more expensive for the healthcare system and potentially the patient. Invasive screening methods also come with more risk of complications.

The question may arise, 'why are we not capable of paying attention to the sound of our own voice for signs of depression?' Reporting bias is one of these reasons, however another reason is due to a process known as corollary discharge. This essentially means that the auditory cortex shuts down when we speak, meaning a person will hear their voice but will not actually listen to it. Evidence suggests that this could be to reduce the energy that the brain would expend in analysing the voice, as an individual already is aware of how they sound (Kleinberger, 2017).

On the other hand, detecting depression by the voice is quick, low-cost, objective, scalable and accessible by anyone, even from remote locations with limited healthcare and technological infrastructure. When implemented into automated systems in healthcare practices with the assistance of Artificial Intelligence (AI), this would enable doctors to have access to the patient's voice data before the patient attends an appointment. "By being continuously available for patient's, AI precludes the need to schedule an appointment. By accurately pre-screening patients, it saves precious bandwidth in the mental health system" (Kesari, 2021).

c. Acoustic Features and Depression

Various acoustic features have shown potential for diagnosing depression. Whilst there are a considerable number of research papers which do not disclose the specific acoustic features that were measured in their depression detection models, in this section the plethora of features that have shown correlations with identifying depression will be explored from those papers that do disclose the acoustic features. Understanding what specific acoustic features have been used in other research not only allows for comparisons to be made with this thesis, but it also gives an understanding of how results were achieved by others to detect depression.

A paper published in 2019 (Wang et al., 2019) found loudness to be the most significant indicator of depression when the voice was used to determine the differences in vocal characteristics between individuals with depression and healthy individuals. F0 was also observed to show a correlation when looking at the differences between the sample and control group. In this study, 47 patients with a form of depression, known as Major Depressive Disorder (MDD), were recruited from a hospital specialising in mental health. These patients had been diagnosed with MDD by mental health professionals, which ensured there was a reduced chance of reporting errors. 57 healthy individuals were recruited as part of the control group. There was not significant variance between the age and gender of participants in both groups, however the control group had a higher education level, which resulted in education being implemented as a co-variate. To obtain the acoustic features measurements, participants completed vocal tasks which were a combination of spontaneous speech and read back speech. The first vocal task included participants watching a video and then explaining which character had the strongest impact on

them. This was then followed by a question-and-answer (QA) task which focused on describing the emotional scenes in the video. Participants then read back three paragraphs of text, each assigned a single emotion type and containing 140 words each, and finally describing or commenting on pictures displayed to them. Results demonstrated the loudness levels of individuals suffering from depression were significantly lower than healthy individuals. For F0, results depicted the acoustic feature was notably lower in those individuals with depression. Researchers predicted the reason for this could have been related to the reduced muscle tension in depressed individuals. As the tension in the vocal fold's increases, so does the velocity of the vibration, leading to a higher F0. This would be the opposite when there is a reduction in tension, leading to a lower velocity and a lower pitch. Awareness was given to the fact that the sex of the individual influences the F0 and therefore could alter the relationship between F0 detecting depression, as the acoustic features of males and females was done together.

In a study (Calić et al., 2022) aimed to uncover the relationship between acoustic features and the presence of depression, Jitter and Voice Amplitude Variation (vAm) was seen to manifest differently in those individuals suffering from depression. A total of 51 participants enrolled onto the study, with 18 participants having been diagnosed with depression (mean age = 51.83, SD = 9.357), and 33 participants without depression. The control group was further split into two, with 24 participants classified as healthy individuals, making up control group 1 (CG1) (mean age = 24.25, SD = 4.286), and nine individuals diagnosed with a psychogenic voice disorder, making up control group 2 (CG2) (mean age = 46.44, SD = 9.671). Dividing the control group into two could have helped understand the differences in acoustic features across the three groups, and potentially helped to understand with more accuracy whether depression was a cause of the changes in acoustic features. VTs performed by participants included sustaining the vowel /a/ for 3-4 seconds. This VT was chosen to avoid the fluctuations in speech that spontaneous speech may cause. The results observed Jitter values to measure the highest in the sample group at 3.35, followed by CG2 measuring 1.63 and CG1 measuring 0.55, meaning individuals with depression partaking in this study had a greater sway in pitch. The acoustic feature vAm also was the highest in the sample group, indicating more variation in amplitude in those with depression. The researchers indicated needing a larger sample size to generalise results. It was also noted that subgroups for depression severity, considering variables in the acoustic signatures of an individual's sex, history of smoking and medication use could have been considered to improve the accuracy of the results. Looking at the wide differences in the age of the participants across the three groups, this could have also been considered, as it is known that there is a reduction in harmonics with an increase in age (Singh, 2019).

Whilst most studies have analysed the voices of male's and female's together, a research paper (Low et al., 2011) analysed the sex of participants separately when detecting individuals suffering from depression. As a result, researchers observed differences between the male and female results. The study focused on detecting depression in adolescents from 14 to 18 years old. A total of 139 adolescents were recruited (93 females and 46 males). 68 (49 females and 19 males) out of the 139 participants had

MDD according to the DSM-IV, making up the sample group. The other 71 participants (44 females and 27 males) were healthy individuals not suffering from depression. VTs were designed to capture spontaneous speech in a natural environment from three varying interactions, Expressed Positive Interactions tasks (EPI), Problem-solving Interactions tasks (PSI), and a Family Conflict Interactions task (FCI). The model used to identify those with depression analysed acoustic features in categories. The first category being TEO (Teager Energy Operator) based features, which can measure the energy of a sound signal at a specific point in time. The second category being Cepstral features, which infers representing the short-term power spectrum of the frequencies in the sound signal. The third analysed Prosodic features, which would include acoustic features such as F0, Formant Frequencies, Jitter and Shimmer. The fourth analysed Spectral features such as Spectral Flux and Power Spectral Density. The final category looked at Glottal features, which would look at acoustic features such as Glottal Source Spectrum, which relates to analysing the sound of the voice before it is impacted by the vocal folds. TEO based features identified those participants with depression with the greatest accuracy when compared to all other categories of features, achieving a classification score of up to 86.64% in males and up to 78.87% in females. Having a much larger number of female participants in the sample population could have influenced these results however, therefore a similar number of males to females would have enabled a more accurate picture as to whether the model was better at detecting depression in males.

Further studies have looked in more detail at the differences between detecting depression in males and females. One study (Cummins et al., 2017) investigated this difference, with a specific focus on the differences in the Formant features between male and females extending English vowels when detecting depression. A total of 28 individuals (12 males and 16 females) with depression (assessed using the Patient Health Questionnaire – PHQ-8) and 114 individuals (67 males and 47 females) without depression were recruited. Upon participants voicing English vowel sounds, F1 and F2 acoustic features were extracted. Results found that male participants with depression had a lower F1. The opposite was true for females, however, where the F1 in those with depression was observed to be higher. To build on this, another study published in the same year observed identical results (Vlasenko et al., 2017). A study (Hashim et al., 2017) came to discover general differences between the sexes also, when their model for detecting depression was observed to be more accurate when identifying the presence of depression in male's when compared with female's. Researchers noted that there needs to be separate models used for depending on the sex of the individual.

The prevailing practice in the literature is to analyse male and female participants in the same study, whether that is done together or analysed separately during the data analysis stage, with most studies having a mixture of males and females in the sample population. Few studies have focused entirely on the analysis of one sex. A study (Pan et al., 2019) focused on analysing female only voices to detect depression, due to the physiological and biological differences between males and females, resulting in

changes in acoustic features between the two groups. The study gathered responses from two datasets. In the first dataset there was a total of 1132 participants, 584 individuals with depression and 548 individuals considered healthy and without depression. The second dataset had a total of 904 participants, 500 with depression and 404 healthy individuals. There was some overlap in these datasets, with 410 participants appearing in both datasets. All participants were Chinese and between 30 and 60 years old. Demographical data was gathered from each participant, including their age, accent, education, occupation, marital status, and social class. VTs varied depending on the dataset, but included a variety of tasks, such as answering questions, text reading and describing pictures. The average VT length measured 10 seconds. Acoustic features were extracted from the VTs, including intensity, loudness, zero-crossing rate, voicing probability, F0, F0 envelope, eight-line spectral pairs (LSP), and MFCC's. All these acoustic features were built into the model, so there was no information related to how each acoustic feature performed and which feature detected depression with the most significance. The detection of depression was first done based on demographical data alone (based on the fact that certain demographic groups are more or less at risk to developing depression), which was able to identify depression with an accuracy between 69-75%. The detection of depression was then done based on voice data alone, which achieved an accuracy of between 75-80%. Finally, the demographic and voice data were used together to detect participants with depression. This achieved a classification accuracy between 78-81%, demonstrating that the study was able to detect depression to a significant degree, even when voice data alone was used to detect depression.

Various other acoustic features have shown significant results when being used to detect depression. One such feature is Mel-frequency Cepstrum Coefficients (MFCC's), which in broad terms, measures the distribution of energy across the frequency range of a sound signal. This shows where the prominent energy is in the frequency spectrum of the voice, and similarly where there is less energy. One of the studies which observed this (Taguchi et al., 2018) investigated the discrimination between healthy individuals and those with MDD. A total of 72 participants took part, with 36 having MDD (22 males and 14 females) who were recruited from a psychiatric hospital, and the other 36 classified as healthy individuals (16 males and 20 females). The mean age of those with depression was 44 years old with an SD of 16.3. The mean age in the healthy individual's group was 38 years old with an SD of 10.4. Two VTs were carried out in Japanese, the first VT was a Number Reading Task, involving participants reading a sequence of ten numbers. The second was a Verbal Fluency Task (VFT), where participants voiced as many words beginning with the letters /a, /u and /o in a timeframe of 30 seconds. The voice was also analysed before and after this task, as VFT was observed to activate the frontal lobe of the brain. This area of the brain is connected with depression due to it being related to emotional expression. Results of this study observed that, upon measuring MFCC 2, there was less spectral energy around 2000-3000Hz in the sample group when compared to the control group, which aligns with the clinical perception of individuals with depression sounding 'muffled'. Using the MFCC 2 feature alone to detect depression

exhibited an accuracy of 80%. This marks an interesting area of the spectrum of frequencies as the human auditory system is naturally more sensitive to frequencies around 2-4kHz, meaning it could be easier for humans to pick up on changes in this area of the spectrum than it would be for other frequency bands. A reduction in frequencies around 2-3kHz was also seen to link to the sound of the voice being perceived as muffled by listeners. F0 was also analysed in this study, however no relationship was found between this feature and the detection of depression. Researchers noted however that the silence between voice samples could have impacted F0 measurements.

Despite some studies being published several years ago, they still hold relevance even today. One study (Kuny & Stassen, 1993) which stood out, investigated the changes in individual's voices suffering from depression as they recovered. The sample group consisted of 30 hospitalized patients between the age of 24 to 85 (12 male and 18 female) with a mean age of 47.9 years old and a SD of 16 years. The inclusive criteria meant that anyone with any type of depressive disorder was recruited, resulting in the recruitment of patients suffering from affective psychoses, schizoaffective psychoses, and depressive states. Whilst the study measured the differences in those recovering from depression and how these acoustic features changed as they recovered, there was also a control group of 192 healthy participants which acted as a baseline for those recovering from depression to be analysed against. The VTs included counting from 1-30, and reading back a simple passage of text from a children's book which was defined as emotionally neutral. Voice data was taken over a two-week period, with VTs performed on Monday, Wednesday, and Friday mornings. Various acoustic features were then extracted from the voice data, including F0, F0 amplitude, F0 6dB bandwidth, Loudness and Pause Length (PL). When analysing PL, pauses shorter than 250 milliseconds were disregarded. The most significant finding demonstrated that as a patient recovered from their depressive condition, there was a reduction in the mean PL, which also led to a shorter overall recording time. No results were mentioned which directly related to how the mean pitch of F0 differed as patients recovered, however the F0 intensity was seen to decrease and the dynamic variation of the F0 pitch increased as patients recovered from depression. Researchers noted that the antidepressants patients received during their recovery could have led to changes in the voice, as the medicine had a high anticholinergic potency, potentially impacting results. It should also be noted that the wide age range used in this study could have impacted results due the acoustic signature of an individual being somewhat dependent on age.

A further study (Liu et al., 2017) focused entirely on analysing whether the PL acoustic feature was effective when detecting individuals suffering from depression. As a secondary aim, the study attempted to uncover which VT was best for measuring PL when detecting depression. A total of 92 participants made up the sample group (38 males and 54 females), all of whom were all recruited from a hospital. The control group was again made up of 92 participants with the same amount of male and females, along with matching ages and education level. All participants were Chinese and aged between 18-55 years

old. If patients were currently suffering or did have any sort of psychotic disorder, severe somatic disease, drug, alcohol abuse or were pregnant, they were excluded from the study. There was a total of four VTs, which included an interview, reading a passage, reading out words and describing pictures. All VTs were conducted in Chinese. VTs showed a mixture of both spontaneous speech and readback speech. Read back speech involved reading a passage of text from Aesop's Fables, which has been used in various other multilingual research papers. A total of three acoustic features were measured from the voice data, Recording Time (RT), Phonation Time (PT) and Speech Pause Time (SPT), which is identical to PL as referred to in this thesis. As in the previous study analysed, pauses shorter than 250 milliseconds were disregarded when calculating SPT. When analysing results obtained for the VT which involved interviewing participants, the sample group measured a higher mean and SD for all acoustic features (RT, PT and SPT) when compared to the control group, with RT and SPT showing the greatest significance. VTs for read back speech and the picture description observed differences between male and female measurements, but no significant findings related to detecting individuals with depression. The final VT which included participants reading back words showed no statistical differences when it came to all acoustic features, and thus it was concluded that this VT is not suitable in detecting individuals with depression. A further analysis was done to understand the effect of the medicine that some participants in the sample group were currently taking. It was observed that those taking medication exhibited a shorter SPT when compared to depressed individuals that were not taking medication, and this was more pronounced for female participants. Conclusions drawn state that SPT was the most effective acoustic feature to diagnose depression when measured from spontaneous speech.

d. Detecting Depression Severity

The presence of depression occurs in varying degrees, with symptoms ranging from mild, to moderate, to severe. Whilst detecting the severity of depression was not the focus of this thesis, it is important to mention, as being aware of the acoustic features best suited for the detection of depression severity could explain the results from this study in more detail.

Various studies have investigated the identification of depression severity. For example, a study (Yang et al., 2013) used various prosodic features (F0 mean, F0 Variability, Pause Length, Switching Pause mean) to estimate depression severity. The 57 participants that took part in the study all had MDD and were recruited from a clinical trial for the treatment of depression. Participants had ages varying from 19 to 65 (mean age = 39.65 years) with a split of 34 women and 23 men. One of the most accurate acoustic features used in the depression severity detection model was Switching Pause (the time in between when one person stops speaking and another person starts during a conversation), which observed that as depression severity increased, an individual's Switching Pause in a conversation becomes longer and more variable in length. This feature alone was able to predict depression severity with an

accuracy of 69.5%. This was compared with listeners being able to perceive the severity of depression with a 73% accuracy, inferring more work on the detection model would be needed to outperform the human perception of depression severity.

A significant study (Mundt et al., 2007) which aimed to investigate the differences in vocal characteristics of individuals suffering from various degrees of depression (from mild to severe) observed significant differences related to the severity of depression. A secondary aim of their study was to understand the impact that depression treatment had on vocal characteristics. This was a longitudinal study which took place over a period of six weeks, with voice data captured each week. A total of 35 participants (mean age = 41.8 years) took part in the study, which comprised 15 men and 20 women. All participants had depression and were referred by a physician. They had also just started on pharmacotherapy and/or psychotherapy and were all over 18 years old. VTs included an interview, which asked questions related to the participants mental, physical, and emotional health. The second VT involved participants counting from 1-20 and reciting the alphabet. Other VTs included extending the vowel sounds /a, /i and /u for five seconds, rapidly voicing the syllables /pa ta ka/ for five seconds and reading back the '*Grandfather Passage*' (as used in previous studies). This was therefore a mixture of spontaneous speech, read back speech and extending sounds. The Coefficient of Variation (COV) from F0 in the read back passage was measured. Acoustic features extracted from other VTs included PL, Total Pause Time, Total Recording Duration, Vocalisation Time, Percent Time Pausing, Speaking Rate, Syllable Durations, Vocal Intensities and Syllable Rate. Results show that COVs of F0 and F2 did not exhibit any statistically significant results, F2 COV however, had a greater variance as depression severity increased. This did not always exhibit statistically significant measurements, making it unreliable as an acoustic feature to detect depression severity without conducting further research. Other acoustic features which did show a correlation with depression severity included Total Pause Time and Pause Variability. Total Pause Time exhibited a higher correlation during automatic VTs (counting, reciting the alphabet and passage reading) when compared to spontaneous VTs. For Pause Variability however, the opposite was true, with it exhibiting a greater correlation with depression severity during the spontaneous VTs when compared to automatic VTs. Overall patterns observed stated that individuals with a higher severity of depression had longer recording times, due to an increase in PL and a greater variability of their PL measurements. Slower speaking rates were also seen to have a correlation, meaning the slower a participant spoke the more severe their depression was. The secondary aim of the study was to observe the changes of acoustic features in response to depression treatment, F0 COV increased as participants that were receiving medication for depression responded to the treatment. This would infer that as an individual recovers from depression, there is a greater variability in their F0.

Another study (Mundt et al., 2012) which aimed to uncover the acoustic features of the voice that detects the severity of depression did so by measuring various features, including PL, Total Pause Time,

Speaking Rate Number of Pauses, F0, F1, F2 and several other acoustic features. A total of 165 participants (mean age = 37.8 years, SD = 12.5 years) took part, all of whom had MDD. The sample population comprised of 61 males and 104 females and were all between the ages of 18 and 65 years old. From the 165 participants, 39 did not continue the study, and a further 20 had missing voice data, leaving 106 participants with voice data to measure from. Voice data was taken over the course of several weeks, making it a longitudinal study. There were four VTs performed by participants. The first VT was free speech, otherwise known as spontaneous speech, where participants described various experiences they had had over the week prior to recording. The second VT was automated speech, which involved reciting the alphabet and counting from 1-20. The third VT was the read back of the '*Grandfather Passage*'. This passage contains 175 syllables and 132 words. The final VT involved participants extending the vowels /a, /i, /u, and /ae for approximately five seconds. The extension of these vowels was captured to measure F0, F1, and F2. Significant patterns were observed when analysing the voice. Those individuals with more severe depression had longer recording times, an increased PL, as well as a more variable PL and slower speaking rates. F2 COV variability however was observed to be one of the most important acoustic features when detecting depression severity, with F2 COV increasing as depression became more severe. This aligns with previous research (Mundt et al., 2007) published by the same author some years before.

e. The Impact of Age on the Voice

Age is known to have an impact on the voice, with a change in acoustic signatures at various points in life. During a person's youth, the voice is rich in harmonics because of the presence of collagen and elastin in the body. When a person ages, collagen and elastin lessen, reducing the harmonics of a person's voice (Singh, 2019). It is said that "at the beginning and the end of our life, male and female voices are very similar and are difficult to distinguish. In the middle of our lives, our voice becomes a marker of our fluid identity" (Kleinberger, 2017). The fluid nature of the voice at different ages can present difficulties when measuring the voice, which could lead to confounding results, therefore the impact of age on the voice needs to be considered.

One study (Kuny & Stassen, 1993) where the primary aim was to investigate how the characteristics of the voice changes in individuals recovering from depression, haphazardly observed that as the age of an individual increased, there was a decrease in the F0 pitch and amplitude, which suggests as an individual ages, the pitch of their voice becomes lower. On the contrary, the F0 6decibels(dB) bandwidth was seen to increase with age. The F0 6dB band is related to the range of frequencies around the F0 peak level where it drops by no more than 6dB. This would infer that the pitch of an individual's voice varies more as they get older, meaning there is less stability and clarity in the pitch of their voice. This logic can be

understood, as the speaker could have less control of the voice due to parts of the vocal system naturally ageing.

Another study (Lortie et al., 2015) analysed the effect of age on the voice, specifically the impact of age on amplitude and frequency acoustic features of the voice. A total of 81 participants (46 men and 35 women) took part, all of whom were Canadian French native speakers. All participants were between the ages of 20-75 years old (mean age = 54.63 years, SD = 17.57 years). The education level was also stated, with the mean years of education being 17.76 and an SD of 3.5 years. There were two VTs completed in this study. For the first VT participants were asked to sustain the vowel /a/ in their normal voice and amplitude. They were then asked to voice the same vowel (/a/) for an extended time of around 3 seconds with the lowest amplitude possible (whilst still voicing the vowel), the highest amplitude (without introducing distortion), the lowest pitch possible and the highest pitch possible. When measuring the extended vowel sound, a stable section of the vowel was chosen, which would avoid any unnatural sways in pitch during the process of recording. The second VT consisted of participants narrating the stories of 'Red Riding Hood' and the 'Three Little Pigs'. This was not read from a book, but rather the participants were asked to recall the stories from their own memory. This VT could be considered a mixture of both readback and spontaneous speech. However, this VT is also heavily reliant on the individual's memory of the stories, which may introduce some inconsistencies in reporting if acoustic features such as PL (or similar features) are measured, as some participants may recall the stories better than others. Several acoustic parameters were measured from the two VTs, including Minimum F0, Maximum F0, Mean F0, F0 SD, Mean Amplitude, Amplitude SD, Duration of Voiced Utterances, Jitter, Shimmer and HNR. When data analysis was done, participants were split up into three groups, young (from 20-39 years), middle-aged (from 40-65 years), and old (from 66-75 years). The general patterns observed was that as age increases, there is less stability in the voice, inferring a lack of control, which could link to ageing of the vocal apparatus. More noise is also measured in the voice as age increases, which suggests abnormal vibration of the vocal folds and turbulent airflow in the vocal system. Having an increased amount of noise in the voice signal (lower HNR values) would mean the clarity of an individual's voice is reduced as they age. Specific findings include an increase in Jitter with an increase in age, as well as an increase in Shimmer and F0 SD. This would mean the voice has inconsistent fluctuations in frequency and pitch, again suggesting less stability in the voice. When analysing the voices of men and women separately, there was a considerable drop in F0 in women, but no change in the F0 of men. There were no clear changes observed in features related to amplitude however, which suggests little hinderance in producing loud sounds as an individual's age increases.

f. The Impact of Educational Level on the Voice

Whilst no academic research has been found on the specific impact of education level on the voice, there are multiple voice studies (Wang et al., 2019, Lortie et al., 2015) which investigated changes in acoustic features upon the presence of depression or from other variables that have used education level as co-variate to avoid potential confounding results. This infers that depression levels could vary depending on the educational level of an individual.

Graduate students are known to be one of the demographics most at risk to developing depression. A statistic which states 39% of graduate students studying a PhD or Masters have depression (Evans et al., 2018). This is a significant finding, and already supports the evidence to control for education level in voice studies. In a study of 50 PhD students (Gin et al., 2021), it was observed that students noted research increased levels of depression, whereas teaching during their graduate education had a positive effect. A further study (Matacotta, 2020) observed that students in the first year of their undergraduate studies, as well as the last year of undergraduate, and graduate students had significantly higher levels of depression. The study also draws links to how students use alcohol and other substances to cope with the stress and anxiety they are experiencing. This would mean further research is needed to understand the level of impact drugs and alcohol use could be having on the number of registered cases of depression in higher education, and whether this could be a major cause. Besides this, with various links observed between the change in acoustic features of the voice and the presence of depression, it warrants a focused study of depressed individuals in higher education, due to the rates of depression being much greater than in the general population.

g. The Impact of Other Variables on the Voice

There is a myriad of variables that can affect the human voice. These variables are important to consider as anything which could affect the voice could in turn interfere with the results of this study. Isolating the effects of variables is needed as this will give more confidence that any changes in acoustic features are a result of the presence of depression or the health condition which is being tested. As discussed previously, these variables could be the presence of depression or other health conditions, as well as the sex and age of an individual. Several other variables can also affect the acoustic signature of an individual's voice, such as accents, personality, and the presence of drugs (Singh, 2019). Many variables impact the nervous system, and due to the nervous system and the human voice being tightly connected, an investigation on the effect on the voice is warranted (Singh, 2019). Influencing variables on the voice can be divided into short and long-term influences. Short-term influences are those which can alter the voice for a certain period but may not have a lasting impact. These could include variables such as caffeine, drugs, medical conditions, or fatigue. For example, as seen with depression, characteristics of the voice can revert to a healthy state upon the response to medication (Mundt et al., 2007). Long-term

influences, however, are usually those factors which have a lasting impact on the voice and could remain somewhat the same for a lifetime. These are factors such as the sex of an individual, the accent or native language.

Native language has been observed to have a significant impact on the acoustic signature of an individual's voice. This is heavily based on whether a person's native language is a tonal or stressed language. In a study (Eady, 1982) conducted to investigate these differences, significant variations between tonal and stressed languages were found. A specific focus in this paper was to uncover whether the patterns of F0 were different between the two groups. The study recruited a total of 14 male students between the ages of 22 and 33 years old. Half of the sample populations native language was American English (a stressed language), whilst the other half were native speakers of Mandarin Chinese (a tonal language). There was one VT performed in this study, which involved both groups reading back a narrative text. Native English speakers voiced the passage in their own language, which contained 400 syllables, whilst the native Mandarin speakers voiced a translated version in their own language, which contained 350 syllables. Upon recording, F0 measurements were taken at 10 millisecond intervals throughout the text. Results demonstrated that the range which the F0 varied in pitch by was similar for Mandarin and English, however Mandarin speakers exhibited fluctuations in the F0 more frequently than English speakers. It should be noted that this was observed in the read back of a passage of speech that was largely devoid of emotion. Researchers highlighted that this should also be tested across multiple VTs to see if similar patterns are observed. This study had a low sample population, meaning for these results to hold a greater significance, further studies should be carried out to validate these findings. In general, the results from this study show that acoustic features could be dependent on the native language of the speaker to some degree. Another study (Lab, 2020) which observed the variation in F0 between men and women saw a greater difference in Japanese speakers, whereas less of a difference was observed in English and German. Further analysis saw that a lower F0 amongst Japanese men is seen as more attractive. When they looked at this for Dutch however, no link was found.

When looking at fatigue, a study (Baykaner et al., 2015) was conducted to measure how levels of fatigue can impact acoustic features of the voice. Participants were kept awake for 24 hours when researchers observed an increase in PL during a read speech VT. The variation of the 4th Formant Frequency was also seen to decrease during a VT which involved sustaining vowels. In a similar study (Icht et al., 2020), acoustic features of participants' voices were analysed after both 24 hours of sleep deprivation, and after nocturnal sleep, which is considered a regular night's sleep. A total of 47 healthy undergraduate students took part ranging from 18 to 25 years old. The sample population consisted of 24 males with a mean age of 24, and 23 females with a mean age of 23. Participants spoke Hebrew as their native language. Voice data was gathered from participants at 8am following either 24 hours of sleep deprivation or a night of nocturnal sleep. Three VTs were included in this study, the first, sustaining the vowel /a/, the second,

repeating specific words after hearing them, and the third, repeating sentences after hearing them. Various acoustic features were measured from the voice data, including F0, Voice Intensity, HNR, Jitter and Shimmer. Male and female data was analysed separately, with results showing that HNR values were considerably higher after 24 hours of sleep deprivation than they were after nocturnal sleep amongst females. A higher HNR after sleep deprivation was observed for all three VTs. Higher values of HNR means the harmonics in the voice signal are greater than noise components, meaning the voice would be considered having more clarity, A reduced mean intensity level was also observed in males and females after sleep deprivation in the repeating sentences VT. This study produced significant findings comparing acoustic feature changes after sleep deprivation compared to a normal night's sleep, however, researchers highlighted the need to carry out studies which focus on the changes in the voice after different durations of sleep.

The location where an individual is from is also known to impact acoustic features. A study (Jacewicz et al., 2009) analysed the differences in speakers from Northern (Wisconsin) and Southern (North Carolina) America. The study observed the articulation rate of Northern speakers to be 8% higher in reading and 12.5% greater in spontaneous speech, inferring that those in the Northern state of Wisconsin spoke faster than those in North Carolina. This may be largely dependent on the location the speaker is from, and with whom the speaker is compared against. In a contrasting study, residents from the rural Island of Orkney were compared against those living in the city in Edinburgh. In this study there was no significance found in the articulation rates of participants, suggesting that articulation rates may be dependent on a multitude of factors such as social variables, education, place of residence or occupation (Jacewicz et al., 2009). The influence of the company that a person keeps in their younger years is also seen to have a significant influence on voice. A study found "girls in Glasgow from working-class backgrounds had pronunciation that was more like that of their peers, whereas middle-class girls patterned more with adult women (Lab, 2020).

External sounds can also modify the voice at the time of recording, causing an imprint on the phone signal. This could be anything from birds, to planes, to road traffic, leading to frequency masking or constructive or destructive interference of the sound signal, resulting in the voice not being represented clearly. Some researchers have resorted to carrying on voice recordings in controlled environments such as recording studios to avoid the impact of background sounds. However, this approach does not clearly represent real-world environments where we often do not have control of the sound environment (Gideon et al., 2016). Another study (Higuchi et al., 2017) focused on the impact of environmental sounds, specifically when analysing a person's mental health state. The study conducting this research by measuring utterances from a simulated environment with various types of background noises such as subway trains, aviation noise and restaurant environments. Real-world environments were also used, with sources ranging from road noise to a university environment. Depending upon the environment, the

Signal to Noise Ratio (SNR) alternated from low to high. A high SNR would indicate a large proportion of the sound reaching the receiver is desired and a useful sound signal. A low SNR would indicate high levels of unwanted noise, with not much useful information. It was observed that for a clear measurement of voice data, the threshold was at or below SNR 20dB in all simulated environments tested. Over an SNR of 20dB, voice data was poorly detected. This trend continued when detecting voice data from real-world environments. The results uncovered that sometimes background noise was misconstrued as an utterance, introducing unintentional falsified data.

Any sound signal naturally has an acoustic signature, and therefore can be recognized (Elizalde et al., 2018) Carnegie Mellon University have initiated a database called the Never-Ending Learner of Sounds (NELS) which classifies and stores a wide variety of sounds from five broad categories, including animals, natural soundscapes, water sounds, human non-speech sounds, interior/domestic sounds, and exterior sounds (Elizalde et al., 2018). This NELS database would be beneficial for voice studies as once the acoustic signature of these sounds is known, audio separation can be completed by subtracting the unwanted noise from the sound signal.

As discussed, the voice can be impacted by both internal (age, sex, native language) and external variables (education level, environmental sounds, drugs). However, there can also be a crossover between the two, when an external event causes internal changes in the physiological or psychological state of an individual, impacting acoustic features of the voice. This links to situational and cross-situational events, which has been referred to in another study (Wang et al., 2019). This paper poses the question as to whether abnormalities in those with depression can be detected only in specific situations (situational), or in any situation (cross-situational). Researchers conducted measurements in 12 different speech scenarios and concluded that these vocal abnormalities caused by depression can be detected in cross-situational events.

Examples of external events that could impact the internal characteristics of the voice could be humidity, temperature, times of the day, also events which can have an emotional impact on a person, such as a fear evoking event. For example, increasing temperature conditions causes an increase in air absorption, due to particles being closer together, resulting in more friction, and thus, increased absorption. High frequencies are the most prone to air absorption due to their shorter wavelengths when compared to low frequencies. Air absorption could result in a misrepresentation of the high frequency content in the voice if the meteorological conditions match that of the conditions described. A scenario of when an external event could impact the internal state of a person that would evoke an emotional response, and therefore potentially their voice, was in a study (Tokuno et al., 2016) which analyses the effects of an earthquake on nearby residents' health. Immediately following the earthquake, the number of depressive states measured were highest closest to the epicenter. The number of depressive states then gradually

increased away from the epicenter in small remote areas, meaning as news spread outward from the epicenter, so did the number of depressive states. Knowing the impact that depressive states can have on the voice, this could have resulted in changes in acoustic features. Therefore, it is vital when measuring voice data, that researchers maintain an awareness of the current situation that the speaker is in at the time of the recording, as numerous variables are constantly impacting the voice, which could lead to confounding results.

h. Future Directions

A multitude of future directions for voice research related to the detection of depression have been stated in a recent review of bio-acoustic features of depression (Almaghrabi et al., 2023). The review highlights the need for standardised processes when collecting voice data, with various methods currently used to collect, measure, and analyse voice data. Having a standardised approach to this could mean identical VTs, identical equipment (such as microphones), consistent recording environments with a specific background noise level and voice data formats (uncompressed WAVs etc.). Standardised VTs have been referred to in other studies when researchers have been left unsure about what VTs to measure. “The literature is inconsistent regarding what type of voice sample (e.g., sustained vowels and continuous speech) is most appropriate for acoustic analysis” (Icht et al., 2020). Having this uniformity not only in VTs but across the board would ensure voice studies can be compared to other studies with much more ease, knowing that the recording, measuring and analysis of voice data is consistent.

Another direction in the review (Almaghrabi et al., 2023) highlights the need to monitor longitudinal voice data from participants to ensure changes over time are controlled for. This would ensure that there are no abnormalities in the participants acoustic features on the specific day of recording that might not have been caused by the presence of depression, but rather could be caused by an external variable as discussed earlier. It is stated in the review that this longitudinal data can be measured across intervals of time that are clinically meaningful, rather than random intervals conceived by the researcher which has no significance. Intervals which are clinically meaningful, due to vast amounts of knowledge mental health professionals have, could be chosen based on the onset of depression, the progression of the health condition, various significant times in the treatment of depression, or the onset of response to medication for depression.

The paper (Almaghrabi et al., 2023) also proposes various other areas to focus on in future voice research, such as the importance of larger sample sizes, inferring that future studies within the field should recruit larger numbers of participants to ensure there is an increase in statistical power, resulting in more accurate results. A higher statistical power would also mean the study has more generalisability and attenuates the likelihood of any sampling errors. Finally, the review mentions the implementation of interfaces such as Amazon Alexa and Apple Siri when conducting monitoring of the voice. This would

increase the amount of, and access to voice data available to researchers. However, these technologies introduce many ethical considerations that would need to be addressed, specifically related to the handling of patient data and the access control to this, as for the most part, medical data is confidential. Rigorous regulations would need to be implemented to ensure this data is handled properly, and most importantly, ethically.

i. Summary of Literature Review

Investigation on the impact of depression and other external factors in this section of the paper has shown that changes in acoustic features of the voice can uncover the cause behind these fluctuations if rigorous scientific studies are completed with statistically significant results. Below is a review of the vocal patterns observed because of depression and other variables.

Vocal Patterns Signalling the Presence of Depression:

- Reduced loudness levels (Wang et al., 2019)
- Lower F0 frequencies (Wang et al., 2019)
- Higher measurements of Jitter (Calić et al., 2022)
- Increased Voice Amplitude Variation measurements (Calić et al., 2022)
- Changes in TEO-based features (Low et al., 2011)
- Lower F1 frequencies in males (Cummins et al., 2017)
- Higher F1 frequencies in females (Cummins et al., 2017)
- Less spectral energy around 2000 – 3000 Hz (Taguchi et al., 2018)
- Increased Pause Length (Kuny & Stassen, 1993), (Liu et al., 2017)
- Reduced dynamic variation of F0 pitch (Kuny & Stassen, 1993)
- Increased Total Recording Time (Liu et al., 2017)
- Increased Phonation Time (Liu et al., 2017)

Vocal Patterns Indicating the Severity of Depression:

- Increased length and variability in Switching Pause (Yang et al., 2013)
- Increase in Total Pause Time (Mundt et al., 2007)
- Increase in Pause Length (Mundt et al., 2012)
- Increase in Pause Variability (Mundt et al., 2007), (Mundt et al., 2012)
- Slower Speaking Rates (Mundt et al., 2007)

- Reduction in F0 COV (Mundt et al., 2007)
- Increase in F2 COV (Mundt et al., 2012)

Other variables impacting acoustic features:

- Decreasing F0 pitch and amplitude with an increase in age (Kuny & Stassen, 1993)
- An increase in F0 6dB bandwidth with an increase in age (Kuny & Stassen, 1993)
- Reduction in HNR with an increase in age (Lortie et al., 2015)
- Increase in Jitter, Shimmer and F0 SD with an increase in age (Lortie et al., 2015)
- Drop in F0 pitch with an increase in age amongst women (Lortie et al., 2015)
- Potential more frequent fluctuations in F0 in tonal languages (Eady, 1982)
- Reduced F4 variation as a result of lack of sleep (Baykaner et al., 2015)
- Higher HNR values after sleep deprivation in females (Icht et al., 2020)
- More frequent fluctuations in F0 amongst Mandarin Chinese speakers (Eady, 1982)

Overall, the current state of the art in the detection of depression via the human voice shows promise, with several acoustic features of the voice (such as reduced loudness levels and longer pause times) exhibiting correlations with the presence of depression. However promising results referred to in the literature review are, these studies still lacked the large numbers of participants needed in order to obtain statistically significant results that would be accurate enough to be used in a clinical setting. Studies related to the detection of depression did utilise age as a co-variate. Evidence has shown age to lower the F0 pitch and amplitude as age increases (Kuny & Stassen, 1993). Precision is needed in the study design to ensure all other factors that could impact acoustic features, and subsequently the study's results, have been accounted where possible.

Directions for future research related to the diagnosis of depression via the human voice pose many challenges, due to the variety of factors that could impact the voice, as well depression being a complexed condition, due to depression potentially manifesting differently in each individual. However, once these challenges are overcome, there is potential for the creation of powerful voice diagnosis models, that have a dramatic impact on the wellbeing of patients, whilst reducing the burden on healthcare professionals.

4) **Methodology**

a. Background

A study was carried out to investigate the potential of the human voice to be used as an indicator of depression. Voice data was gathered remotely from participants. The study was run and approved in accordance with the ethics committee at the University of Huddersfield (UoH) (appendix 3). Participants partaking in the study gave their consent via an online questionnaire before commencing, which can be found here; <https://forms.office.com/r/uHsa37gcJy>. Rather than fully disclosing the full purpose of the study, participants were informed that the study was to investigate differences in the human voice in order to avoid any reporting bias. All data gathered was kept on an encrypted hard drive for the duration of the study, and as agreed with participants, will be deleted upon receiving the final grade for the master's degree (or up to 6 months after the submission date).

b. Participants

A total of 17 participants took part in the study. Inclusion criteria included being a biological female between the ages of 18 and 26, and either currently suffering from depression (sample group), or never having suffered from depression (control group). No participants registered that were under 20 years old, resulting in the age range being 20-26 years. The mean age amongst the sample group measured 23.5 years with an SD of 1.77 years. The mean age amongst the control group measured 23.88 years with an SD of 1.96 years. Depression was self-reported, although most participants gave the date in which they were diagnosed with depression by a mental health professional and confirmed that they were still suffering from depression. If the participant met the inclusion criteria, they were sent information on what would be required from them and instructions on how to complete the vocal tasks (refer to Appendix 7.1 for the participant information sheet sent out).

The study focused on biological females, due to the pitch of a female's voice being higher than that of men. As we have discussed, the vocal folds of women are smaller and lighter, meaning they vibrate at a faster rate (Lab, 2020). "In adult speakers, F0 is usually higher in women (200-220 Hz), who typically have short and thin vocal folds, compared to men (100-120 Hz), who have long and thick vocal folds" (Almaghrabi et al., 2023). Analysing the voice of male's and female's side by side would have meant for a more complex analysis, which was beyond the scope of this study based on the low number of male's that registered their interest in the study. Therefore, the study focused on female's only in order to narrow the scope of the analysis and obtain results that were more focused and specific. Age is also known to impact the voice, and thus a specific 18-26 age bracket was chosen. During a person's youth, the voice contains a greater amount of harmonics caused by the presence of collagen and elastin in the body.

When a person ages, the amount of collagen and elastin decreases, reducing the harmonics of a person's voice (Singh, 2019). Therefore, the criteria was set between 18-26 years old to attenuate the influence of age on acoustic feature measurements. Young adults are also one of the age brackets most at risk to suffering from depression, in particular women, with 43% of women aged between 16 and 29 experiencing depressive symptoms, compared with 26% of men in the same age bracket (Office for National Statistics, 2021).

Inclusion into the sample group was met if the participant was currently self-reported that they were suffering from depression. Nine participants made up the sample group, of whom eight had been clinically diagnosed with depression, whilst the other simply stated they were currently suffering from depression but had not been diagnosed by a doctor. All those within the sample group currently suffered from depression, even if they were diagnosed many years previous. Inclusion into the control group was met if the participant had never suffered and was not currently suffering from any form of depression.

Participants were recruited via the UoH, which subsequently led to all participants being higher education students completing a variety of degrees, from undergraduate to doctoral level. Participants were contacted initially via their university email, where they were asked to fill out an initial questionnaire on Microsoft Teams to gather their information to see whether they met the inclusion criteria for the study. Contact was then made if they met this criterion and they were sent further information. The study was designed to anonymise all the participants by removing their name, rather identifying each participant by their participant ID and demographical data. This demographical data was essential, due to the various factors that are known to impact a speaker's voice. Participants were made aware that any personal details including their name and contact information will be deleted, and it was at their discretion if they wanted to disclose this information from the start.

Demographical data was gathered from participants to allow for a greater understanding of the participants current situation, their education level, native language, living situation and even if they had financial worries. Table 1 and 2 outlines the full spectrum of data that was gathered from each participant in the sample and control group. Additional data from participants gave a more holistic overview of the various factors that could have been influencing the participants acoustic measurements at the time of recording. Knowing how each factor imprinted upon the acoustic features in the participants voice was beyond the scope of this study, however, a general awareness of how each factor could impact the voice was maintained. For example, fatigue is known to cause the pause length to gradually increase in read speech (Baykaner et al., 2015), whilst the consumption of caffeine has shown links to the dehydration of the vocal folds, "which would manifest as abnormal voice production" (Georgalas et al., 2023).

Questions on financial worries, living alone and having a job alongside studying are nuanced and therefore there was no current literature which demonstrated evidence that they had a direct impact upon the voice, however, these factors could lead to an increase in stress. This information is important to know as in the field of voice diagnosis it is said that “everything that could influence your physiology and your mind could impact your voice, as speech is a complexed biomechanical process” (Singh et al., 2016). Further investigations into the specific impact these individual factors had on the participants voice could have taken place if there were acoustic feature measurements which stood out as anomalies or outliers.

Having more data about each participant also allowed for the potential to use certain parameters as co-variates during the analysis of the data. Participants could have been grouped based on if they were living alone or with someone they knew for example. However, due to the limited number of participants, no co-variates were utilized in the study, as this would have rendered the group sizes unfit to observe any potential patterns or correlations.

c. Vocal Tasks

When gathering voice data, participants recorded and submitted two vocal tasks which took them approximately 2-5 minutes to complete. Instructions were given as to how to voice the vocal tasks and what to say. Participants were advised to not keep the microphone too close or too far away from their mouth, also to ensure the acoustic environment they were in was low in background noise to avoid any interruptions which could have impacted the reliability of the voice data.

The first vocal task consisted of voicing extended vowel sounds /a, /u and /m for between 5-7 seconds each. Using extended vowel sounds has been popular in past literature due to their ability to go beyond language (Higuchi et al., 2018). The vowel sounds used are universal and can be found in all major languages, meaning they are generally uttered and voiced in the same way, irrespective of each participants native language. The vowel sound /a has been used in studies (Patel et al., 2011) when classifying emotions, as well as another study when /a was found to detect depressive status with greater accuracy than /e and /u. Utilising sounds which do not require the action of the tongue creates a level playing field when analysing the voices of participants with varying native languages (Omiya et al., 2019). With 3 participants in the control group and 1 participant in the sample group having a native language other than English, this was an important consideration.

The second vocal task included the readback of a pre-defined passage known as the ‘rainbow passage’, which has been commonly used in previous voice studies (Shinohara et al., 2016) due to the wide variety of sound combinations from the English language that it utilises (Hidden Braces, n.d.). Having readback voice data allowed for a uniform, standardised approach to the analysis of participants voices, and

enabled a certain degree of control over the words, sounds, and thus frequency content which participants produced. This maintained consistency across speakers, making the voice measurements easier to analyse due to no differences in the content and sounds in the speech. Gathering free form speech data would have led to differences in the words uttered, thus changing the frequency range and content of each person's vocal sample. The rainbow passage which participants read includes a total of 97 words and 96 pauses. The passage can be found below.

'When sunlight strikes raindrop in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.'

Participants recorded all vocal tasks remotely via their personal smartphone. A decision was made that smartphones would be sufficient to capture the voice data and allow for analysis due to modern mobile phones now coming with wideband technologies which transmit frequencies as low as 50Hz and as high as 14kHz (Cox et al., 2009). This is due to mobile phones containing greater amounts of memory, leading to higher sampling rates. The greater the sampling rate the more faithfully the analogue sound signal (participants voice) will be represented in digital form. Shannon's sampling theorem states that the sampling rate/2 will determine the upper frequency limit of the sound signal we are sampling, meaning a mobile phone would require a sampling rate of 28k to represent frequencies up to 14kHz faithfully. Not having a high enough sampling rate can result in frequency aliasing, where high frequencies are falsely represented as low frequencies. With human speech falling between approximately 125Hz and 8kHz (Bigpi, 2020), the upper frequency limit in mobile phones of 14kHz proved reliable for the purpose of this study.

Vocal task recordings were submitted via WhatsApp, Telegram, or email. By using Telegram, participants had the option to remain anonymous, as mobile numbers can be hidden on this platform. Other options to submit voice data were explored, such as uploading voice data to an online platform, but ultimately this was decided against, as these platforms were operated by third parties and relied on the user to sign a contract which permitted the third party to use the voice data on any of their own studies. This would have broken ethical agreements with participants in the study, therefore direct contact via social media/email was utilised.

d. Acoustic Features

Fundamental Frequency (F0), Formant Frequencies 1 and 2 (F1 & F2), Voicing Onset Time (VOT), Pitch Variability (PV) and Pause Length (PL) were the acoustic features chosen to analyse. F0, F1, F2, VOT and PV are all well referenced in the academic literature and have shown positive links when used to detect depression. The less studied acoustic feature was PL, which meant there was room to establish new connections between PL being used in the detection of depression. Average F0 was measured for VT1 and VT2. F1 and F2 were measured for VT1. All other acoustic feature measurements (VOT, PV, PL) were extracted for VT2 only.

e. Acoustic Feature Extraction

Acoustic features were measured using a combination of the audio analysis software Praat version 6.2.23, as well as Python code running Librosa. Before any acoustic features were measured the audio files received from participants were normalised to -0.1dB to ensure the gain settings were consistent for each recording. Audio files were trimmed to ensure silence at the beginning and end of each recording was kept to a minimum. To measure average F0, PV and PL, Python code was used to extract measurements. This was then crosschecked with manual measurements taken from Praat to ensure that the code was accurately measuring the acoustic features, of which it was. To measure F1, F2 and VOT, Praat was used instead of Python, due to these features being more complexed to extract with code.

Extracting Fundamental Frequency (F0)

The code used to measure the average Fundamental Frequency (F0) of VT1 and VT2 can be found in appendix 2.5. The code operates by first loading the audio file, it then reduces background noise to focus on the sound signal in question. Harmonic and rhythmic content of the sound are then separated. The rhythmic parts of the speech are disregarded, due to this generally containing the unvoiced parts of speech. Only the voiced parts of speech are needed, as this contains the pitch, and thus F0 information. The code finds the F0 for every 23 milliseconds of audio. The average F0 is then calculated and printed in Hertz (Hz).

Extracting Pitch Variability (PV)

Code used to extract the PV can be found in appendix 2.6. The code first loads and reads the sound file. A high-pass filter is run to discard unwanted frequencies such as rumbling sounds caused by movements of the speaker, microphone, or other sound sources in the vicinity at the time the speaker is recording their voice. The audio is then separated into frames of 23 milliseconds and the pitch is calculated for each frame. If the pitch is measured above 0, the values are collected and the amount the pitch changes over time is calculated. When the mathematical formula for variance is calculated, values are squared,

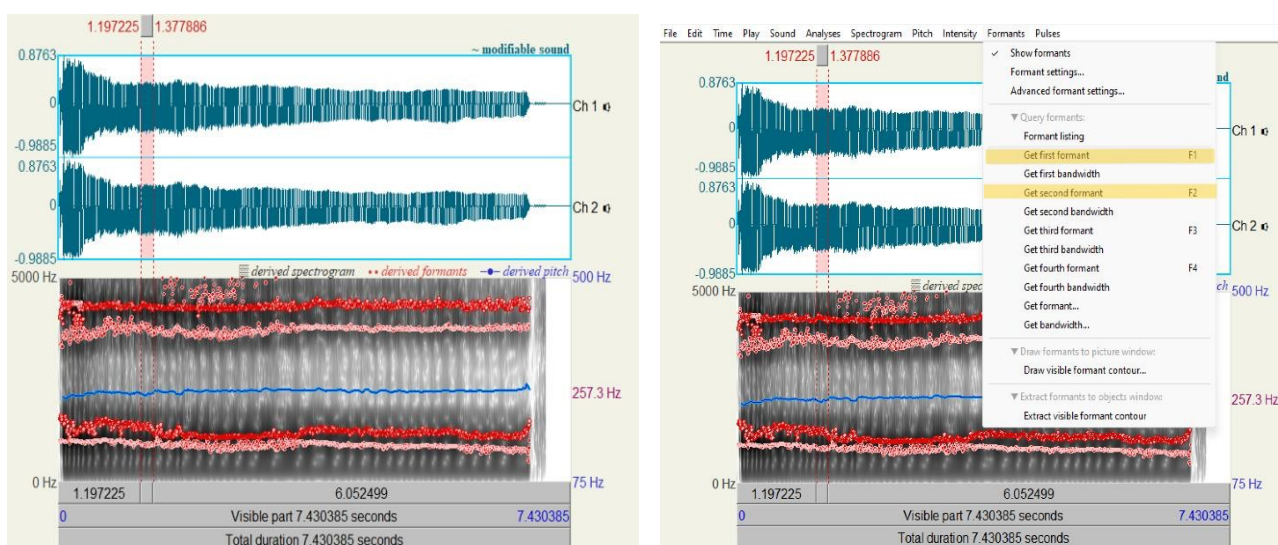
therefore PV is measured in Hz^2 . If the audio file has a high variability, this indicates the pitch of the voice changes considerably, if it has a low variability, this indicates not much change in the pitch of the participant's voice. PV was calculated for vocal task 2 only, due to vocal task 1 based on extending a vowel sound with one pitch.

Extracting Pause Length (PL)

Code used to measure the average PL can be found in appendix 2.7. The code is designed to calculate the amount of time the participant stops talking in between words in the rainbow passage (VT2). The audio file is first loaded, a high-pass filter is then applied to reduce unwanted low-frequency sounds. A silence threshold is then calculated, and any sound which falls under this threshold is considered as silence. The silence threshold which indicated speech marked sounds over 40dB. The points of silence and noise are then marked, and the average length of pause time across all points is calculated in milliseconds.

Extracting Formant Frequencies (F1 and F2)

F1 and F2 were measured using the voice analysis software Praat for vocal task 1. Upon loading the audio file into Praat, the waveform editor was opened, and a steady formant section was chosen to ensure co-articulation was avoided, also to ensure the section chosen had a continuous unchanging pitch, with the vocal task's instructions being to extend the single pitched vowel sound /a/, /u/ and /m/ (separately). The length of the section chosen usually ranged from between 0.1s and 0.2s. F1 and F2 measurements were then recorded for this section. The two images below show the steady formant section being highlighted on the waveform graph in red (left – top), followed by then displaying the F1 and F2 (right).

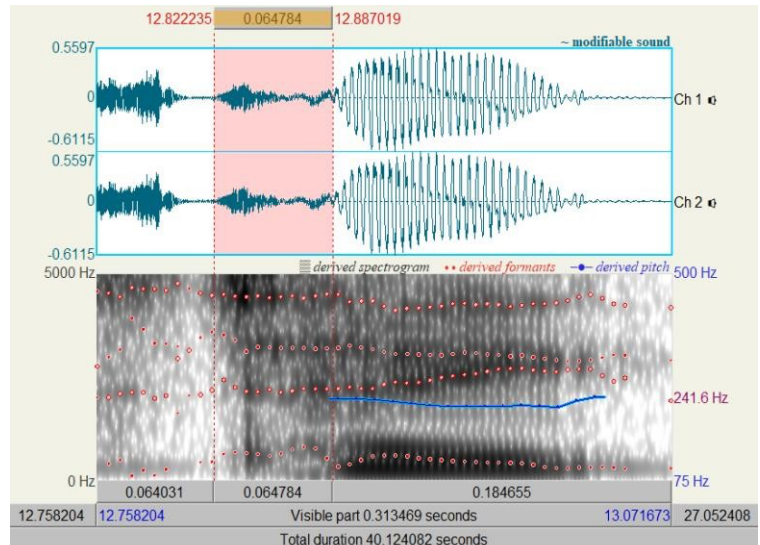


These screenshots are taken from Praat.

This screenshot is taken from Praat.

Extracting Voicing Onset Time (VOT)

VOT was measured using Praat for vocal task 2, which provided a plethora of unvoiced plosive sounds followed by voiced vowel sounds to measure from. The words /take and /pot were chosen from the *rainbow passage* to measure the VOT from. The audio file was first loaded into Praat and the chosen words were located. In the image to the right the word /take is displayed on the waveform graph (top). The VOT was



measured from the start of the unvoiced plosive /t to the onset of the voiced vowel /a. In this instance, the VOT was measured at 64.784ms (highlighted in orange). The same method to extract VOT measurements was again utilized for the word /pot.

f. Data Analysis

Once the measurements were taken, all data was recorded on Excel. The mean values for all 5 acoustic features measured were calculated for both the sample and control group separately. The standard deviation was also calculated for each group, which highlighted to what degree the measurements for each group deviated from their mean value. Finally, the Coefficient of Variation (COV) was calculated to show how much the data deviated away from the mean. This is given as a percentage. A high COV would indicate that data has a wide dispersion around the mean, whereas a low COV would indicate data points are dispersed closely around the mean. COV was also calculated in several studies which are referred to in the literature review. This meant the COV from this study could be analysed against other studies with much more ease.

Due to the number of participants in the study, an observation was made prior to analysing the data to conclude there was not enough statistical significance to draw any meaningful conclusions. A t-test was still carried out to uncover whether any of the acoustic feature measurements exhibited a p-value of less than 0.05. This would demonstrate that there is statistical significance and that the H0 can be rejected (Morris & Elston, 2011). If significant changes are observed between the sample and control group, this would suggest the acoustic feature is an indicator of the presence of depression. Cohen's d was also calculated to understand the effect size. This would show the magnitude of the change (if there was one) between the sample and control group. When analysing Cohen's d, a score of more than 0.2 signifies a

small effect size, over 0.5 a medium effect size, and over 0.8 a large effect size (Panjeh et al., 2023). This gave more insight on the specific acoustic feature was causing a considerable change amongst those individuals with depression.

5) **Results**

The H1 of this study was that the presence of depression will cause changes in one, if not all the acoustic features measured when comparing the sample group against the control group. The H0 is that the presence of depression will cause no changes in any of the acoustic features measured from the voice. The sample population of this study was low; therefore, a t-test was run to observe the statistical significance that the results of each acoustic feature had. The t-test and effect size results can be found below in Table A below.

Table A: Results of t-tests and Cohen's d for acoustic features

Logistic parameter	Sample Group		Control Group		<i>t</i> (40)	<i>P</i>	Cohen's
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			<i>d</i>
F0 VT1 /a, in Hz	217.30	23.79	203.31	22.45	1.25	0.604	0.23
F0 VT1 /u, in Hz	225.95	22.99	203.17	22.7	2.06	0.997	0.06
F0 VT1 /m, in Hz	222.11	26.09	203.87	13.9	1.83	0.857	0.10
F0 VT2, in Hz	190.72	22.99	174.63	22.11	1.47	0.712	0.16
F1 VT1 /a, in Hz	725.21	123.55	706.51	81.62	0.37	0.176	0.72
F1 VT1 /u, in Hz	399.14	54.66	380.64	69.60	0.6	0.298	0.55
F1 VT1 /m, in Hz	312.74	56.53	283.56	56.41	1.06	0.517	0.30
F2 VT1 /a, in Hz	1119.65	219.40	1348.30	157.12	-2.49	-1.186	0.03
F2 VT1 /u, in Hz	1375.02	426.80	1128.67	252.66	1.47	0.691	0.18
F2 VT1 /m, in Hz	1565.44	384.36	1571.96	315.92	-0.04	-0.018	0.97
VOT VT2 /pot, in ms	53.44	13.78	66.16	55.47	-1.13	0.84	0.26
VOT VT2 /take, in ms	57.72	26.55	29.09	15.04	0.22	0.102	0.84
PV VT2, in Hz ²	2402.45	1167	2994.34	964.32	-1.14	-0.549	0.28
PL VT2, in ms	312.56	115.98	676.25	444.64	-2.25	-1.153	0.03

Fundamental Frequency (F0)

The analysis of the F0 acoustic feature (see Table B and C) revealed that the sample group consistently had a higher average F0 when compared with the control group across both vocal task 1 and 2. The greatest difference in average F0 was observed in vocal task 1, when participants voiced the extended vowel sound /u/. The t-test for VT1 and VT2 did not meet the criteria of statistical significance, however for VT1 when extending the vowel /u/, the p-value was calculated at 0.06, only slightly higher than the criteria of 0.05 to show statistical significance. The effect size was also considerably large, with a Cohen's d measurement of 0.997. A large effect size was also observed when participants extended the sound /m/.

Table B: Fundamental Frequency (F0) for Vocal Task 1

Variables	Mean ± SD ± COV (/a)	Mean ± SD ± COV (/u)	Mean ± SD ± COV (/m)
Sample Group (SG)	217.3 ± 23.79 ± 10.95	225.95 ± 22.99 ± 10.17	222.11 ± 26.09 ± 11.75
Control Group (CG)	203.31 ± 22.45 ± 11.04	203.17 ± 22.7 ± 11.17	203.87 ± 13.9 ± 6.82

*Measurement Unit for mean and SD = Hertz (Hz) // *Measurement Unit for COV = %

Table C: Fundamental Frequency (F0) for Vocal Task 2

Variables	Mean ± SD ± COV
Sample Group (SG)	190.72 ± 22.99 ± 12.05
Control Group (CG)	174.63 ± 22.11 ± 12.66

*Measurement Unit = Hertz (Hz) // *Measurement Unit for COV = %

Formant Frequency 1 (F1)

F1 results (see Table D) demonstrated a similar pattern to F0 when comparing the average F1 between the sample and control groups during VT1. The sample group had a consistently higher average F1 than the control group for every extended vowel, with the extended vowel /m/ displaying the greatest difference. However, all F1 measurements exhibited p-values that were greater than 0.05, meaning there were no statistically significant results found. A low effect size was also observed across two extending sounds, with a medium effect observed for the extended sound /m/.

Table D: Formant Frequency 1 (F1) for Vocal Task 1

Variables	Mean ± SD ± COV (/a)	Mean ± SD ± COV (/u)	Mean ± SD ± COV (/m)
Sample Group (SG)	725.21 ± 123.55 ± 17.04	399.14 ± 54.66 ± 13.69	312.74 ± 56.53 ± 18.08
Control Group (CG)	706.51 ± 81.62 ± 11.55	380.64 ± 69.60 ± 18.28	283.56 ± 56.41 ± 19.89

*Measurement Unit = Hertz (Hz) // *Measurement Unit for COV = %

Formant Frequency 2 (F2)

F2 results (see Table E) did not display any consistent patterns across the three extended vowels when comparing the average F2 mean in the sample and control group during VT1. When analysed individually however, extended vowel /a/ exhibited a lower mean in the sample group, a higher mean for extended vowels /u/ and an almost identical mean for the extended vowel /m/. When looking at the t-test results, VT1 when participants extended the vowel /a/ met the criteria to signal statistical significance, with a p-value of 0.03 and an effect size of -1.186.

Table E: Formant Frequency 2 (F2) for Vocal Task 1

Variables	Mean \pm SD \pm COV (/a/)	Mean \pm SD \pm COV (/u/)	Mean \pm SD \pm COV (/m/)
Sample Group (SG)	1119.65 \pm 219.4 \pm 19.6	1375.02 \pm 426.8 \pm 31.04	1565.44 \pm 384.36 \pm 24.55
Control Group (CG)	1348.3 \pm 157.12 \pm 11.65	1128.67 \pm 252.66 \pm 22.39	1571.96 \pm 315.92 \pm 20.1

*Measurement Unit = Hertz (Hz) // *Measurement Unit for COV = %

Voicing Onset Time (VOT)

The mean VOT (see Table F) from the letter /p/ in /pot/ in the control group was longer than that of the sample group. For the VOT from the letter /t/ in /take/, the opposite was true, with the sample group measuring a higher mean value by a marginal amount. T-test measurements confirmed there was no statistical power amongst these acoustic measurements, however the VOT from the word /pot/ displayed a large effect size of 0.84.

Table F: Voicing Onset Time (VOT) for Vocal Task 2

Variables	Mean \pm SD \pm COV (/pot/)	Mean \pm SD \pm COV (/take/)
Sample Group (SG)	53.44 \pm 13.78 \pm 25.79	57.72 \pm 26.55 \pm 46
Control Group (CG)	66.16 \pm 29.09 \pm 43.97	55.47 \pm 15.04 \pm 27.11

*Measurement Unit = Milliseconds (ms) // *Measurement Unit for COV = %

Pitch Variability (PV)

Results can be found in Table 5.1 below.

During VT2, the average PV (see Table G) was observed to be greater in the control group than the sample group by a considerable amount. However, the t-test observed no statistical significance, with a p-value of 0.28. The Cohen's d test demonstrated a medium effect.

Table G: Pitch Variability (PV) for Vocal Task 2

Variables	Mean \pm SD \pm COV
Sample Group (SG)	2402.45 \pm 1167 \pm 48.58
Control Group (CG)	2994.34 \pm 964.32 \pm 32.2

*Measurement Unit = Hertz Squared (Hz²) // *Measurement Unit for COV = %

Pause Length (PL)

For VT2, the average PL (see Table H) between words was greater in the control group. T-test results also showed these measurements to be statistically significant, with a p-value of 0.03. A large effect size was also exhibited, with a Cohen's d of -1.153.

Table H: Pause Length (PL) for Vocal Task 2

Variables	Mean \pm SD \pm COV
Sample Group (SG)	312.56 \pm 115.98 \pm 37.11
Control Group (CG)	676.25 \pm 444.64 \pm 65.75

**Measurement Unit = Milliseconds (ms) // *Measurement Unit for COV = %*

Bonferroni Correction

The results above have been analysed using the standard level of statistical significance, 0.05, meaning there is a 5% chance of attaining a false positive result. Due to the number of t-tests ran, the Bonferroni Correction was applied in order to minimise the chances of a Type 1 error. 14 t-tests were conducted in total. The standard level of statistical significance (0.05) was divided by the number of t-tests (14) to obtain a Bonferroni Correction of 0.0036. Once data is analysed using this correction, this would mean that statistical significance for p-values above 0.0036. If the Bonferroni Correction is used in the analysis of F0, F1, F2, VOT, PV and PL, no p-values portrayed those needed to classify the results as statistically significant.

6) **Discussion**

The study sought to understand whether the presence of depression will cause changes to acoustic features of the voice amongst biological females in higher education. Two VTs were conducted (extended vowel sounds and readback speech) to measure six acoustic features of the voice (F0, F1, F2, VOT, PV, PL), comparing a sample group consisting of nine individuals with depression, against the control group of eight healthy individuals without depression. The mean, SD and COV was measured for each acoustic feature. The strengths of the study include the focus on a specific demographic which has not previously been studied in detail, along with the analysis of acoustic features (VOT) not referenced regularly in the literature related to detecting depression via the voice. A key finding in this study was that notable differences between healthy and depressed individuals can be observed across two acoustic features measured. However, due to the low power, further studies are required to confirm these findings. It should be noted that all results found were applicable to females only and between the age range highlighted.

Fundamental Frequency (F0)

When observing the results obtained from measuring the F0, the mean average F0 was consistently lower in the control group across both VT1 (extended sounds) and VT2 (read back speech) when compared to the sample group. This contradicts previous research, which stated that depressed individuals exhibit a lower F0 than healthy individuals (Wang et al., 2019). The reasons given for a lower F0 amongst depressed individuals in the study referenced was based on an inference that a lower F0 is attributed to a reduction in muscle tension caused by the presence of depression, after an alternative study stated there was a positive relationship between F0 and muscle tension (Scherer, 1986). However, the study (Wang et al., 2019) which observed a lower F0 in depressed individuals analysed both male and female voice data together. With substantial evidence to confirm the differences in acoustic features between male and female's voices, especially related to F0 differences, analysis of this acoustic feature should be done separately to avoid any confounding results. Therefore, this study concludes the mean F0 to be higher in depressed females across both readback speech and extended sounds, however further research needs to be carried out to conclude this acoustic feature as a diagnostic indicator of depression, due to contradictory results and a low power in this study. Future studies should focus on the extension of the vowel /u/ when observing conducting VTs to uncover acoustic differences in the voice caused by depression, due to this feature measuring 0.06 in VT1 /u/ on the t-test, just slightly above the normal criteria for significance of 0.05. This was, however, considerably above the Bonferroni level of 0.0036.

The F0 variation in the sample group on the other hand closely resembled that of the control group in the readback speech task, and when extending the vowel sounds /a and u/. Across these VTs, the COV was slightly lower in the sample group, which is consistent with another study (Mundt et al., 2007) that observed with an increase in depression severity, there is a decrease in F0 COV. This would infer healthy individuals have a higher F0 COV. When extending the sound /m/ however, this pattern was not observed,

with sample group exhibiting a significantly higher variation. Further studies with a greater power need to be run in order to uncover whether the extended sound /m produced confounding F0 COV due to it being a recording error or sample size limitation, or whether there are differences in the acoustic signatures of this sound which explains the reason for the inversed F0 COV pattern observed in the extended sound /m.

Formant Frequency 1 (F1)

F1 was measured from VT1 alone. Across the 3 extended sounds, the F1 mean was higher in those suffering from depression when compared against healthy individuals. This supports the findings of a previous study (Cummins et al., 2017), which stated that higher F1 measurements are found in females with depression. The same study found the opposite was true for males. This indicates the importance of analysing F1 separately when detecting depression. When looking at SD and COV measurements, no clear patterns were observed. Variance was greater in the sample group when extending the sound /a, less when extending the sound /u, and similar when extending the sound /m, however these were all by a marginal amount. With no clear patterns observed in the reviewed literature, a relationship cannot be drawn between the F1 SD and COV for the detection of depression. A greater F1 mean amongst females with depression however does show potential to be utilised as a diagnostic indicator of depression but would need further studies to be carried out as the t-test did not conclude these results as statistically significant, with p-values between 0.3 and 0.72 measured depending on the extended sound.

Formant Frequency 2 (F2)

F2 was also measured from VT1 alone. Previous studies (Mundt et al., 2012) in the literature related to the detection of depression severity have observed an increase in F2 COV as depression severity increases. This would infer healthy individuals would exhibit a lower F2 COV than depressed individuals. When we apply this inference to F2 COV measurements captured in this study, this inference seems to hold true. F2 COV was consistently higher in the sample group across all 3 extended sounds, essentially meaning the F2 expressed a higher variability around the F2 mean amongst those individuals with depression. The t-test measured the extended vowel /a to exhibit statistical significance below 0.05, with a power of 0.03 and a large effect size of -1.186. This demonstrates F0 COV for VT1 could be a promising acoustic feature to detect depression when extending the vowel /a. However, once the Bonferroni Correction was applied, 0.03 was considerably over the 0.0036 significance threshold.

Voicing Onset Time (VOT)

Measurements of VOT were taken after the plosive sound /p and /t and before the onset of the voicing in the words /pot and /take. This ensured that results were achieved from different types of plosive sounds, with /p, being a bilabial plosive, produced by bringing the lips together to create an obstruction, and /t, being an alveolar plosive, produced by touching the tongue against the point behind the upper front row

of teeth. The third category of plosive sounds is a velar plosive when the tongue touches the back of the upper mouth. This type of plosive was not measured.

Results show that the mean VOT and variability was considerably higher amongst healthy individuals for the bilabial plosive /p/. This is supported by the Cohen's d test also, which highlighted that these differences measured a large effect size of 0.84. The VOT was slightly higher in depressed individuals for the alveolar plosive /t/ and the VOT variation was considerably higher for depressed individuals. These results could hypothesise that depressed individuals exhibit a lower VOT for words beginning with a bilabial plosive, whilst a higher VOT for words beginning with an alveolar plosive. However, generalising these findings to all alveolar and bilabial plosives would require a study measuring the VOT of words beginning with /b/ (bilabial) and /d/ (alveolar) also, to ensure that all bilabial and alveolar plosive words have been analysed. However, with no referenced literature related to the connection between VOT and the detection of depression, and with the low power in this study, results should be seen as speculative, and any conclusions would require further research to support the initial findings stated in this study.

Pitch Variability (PV)

The PV was calculated from VT2 to demonstrate the range of fluctuation in pitch in the '*Rainbow Passage*'. PV was not extracted for VT1 as this involved the extension of three sounds with a consistent individual pitch. Results from VT2 demonstrated that PV was higher in healthy individuals. Another study as mentioned in the literature review observed a reduced dynamic range of the F0 pitch in depressed individuals (Kuny & Stassen, 1993). Although this cannot be directly comparable, due to the referenced paper focusing on the deviation from the F0 mean rather than the full range of fluctuations in pitch, a pattern can still be observed to state that depressed individuals are less expressive when voicing readback passages of speech. Various other factors that could influence PV should be considered in future studies. For example, there were two students in the sample population that were studying drama and performing arts. This could have resulted in these participants being more comfortable and naturally more expressive when reading back passages of speech.

Pause Length (PL)

PL was measured from VT2. This acoustic feature was one of the most referenced in the literature and related to both the detection of depression and the severity of depression. One study (Kuny & Stassen, 1993) observed an increased PL amongst those with depression. This was measured for both counting from 1-30, and also reading back a passage of text. A second study (Liu et al., 2017) found that as patients recover from depression, the variation in their PL reduces. This would infer healthy individuals have a lower variation in PL, and this variation increases with the severity of depression. Another two studies echoed these two findings when they investigated the change in acoustic features in response to varying levels of depression, with an increase in PL correlated to an increase in depression severity found (Mundt

et al., 2012), as well as an increase in the variability of PL correlated to an increase in depression severity observed (Mundt et al., 2007, Mundt et al., 2012).

These findings, however, did not align with the measurements observed in this study. Individuals with depression exhibited a considerably shorter mean PL and also a considerably shorter variation in PL when reading back pre-written text during VT2. These results contradict earlier voice studies published in the literature, as is of particular interest due to the measurements exhibiting a p-value of 0.03, to signify statistical significance. The contradiction in results when compared to previous studies could also be due to other variables affecting the voice. Four participants out of the sample population had a native language other than English. One of these participants were in the sample group, whilst the other three participants were in the control group. When looking at individual PL measurements, two out of the three participants with varying native languages in the control group had considerably higher PL measurements that measured over five times greater than some participants of whose native language was English. The mean PL for these participants measured 1160ms and 1100ms. The lowest mean PL measured amongst participants in the control group was 197ms respectively. This observation could have been down to many variables, for one this could be an anomaly, or it may also be related to the English-speaking level of an individual. A conclusion that non-native speakers exhibit longer PLs when reading back a passage of speech in English however, as no evidence was found to back up these findings in the literature. The number of non-native speakers recruited in this thesis would also not warrant a conclusion such as this to be drawn. Therefore, this would point to a future study that controls for native language, the level of English speaking, or focuses entirely on a sample population with one native language, to understand more on how these results were achieved.

7) **Limitations**

One of the major limitations that had an impact on the conclusions drawn from this study was the number of participants in the sample and control group. This reduced the power of the statistical analysis' carried out and meant that the statistical significance of the results was not strong enough to be confident that certain changes in acoustic features could be used in the detection of depression. When calculating the power analysis for this study, the context that this technology would be used in needed to be considered when determining whether the conclusions drawn would be significant. This thesis used a power value of equal to or less than 0.05 to signify results that were statistically significant. As a result, only two out of the six acoustic features proved to have statistical significance in specific VTs. Medical devices that are used in the detection of health conditions require a high accuracy to be sure that the diagnosis is valid. Whilst the percentage accuracy rating of a medical device in diagnosing a clinical condition may be hard to obtain, due to the wide variety of conditions and devices out there, the benefits should always outweigh the risks. An accuracy of close to 100% would ensure the risk of a misdiagnosis is considerably attenuated.

The systems of including or excluding participants into the control and sample group had certain limitations. Participants included in the sample group had been clinically diagnosed with depression by a doctor. However, there were no background checks done to ensure that this information provided by participants was truthful. Participants were trusted that this information was provided with honesty. Certain studies (Wang et al., 2019) collaborated with medical centers when recruiting participants with depression, which ensured medical information was accurate. Furthermore, in the sample group, one participant out of the nine was accepted into the sample group upon self-diagnosing themselves with depression. There were no further checks done with this participant to ensure that the depression they were experiencing was persistent over a period of time, and it impaired their work, personal life, and relationships. Healthcare questionnaires such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) could have been used here to ensure there was some evidence to state that this participant was suffering from depression and therefore should be permitted into the sample group. The DSM-5 is best used by a doctor however, meaning collaborating with medical professionals would have been needed to ensure the results were accurate and participants safety and comfort had been accounted for.

Another limitation was that the severity of depression that participants in the sample group was not considered. Depression can be classified as mild, moderate, or severe. Participants with severe depression could have displayed vastly different scores than participants with mild depression. Ideally, implementing sub-groups to enable a comparison between the severity of depression affecting acoustic vocal features would have sufficed to investigate whether this had any influence on measurements. This

was not done in this study due to the low number of participants recruited to partake. As well as the severity of depression, the type of depression was also not considered. "There are various types of depression, such as Major Depression, Persistent Depressive Disorder... Bipolar disorder and Psychotic Depression" (Williams). All of these can vary in terms of their severity and effects on a person. The type of depression the participants in the sample group suffered from was not recorded in this study. With potential differences in types of depression and the severity, participants in the sample group could have also been prescribed various drugs to treat their condition. This information was not gathered, since the investigation into the impact that these drugs have on the human voice would have required a separate study, due to no current literature on regularly prescribed depression drugs such as SSRIs, SNRIs, NDRIs or antidepressants published. A study also investigating depression detection noted "psychotropics could have confounding effects on the voice", mentioning that measuring the voices of individuals with depression who do not take any drugs for this condition could be effective in understanding the impact of these drugs on the human voice (Taguchi et al., 2018).

Depression is not always black and white. Our mental state can change on a day to day or even a moment-to-moment basis based on the situation we are in. Therefore, detecting depression may come with various complexities. The acoustic features in the human voice may exhibit changes only in a situational specific environment, or they could exhibit changes in a cross-situational environment. Situational specific refers to the acoustic features which vary according to the situation a person is in. A person may be in a relaxed, stressful, fearful, loud, or quiet environment for example. If the acoustic feature only manifests changes in certain situations, then it is referred to as a situational specific feature. If the acoustic feature in the voice manifests changes in all situations, then it is referred to as a cross-situational feature. A question which has arisen in another studies, "are the vocal differences in people with depression cross-situational, or can they be detected in special situations?" (Wang et al., 2019) relates to this. Whether the acoustic features analysed in this study were situational specific or cross-situational was not considered. Ideally it would have been suitable for the acoustic features chosen to be cross-situational, as there was no insight into what situation the participants were in when recording the vocal tasks, due to voice recordings being submitted remotely by the participant.

The participants in this study were all recruited from UoH and completing undergraduate or postgraduate studies. This subsequently meant there was no variety between those who had/were obtaining higher education, and those that did not pursue higher education. Ultimately this meant that the results were not representative of the wider population. Previous studies have controlled the impact of educational level on measurements by using it as a co-variate (Wang et al., 2019). In this study, whilst the focus was specifically on those in higher education, undergraduate and postgraduate levels could have been used as a co-variate to investigate whether there was any variability in acoustic feature measurements within the high education demographic itself. In the sample population of this study at least eight participants

were postgraduate students. Evidence shows that in particular, “graduate students are more than six times as likely to experience depression compared with the general population” (Gin et al., 2021), demonstrating the importance of controlling for undergraduate and postgraduate level.

Native language information was gathered from participants, with four out of the 17 participants in the sample population possessing a native language other than English (see appendix 1.1 and 1.2). Gathering this data allowed for clearer visibility if stand out measurements were observed from participants with differing native languages. Major differences between languages include the tonality. Languages such as Mandarin and Vietnamese are tonal, whilst languages such as English and Portuguese are non-tonal. In tonal languages, one syllable can have a variety of meanings based on the pitch it is voiced at (Liu et al., 2010). Whilst all participants were voicing English vowels and repeating a passage in English, this could still have led to greater perturbations in pitch amongst the three participants that had tonal native languages (Mandarin and Vietnamese). “Speakers of Mandarin Chinese seem to vary their fundamental frequency more rapidly during continuous speech than do speakers of American English” (Eady, 1982). This could have also interfered with PV measurements, as “pitch variation plays an important role in tone languages, as varying F0 patterns communicate different lexical meanings” (Wong et al., 2021), suggesting that tonal native language speakers could have a greater variation in pitch. American English has differences to the style of English spoken by the participants in this study (British English), however, we can infer that it would apply to British English also, due to both styles of English being classified as a non-tonal language. An ideal scenario would have been to use native language as a co-variate or build the study to analyse each native language separately to account for the differences in tonal and non-tonal languages and any other features which vary between languages. With the limited number of participants in the study, statistically significant results would not have been gained upon implementing native language as a co-variate, therefore it not used as a co-variate.

8) **Conclusions**

This study analysed the voices of 17 biological females in higher education between the ages of 20 and 26 years old. The sample population was comprised of eight healthy individuals and nine individuals with depression. The voices of healthy and depressed individuals were analysed against each other after participants completed two vocal tasks. Results confirm out of the six initial hypotheses set, all but H3 was true when using a statistical significance level of 0.05, with an increased variation in F2 observed amongst those with depression. This was observed specifically in VT1 when participant extended the vowel /a/. This aligns with previous research in the literature, which observed a positive correlation between depression severity and an increased F2 variability. PL also showed significant changes, with a decreased PL observed amongst healthy individuals in VT2. However, this opposed H6, with results contradicting earlier research in the field. Further studies that use native language as a co-variate should be carried out to investigate this relationship. We can conclude, however, that whilst the F2 variability of the voice shows promise to aid in the detect of depression amongst females in higher education between the ages of 20 and 26 years old when using a statistical significance level of 0.05, this acoustic feature did not exhibit statistical significance once the Bonferroni Correction was applied. Further studies with a specific focus on F2 variability and increased sample sizes would be needed before any significant conclusions can be drawn, and subsequently, before voice diagnosis for depression can be implemented into clinical practices.

(i) **References**

- Almaghrabi, S.A., Clark, S.R., & Baumert, M. (2023). 'Bio-acoustic features of depression: A Review', *Biomedical Signal Processing and Control*, 85, p. 105020. doi:10.1016/j.bspc.2023.105020.
- American Psychiatric Association (2023). 'Depressive disorders', *DSM-5-TR® Clinical Cases* [Preprint]. doi:10.1176/appi.books.9781615375295.jb04.
- Balter, M. (2015, January 13). Human language may have evolved to help our ancestors make tools [Review of Human language may have evolved to help our ancestors make tools]. *Science*; science.org. [URL]
- Baykaner, K., Huckvale, M., Whiteley, I., Ryumin, O., & Andreeva, S. (2015). The Prediction of Fatigue Using Speech as a Biosignal. *Statistical Language and Speech Processing*, 8–17. https://doi.org/10.1007/978-3-319-25789-1_2.
- Bigpi. (2020, September 26). The Human Voice and the Frequency Range | Big Pi. Bigpi.vc. [URL]
- Bourel, M. (2019, October 7). Something in Your Voice. *Proto Magazine*. [URL]
- Calić, G. et al. (2022) 'Acoustic features of voice in adults suffering from depression', *Psiholoska istrazivanja*, 25(2), pp. 183–203. doi:10.5937/psistra25-39224.
- Cannizzaro, M. et al. (2004) 'Voice acoustical measurement of the severity of major depression', *Brain and Cognition*, 56(1), pp. 30–35. doi:10.1016/j.bandc.2004.05.003.
- Coherent Market Insights. (2019, November). Vocal Biomarkers Market Size, Trends, Shares, Insights, Forecast - Coherent Market Insights. www.coherentmarketinsights.com. [URL]
- Cox, R. V., De Campos Neto, S. F., Lamblin, C., & Sherif, M. H. (2009). ITU-T coders for wideband, superwideband, and fullband speech communication [Series Editorial]. *IEEE Communications Magazine*, 47(10), 106–109. <https://doi.org/10.1109/MCOM.2009.5273816>.
- Cullum Attwell, R.M. (2022) Cost of living and depression in adults, Cost of living and depression in adults, Great Britain - Office for National Statistics. [URL] (Accessed: 25 July 2023).
- Cummins, N. et al. (2017) 'Enhancing speech-based depression detection through gender dependent vowel-level formant features', *Artificial Intelligence in Medicine*, pp. 209–214. doi:10.1007/978-3-319-59758-4_23.
- Davenport, M., & Hannahs, S. J. (1998). 5.2.1. In *Introducing phonetics and phonology* (p. 61). Arnold.
- Dyson, M. (2021, July 13). Medical staffing in the NHS in England report. The British Medical Association Is the Trade Union and Professional Body for Doctors in the UK. [URL]
- Eady, S.J. (1982) 'Differences in the F0 patterns of speech: Tone language versus stress language', *Language and Speech*, 25(1), pp. 29–42. doi:10.1177/002383098202500103.

Elizalde, B., Badlani, R., Shah, A., Kumar, A., & Raj, B. (2018). NELS - Never-Ending Learner of Sounds.

Evans, T.M. et al. (2018) 'Evidence for a mental health crisis in graduate education', *Nature Biotechnology*, 36(3), pp. 282–284. doi:10.1038/nbt.4089.

France, D.J. et al. (2000) 'Acoustical properties of speech as indicators of depression and suicidal risk', *IEEE Transactions on Biomedical Engineering*, 47(7), pp. 829–837. doi:10.1109/10.846676.

Future, M. R. (2021, February). Vocal Biomarkers Market Growth, Trends | Size Estimation, 2027. [Www.marketresearchfuture.com](http://www.marketresearchfuture.com). [URL]

Georgalas, V. L., [et al.]. (2023). The effects of caffeine on Voice: A Systematic Review. **Journal of Voice*, 37*(4). <https://doi.org/10.1016/j.jvoice.2021.02.025>

Gideon, J., Provost, E. M., & McInnis, M. (2016). Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). <https://doi.org/10.1109/icassp.2016.7472099>

Gin, L.E., et al. (2021). 'PhDepression: Examining how graduate research and teaching affect depression in life sciences Phd students', *CBE—Life Sciences Education*, 20(3). doi:10.1187/cbe.21-03-0077.

Graber, M. L. (2013). The incidence of diagnostic error in medicine. *BMJ Quality & Safety*, 22(Suppl 2), ii21–ii27. <https://doi.org/10.1136/bmjqs.2012-001615>

Gustafson-Capkova, S., & Megyesi, B. (2001). A comparative study of pauses in dialogues and read speech. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) [Preprint]. doi:10.21437/eurospeech.2001-181.

Hamid, H., Abid, Z., Amir, A., Rehman, T. U., Akram, W., & Mehboob, T. (2020). Current burden on healthcare systems in low- and middle-income countries: recommendations for emergency care of COVID-19. *Drugs & Therapy Perspectives*. <https://doi.org/10.1007/s40267-020-00766-2>

Hashim, N.W., et al. (2017). Evaluation of Voice Acoustics as predictors of clinical depression scores. *Journal of Voice*, 31(2). doi:10.1016/j.jvoice.2016.06.006.

Hidden Braces, I. (n.d.). Speech Exercise The Rainbow Passage. Retrieved January 16, 2022, from <https://www.surreybraces.co.uk/pdf/the-rainbow-passage.pdf>

Higuchi, M., Shinohara, S., Nakamura, M., Mitsuyoshi, S., Tokuno, S., Omiya, Y., Hagiwara, N., & Takano, T. (2017, November 1). An effect of noise on mental health indicator using voice. *IEEE Xplore*. <https://doi.org/10.1109/ICIIBMS.2017.8279690>

Higuchi, M., Shinohara, S., Nakamura, M., Omiya, Y., Hagiwara, N., Takano, T., Mitsuyoshi, S., & Tokuno, S. (2018). Study on Indicators for Depression in the Elderly Using Voice and Attribute Information. *Communications in Computer and Information Science*, 127–146. https://doi.org/10.1007/978-3-319-93644-4_7

Higuchi, M., Tokuno, S., Nakamura, M., Shinohara, S., Mitsuyoshi, S., Omiya, Y., Hagiwara, N., Takano, T., Toda, H., Saito, T., Terashi, H., & Mitoma, H. (2018). CLASSIFICATION OF BIPOLAR DISORDER, MAJOR DEPRESSIVE DISORDER, AND HEALTHY STATE USING VOICE. *Asian Journal of Pharmaceutical and*

Clinical Research, 11(15), 89. <https://doi.org/10.22159/ajpcr.2018.v11s3.30042>

Huckvale, M., & Beke, A. (2017). It Sounds Like You Have a Cold! Testing Voice Features for the Interspeech 2017 Computational Paralinguistics Cold Challenge. Interspeech 2017. <https://doi.org/10.21437/interspeech.2017-1261>

Icht, M., et al. (2020). The “morning voice”: The effect of 24 Hours of sleep deprivation on vocal parameters of young adults. *Journal of Voice*, 34(3). doi:10.1016/j.jvoice.2018.11.010.

Jacewicz, E., Fox, R. A., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2), 233–256. <https://doi.org/10.1017/s0954394509990093>

Kalyani, A., Tonni, S., & Jayakumar, T. (2021a). A critical review of Swara (voice) in ayurveda. *Journal of Ayurveda Medical Sciences*, 4(2), pp. 475–479. doi:10.5530/jams.2019.4.6.

Kesari, G. (2021, May 24). AI Can Now Detect Depression From Your Voice, And It's Twice As Accurate As Human Practitioners. *Forbes*. <https://www.forbes.com/sites/ganeskesari/2021/05/24/ai-can-now-detect-depression-from-just-your-voice/?sh=7029d344c8d9>

Kleinberger, R. (2017, November 28). Why don't I like the sound of my own voice? | Rébecca Kleinberger | TEDxBeaconStreet. www.youtube.com. <https://www.youtube.com/watch?v=L2XppxqR4P0>

Kuny, S., & Stassen, H. H. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. **Journal of Psychiatric Research*, 27*(3), 289–307. [https://doi.org/10.1016/0022-3956\(93\)90040-9](https://doi.org/10.1016/0022-3956(93)90040-9)

Kwo, D. L. (2021, June 16). Contributed: The Future Use Case of Voice Biomarkers. **MobiHealthNews**. <https://www.mobihealthnews.com/news/contributed-future-use-case-voice-biomarkers>

Lab, L. (2020, October). Speech Acoustics 8 - expression of gender. **YouTube**. <https://www.youtube.com/watch?v=TWRB443YrHI&list=PL6niCBwOhjHhQFfl88fQfdLgiD7QaShBo&index=9>

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. **IEEE Open Journal of Engineering in Medicine and Biology*, 1*, 1–1. <https://doi.org/10.1109/OJEMB.2020.3026928>

Laguarta, J., Hueto, F., Rajasekaran, P., Sarma, S., & Subirana, B. (2020). Longitudinal Speech Biomarkers for Automated Alzheimer's Detection. <https://doi.org/10.21203/rs.3.rs-56078/v1>

Liu, H., et al. (2010). Effect of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations. **The Journal of the Acoustical Society of America*, 128*(6), 3739–3746. <https://doi.org/10.1121/1.3500675>

Liu, Z., et al. (2017). Speech pause time: A potential biomarker for depression detection. **2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* [Preprint]*. <https://doi.org/10.1109/bibm.2017.8217971>

Lortie, C. L., et al. (2015). Effects of age on the amplitude, frequency and perceived quality of voice. *AGE, 37*(6). <https://doi.org/10.1007/s11357-015-9854-1>

Low, L.-S. A., et al. (2011). Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering, 58*(3), 574–586. <https://doi.org/10.1109/tbme.2010.2091640>

Manescu, E. A., et al. (2020). Depression in young women: Clinical features and social interference. *Romanian Journal of Psychiatry and Psychotherapy, 22*(1), 1–4. <https://doi.org/10.37897/rjpp.2020.1.1>

Matacotta, J. J. (2020). Mental Health Concerns of undergraduate and graduate students: Depression, anxiety, eating concerns, and substance misuse [Preprint]. <https://doi.org/10.31234/osf.io/fgb8q>

Morris, N., & Elston, R. (2011). A note on comparing the power of test statistics at low significance levels. *The American Statistician, 65*(3), 164–166. <https://doi.org/10.1198/tast.2011.10117>

Mundt, J. C., et al. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics, 20*(1), 50–64. <https://doi.org/10.1016/j.jneuroling.2006.04.001>

Mundt, J. C., et al. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry, 72*(7), 580–587. <https://doi.org/10.1016/j.biopsych.2012.03.015>

Nakamura, M., Shinohara, S., Omiya, Y., & Tokuno, S. (2015). Correlation between self-administered psychological test and emotion measured by voice analysis.

Office for National Statistics. (2021, May 5). Coronavirus and depression in adults, Great Britain - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusanddepressioninadultsgreatbritain/januarytomarch2021>

Omiya, Y., Hagiwara, N., Shinohara, S., Nakamura, M., Higuchi, M., Mitsuyoshi, S., Takayama, E., & Tokuno, S. (2018). The Influence of the Voice Acquisition Method to the Mental Health State Estimation Based on Vocal Analysis. *IFMBE Proceedings, 327–330*. https://doi.org/10.1007/978-981-10-9035-6_59

Omiya, Y., Takano, T., Uruguchi, T., Nakamura, M., Higuchi, M., Shinohara, S., Mitsuyoshi, S., So, M., & Tokuno, S. (2019). An Attempt to Estimate Depressive Status from Voice. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 168–175*. https://doi.org/10.1007/978-3-030-25872-6_13

Oxford, R. (n.d.). Knight Move Thinking. *Oxford Reference*. <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100040543>

Pan, W., et al. (2019). Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. *PLOS ONE, 14*(6). <https://doi.org/10.1371/journal.pone.0218172>

- Panjeh, S., Nordahl-Hansen, A., & Cogo-Moreira, H. (2023a). Establishing new cutoffs for Cohen's d: An application using known effect sizes from trials for improving sleep quality on composite mental health. **International Journal of Methods in Psychiatric Research** [Preprint].
<https://doi.org/10.1002/mpr.1969>
- Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. **Biological Psychology*, 87*(1), 93–98.
<https://doi.org/10.1016/j.biopsycho.2011.02.010>
- Rondón Bernard, J. E. (2018). Depression: A review of its definition. **MOJ Addiction Medicine & Therapy*, 5*(1). <https://doi.org/10.15406/mojamt.2018.05.00082>
- Scherer, K. R. (1986). Voice, stress, and emotion. In **Dynamics of Stress** (pp. 157–179).
https://doi.org/10.1007/978-1-4684-5122-1_9
- Services, A. W. (2019, March). Your Health Speaks: Vocal Biomarkers and the Potential for Direct Measures of Health from Voice. YouTube. <https://www.youtube.com/watch?v=4HG6oDBJXfA>
- Shinohara, S., Aurelian, D. N., & Tokuno, S. (2018). Evaluation of Suicidal Ideation based on the Pitch Detection Rate.
- Shinohara, S., Nakamura, M., Mitsuyoshi, S., Tokuno, S., Omiya, Y., & Hagiwara, N. (2016). Voice disability index using pitch rate. **2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)**. <https://doi.org/10.1109/iecbes.2016.7843511>
- Silva, W. J., et al. (2021). Voice acoustic parameters as predictors of depression. **Journal of Voice** [Preprint]. <https://doi.org/10.1016/j.jvoice.2021.06.018>
- Singh, R. (2018). Keynote Session: Rita Singh, “Profiling Humans from their Voice.” YouTube; Consult. <https://www.youtube.com/watch?v=RvHxjegc2A>
- Singh, R. (2019). **Profiling humans from their voice**. Springer.
- Singh, R. (n.d). Machine Learning for Signal Processing; MLSP.
<http://mlsp.cs.cmu.edu/people/rsingh/index.html>
- Singh, R., Keshet, J., & Hovy, E. (2016). Profiling Hoax Callers [Review of Profiling Hoax Callers]. Carnegie Mellon University. <http://mlsp.cs.cmu.edu/people/rsingh/docs/Profiling-hoax-Singh.HST2016.pdf>
- Solutions, C. P. (2020). Time Domain Analysis vs Frequency Domain Analysis: A Guide and Comparison. <https://resources.pcb.cadence.com/blog/2020-time-domain-analysis-vs-frequency-domain-analysis-a-guide-and-comparison>
- Taguchi, T., et al. (2018). Major depressive disorder discrimination using vocal acoustic features. **Journal of Affective Disorders*, 225*, 214–220. <https://doi.org/10.1016/j.jad.2017.08.038>

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology, 9*, 1112–1122. <https://doi.org/10.1016/j.protcy.2013.12.124>

Tokuno, S., Omiya, Y., Shinohara, S., & Mitsuyoshi, S. (2016). Psychological impact of Kumamoto earthquake by voice analysis using a smart phone application.

Vlasenko, B., et al. (2017). Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. *Interspeech 2017* [Preprint]. <https://doi.org/10.21437/interspeech.2017-887>

Wang, J., et al. (2019). Acoustic differences between healthy and depressed people: A cross-situation study. *BMC Psychiatry, 19*(1). <https://doi.org/10.1186/s12888-019-2300-7>

Wang, T., & Dong, J. (2017). What is “Zheng” in traditional Chinese medicine? *Journal of Traditional Chinese Medical Sciences, 4*(1), 14–15. <https://doi.org/10.1016/j.jtcms.2017.08.005>

Williams, D. F. (n.d.). *Depression: Dr Faadiel Williams, Claremont Psychiatrist Dr Faadiel Williams*. Retrieved October 29, 2023, from <https://www.claremontpsychiatrist.co.za/Depression>

Wong, E. C., et al. (2021). Pitch variation in children with childhood apraxia of speech: Preliminary findings. *American Journal of Speech-Language Pathology, 30*(3S), 1511–1524. https://doi.org/10.1044/2021_ajslp-20-00150

World Health Organization (2023). Depressive disorder (depression). <https://www.who.int/news-room/fact-sheets/detail/depression>

World Health Organization. (2010). *Telemedicine: opportunities and developments in Member States: report on the second global survey on eHealth*. EHealth Series.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing, 4*(2), 142–150. <https://doi.org/10.1109/t-affc.2012.38>

(ii) Appendix

Appendix 1.1: Table 1 (Participant Information): Control Group

Participant ID	8	18	23	31	35	36	37	38
Sex	Female	Female	Female	Female	Female	Female	Female	Female
Age	24	22	20	24	25	25	25	23
Ethnicity	White Caucasian	White Caucasian	White Caucasian	Black/ African	Asian	Asian	Black/ African	Black/ African
Native Language	English	English	English	English	Chinese	Vietnamese	English	Portuguese
Stimulants	No	No	No	Yes	No	Yes	No	No
Alcohol	Yes	No	Yes	Yes	Sometimes	No	No	Sometimes
Smoker	No	No	No	No	No	No	No	No
Country of Residence	England	England	England	England	England	Unknown	England	England
Type of Student	Masters	Undergraduate	Undergraduate	Undergraduate	PhD	Unknown	PhD	PhD
Degree	Drama	Animation Production	Chemistry	Pharmacy	Business & Management	Unknown	Psychology	Business & Management
Year of Study	Final	Third	Third (Placement)	Third	Second	Unknown	Second	First
Accommodation	Private Accommodation	Shared House	Shared House	Private Accommodation	Student Accommodation	Unknown	Private Accommodation	Private Accommodation
Living with Others/Alone	People I know	People I know	People I know	People I know	Alone	Unknown	People I know	People I know
Job Alongside Studying	Yes	Yes	Yes	Yes	Yes	Unknown	Yes	No
Financial Worries	Yes	Yes	Yes	Yes	Yes	Unknown	Yes	No
Average Sleep (per night-hours)	7	5-9	7.5	7	6-7	Unknown	6-7	7-9

Appendix 1.2: Table 1 (Participant Information): Control Group

Participant ID	3	34	10	14	2	39	40	41
Sex	Female	Female	Female	Female	Female	Female	Female	Female
Age	26	23	24	20	25	23	24	26
Ethnicity	White Caucasian	Black/ African	White Caucasian	White Caucasian	Asian	White Caucasian	Asian	White Caucasian
Native Language	English	English	English	English	Mandarin	English	English	English
Stimulants	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Alcohol	Yes	No	Yes	Yes	Yes	Yes	No	Sometimes
Smoker	No	No	No	No	No	Sometimes	No	No
Country of Residence	England	England	England	England	England	England	England	England
Type of Student	PhD	Undergraduate	Undergraduate	Undergraduate	PhD	Masters	Masters	PhD
Degree	English Literature	Molecular and Cellular Biology	Graphic Design & Animation	Drama	Unknown	Psychology	English Literature	Psychology
Year of Study	Final	Final	4 th Year	Final	Second	Second	Second	Third
Accommodation	Student Halls	Student Halls	Private Accommodation	Shared House	Student Halls	Private Accommodation	Shared House	Private Accommodation
Living with Others/Alone	People I do not know	People I do not know	People I know	People I know	Alone	People I know	People I know	People I know
Job Alongside Studying	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Financial Worries	Yes	Yes	Yes	Yes	No	Yes	Yes	Sometimes
Average Sleep (per night-hours)	7	8	7	7	6	8	7	2-7

Appendix 1.3: Raw Voice Data – F0 / Sample Group

Sample Group			Fundamental Frequencies (F0)			
No.	Mental Health Status	Sex	F0 – VT1 /a (Hz)	F0 – VT1 /u (Hz)	F0 – VT1 /m (Hz)	Mean F0 – VT2 (Hz)
3	Diagnosed Depression	Female	188.78	195.56	196.82	161.09
22	Diagnosed Depression	Female	234..73	245.57	238.13	228.22
34	Diagnosed Depression	Female	212.24	241.47	216.10	189.13
10	Diagnosed Depression	Female	199.24	205.86	208.82	194.7
14	Diagnosed Depression	Female	195.08	209.39	193.2	166.1
2	Self-Diagnosed Depression	Female	263.23	261.84	272.06	214.64
39	Diagnosed Depression	Female	232.51	233.88	237.77	174.34
40	Diagnosed Depression	Female	225.23	237.01	237.1	209.23
41	Diagnosed Depression	Female	204.67	198.98	198.98	178.99
		Sample Mean (M1)	217.30	225.95	222.11	190.72
		Standard Deviation	23.79	22.99	26.09	22.99
		Variance	565.75	528.69	680.59	528.67
		Count	9	9	9	9

Appendix 1.4: Raw Voice Data – F0 / Control Group

Control Group			Fundamental Frequencies (F0)			
No.	Mental Health Status	Sex	F0 – VT1 /a (Hz)	F0 – VT1 /u (Hz)	F0 – VT1 /m (Hz)	Mean F0 – VT2 (Hz)
8	Not Depressed	Female	207.34	205.86	207.23	201.7
18	Not Depressed	Female	173.91	162.34	188.75	145.33
23	Not Depressed	Female	232.68	233.05	208.54	206.49
31	Not Depressed	Female	193.63	198.82	200.99	187.76
35	Not Depressed	Female	219.79	189.55	215.06	171.22
36	Not Depressed	Female	229.31	222.46	196.98	168.59
37	Not Depressed	Female	190.34	221.45	227.84	153.45
38	Not Depressed	Female	179.35	191.84	185.56	162.48
		Control Mean (M2)	203.31	203.17	203.87	174.63
		Standard Deviation	22.45	22.7	13.9	22.11
		Variance	503.93	515.17	193.22	489.01
		Count	8	8	8	8

Appendix 1.5: Raw Voice Data – F1 / Sample Group

Sample Group			Formant	Frequency 1 (F1)	
No.	Mental Health Status	Sex	F1 – VT1 /a (Hz)	F1 – VT1 /u (Hz)	F1 – VT1 /m (Hz)
3	Diagnosed Depression	Female	616.64	397.01	370.08
22	Diagnosed Depression	Female	736.41	472.93	287.41
34	Diagnosed Depression	Female	745.78	414.8	319.21
10	Diagnosed Depression	Female	712.9	385.57	251.79
14	Diagnosed Depression	Female	686.61	391.68	376.48
2	Self-Diagnosed Depression	Female	1002.01	384.01	289.65
39	Diagnosed Depression	Female	577.28	463.6	277.4
40	Diagnosed Depression	Female	793.44	282.6	244.21
41	Diagnosed Depression	Female	657.86	400.07	398.41
		Sample Mean (M1)	725.21	399.14	312.74
		Standard Deviation	123.55	54.66	56.53
		Variance	15263.58	2987.62	3195.85
		Count	9	9	9

Appendix 1.6: Raw Voice Data – F1 / Control Group

Control Group			Formant	Frequency 1 (F1)	
No.	Mental Health Status	Sex	F1 – VT1 /a (Hz)	F1 – VT1 /u (Hz)	F1 – VT1 /m (Hz)
8	Not Depressed	Female	584.41	374.57	215.47
18	Not Depressed	Female	747.6	495.47	370.13
23	Not Depressed	Female	725.25	440.84	259.94
31	Not Depressed	Female	748.52	390.1	293.4
35	Not Depressed	Female	608.87	258.28	229.81
36	Not Depressed	Female	831.91	357.21	303.83
37	Not Depressed	Female	739.06	385.75	244.23
38	Not Depressed	Female	666.45	342.93	351.7
		Sample Mean (M1)	706.51	380.64	283.56
		Standard Deviation	81.62	69.6	56.41
		Variance	6662.27	4843.63	3182.17
		Count	8	8	8

Appendix 1.7: Raw Voice Data – F2 / Sample Group

Sample Group			Formant	Frequency 2	(F2)
No.	Mental Health Status	Sex	F2 – VT1 /a (Hz)	F2 – VT1 /u (Hz)	F2 – VT1 /m (Hz)
3	Diagnosed Depression	Female	1292.96	1924.83	1857.08
22	Diagnosed Depression	Female	1063.85	1764.52	1649.26
34	Diagnosed Depression	Female	1456.29	1347.61	1723.16
10	Diagnosed Depression	Female	904.42	1280.92	1667.33
14	Diagnosed Depression	Female	1284.62	1930.81	1943.31
2	Self-Diagnosed Depression	Female	1261.39	825.55	1630.08
39	Diagnosed Depression	Female	784.31	1183.7	623.18
40	Diagnosed Depression	Female	1086.69	788.41	1570.35
41	Diagnosed Depression	Female	942.34	1328.82	1425.23
		Sample Mean (M1)	1119.65	1375.02	1565.44
		Standard Deviation	219.4	426.8	384.36
		Variance	48135.49	182159.54	147729.72
		Count	9	9	9

Appendix 1.8: Raw Voice Data – F2 / Control Group

Control Group			Formant	Frequency 2	(F2)
No.	Mental Health Status	Sex	F2 – VT1 /a (Hz)	F2 – VT1 /u (Hz)	F2 – VT1 /m (Hz)
8	Not Depressed	Female	1232.08	975.75	1753.07
18	Not Depressed	Female	1469.12	1154.92	2023.4
23	Not Depressed	Female	1549.96	1154.62	961.57
31	Not Depressed	Female	1212.22	1193.15	1421.54
35	Not Depressed	Female	1399.45	996.75	1556.99
36	Not Depressed	Female	1185.64	802.58	1502.5
37	Not Depressed	Female	1535.69	1669.12	1538.37
38	Not Depressed	Female	1202.24	1082.48	1818.22
		Sample Mean (M1)	1348.3	1128.67	1571.96
		Standard Deviation	157.12	252.66	315.92
		Variance	24687.46	63837.42	99803.16
		Count	8	8	8

Appendix 1.9: Raw Voice Data – VOT / Sample Group

Sample Group			Voicing (VOT) Onset Time	
No.	Mental Health Status	Sex	VOT – VT2 /pot (ms)	VOT – VT2 /take (ms)
3	Diagnosed Depression	Female	58.17	34.64
22	Diagnosed Depression	Female	56.64	54.32
34	Diagnosed Depression	Female	51.37	34
10	Diagnosed Depression	Female	32.41	63.14
14	Diagnosed Depression	Female	64.04	64.04
2	Self-Diagnosed Depression	Female	55.36	122.02
39	Diagnosed Depression	Female	29.54	51.01
40	Diagnosed Depression	Female	66.98	41.01
41	Diagnosed Depression	Female	66.44	55.31
		Sample Mean (M1)	53.44	57.72
		Standard Deviation	13.78	26.55
		Variance	189.83	705.08
		Count	9	9

Appendix 2.0: Raw Voice Data – VOT / Control Group

Control Group			Voicing (VOT) Onset Time	
No.	Mental Health Status	Sex	VOT – VT2 /pot (ms)	VOT – VT2 /take (ms)
8	Not Depressed	Female	102.89	64.78
18	Not Depressed	Female	78.28	64.47
23	Not Depressed	Female	52.59	53.95
31	Not Depressed	Female	85.49	51.87
35	Not Depressed	Female	33.52	34.32
36	Not Depressed	Female	94.72	48.42
37	Not Depressed	Female	22.05	43.12
38	Not Depressed	Female	59.71	82.86
		Sample Mean (M1)	66.16	55.47
		Standard Deviation	29.09	15.04
		Variance	846.08	226.11
		Count	8	8

Appendix 2.1: Raw Voice Data – PV / Sample Group

Sample Group			Pitch Variability (PV)
No.	Mental Health Status	Sex	PV – VT2 (Hz ²)
3	Diagnosed Depression	Female	2694.96
22	Diagnosed Depression	Female	720.23
34	Diagnosed Depression	Female	3087.01
10	Diagnosed Depression	Female	3456.58
14	Diagnosed Depression	Female	3554.19
2	Self-Diagnosed Depression	Female	3588.08
39	Diagnosed Depression	Female	2066.58
40	Diagnosed Depression	Female	597.72
41	Diagnosed Depression	Female	1856.7
		Sample Mean (M1)	2402.45
		Standard Deviation	1167
		Variance	1361895.54
		Count	9

Appendix 2.2: Raw Voice Data – PV / Control Group

Control Group			Pitch Variability (PV)
No.	Mental Health Status	Sex	PV – VT2 (Hz ²)
8	Not Depressed	Female	2009.31
18	Not Depressed	Female	4298.8
23	Not Depressed	Female	1703.77
31	Not Depressed	Female	2911.72
35	Not Depressed	Female	2883.19
36	Not Depressed	Female	3755.03
37	Not Depressed	Female	4060.29
38	Not Depressed	Female	2332.6
		Sample Mean (M1)	2994.34
		Standard Deviation	964.32
		Variance	929920.28
		Count	8

Appendix 2.3: Raw Voice Data – PL / Sample Group

Sample Group			Pause Length
No.	Mental Health Status	Sex	PL – VT2 (ms)
3	Diagnosed Depression	Female	306
22	Diagnosed Depression	Female	332
34	Diagnosed Depression	Female	217
10	Diagnosed Depression	Female	409
14	Diagnosed Depression	Female	349
2	Self-Diagnosed Depression	Female	230
39	Diagnosed Depression	Female	530
40	Diagnosed Depression	Female	130
41	Diagnosed Depression	Female	310
		Sample Mean (M1)	312.56
		Standard Deviation	115.98
		Variance	13451.53
		Count	9

Appendix 2.4: Raw Voice Data – PL / Control Group

Control Group			Pause Length (PL)
No.	Mental Health Status	Sex	PL – VT2 (ms)
8	Not Depressed	Female	318
18	Not Depressed	Female	197
23	Not Depressed	Female	397
31	Not Depressed	Female	518
35	Not Depressed	Female	1160
36	Not Depressed	Female	1100
37	Not Depressed	Female	1330
38	Not Depressed	Female	390
		Sample Mean (M1)	676.25
		Standard Deviation	444.64
		Variance	197707.64
		Count	8

Appendix 2.5: Python Code for Measuring Mean Fundamental Frequency (F0)

```
import numpy as np
import librosa
import librosa.display
from scipy import signal

# Load the audio file
audio_file = "voice data path.wav"
audio, sr = librosa.load(audio_file)

# Apply spectral subtraction for noise reduction
n_fft = 2048
win_length = n_fft
hop_length = win_length // 2
stft = librosa.stft(audio, n_fft=n_fft, hop_length=hop_length, win_length=win_length)
magnitude = np.abs(stft)
phase = np.angle(stft)
mean_magnitude = np.mean(magnitude, axis=1, keepdims=True)
magnitude -= mean_magnitude
magnitude = np.maximum(magnitude, 0)
clean_stft = magnitude * np.exp(1j * phase)
clean_audio = librosa.istft(clean_stft, hop_length=hop_length, win_length=win_length)

# Compute the harmonic-percussive source separation
c, _ = librosa.effects.hpss(clean_audio)

# Compute the harmonic component using the harmonic-percussive source separation
c_harmonic = librosa.effects.harmonic(c, margin=8)

# Compute the fundamental frequency using the auto correlation method
f0, voiced_flag, _ = librosa.pyin(c_harmonic, fmin=librosa.note_to_hz('C2'), fmax=librosa.note_to_hz('C7'),
sr=sr)

# Compute the average fundamental frequency
avg_f0 = librosa.hz_to_midi(f0[voiced_flag]).mean()

# Print the result
print("Average fundamental frequency: {:.2f} Hz".format(librosa.midi_to_hz(avg_f0)))
```

Appendix 2.6: Python Code for Measuring Pitch Variability (PV)

```
import os

import numpy as np
import librosa

def pitch_variability(audio_file, sr=22050, hop_length=512, fmin=50, fmax=300):
    audio, _ = librosa.load(audio_file, sr=sr)
    audio_filtered = librosa.effects.preemphasis(audio, coef=0.97) # Apply high-pass filter
    pitches, magnitudes = librosa.piptrack(y=audio_filtered, sr=sr, hop_length=hop_length, fmin=fmin,
fmax=fmax)
    pitch_values = []
    for t in range(pitches.shape[1]):
        index = magnitudes[:, t].argmax()
        pitch = pitches[index, t]
        if pitch > 0:
            pitch_values.append(pitch)
    pitch_var = np.var(pitch_values)
    return pitch_var

def process_speech_samples(sample_directory, audio_files):
    pitch_var_list = []
    for filename in audio_files:
        audio_path = os.path.join(sample_directory, filename)
        if os.path.isfile(audio_path) and filename.endswith(".wav"):
            pitch_var = pitch_variability(audio_path)
            pitch_var_list.append(pitch_var)
    return pitch_var_list

samples_dir = "voice data path.wav"

read_speech_files = ["voice data path.wav"] # Replace with your read speech filenames

read_pitch_var = process_speech_samples(samples_dir, read_speech_files)

print("Read speech pitch variability:", np.mean(read_pitch_var))
```

Appendix 2.7: Python Code for Measuring Pause Length (PL)

```
import librosa
import numpy as np

def average_pause_length(audio_file_path, silence_threshold=-40, min_pause_length=0.1,
high_pass_cutoff=100):
    y, sr = librosa.load(audio_file_path, sr=None)

    # Apply high-pass filter
    y_filtered = librosa.effects.preemphasis(y, coef=high_pass_cutoff / (0.5 * sr))

    amplitude_threshold = librosa.db_to_amplitude(silence_threshold)

    is_silent = np.abs(y_filtered) < amplitude_threshold
    pause_starts = np.where(np.diff(is_silent.astype(int)) == 1)[0]
    pause_ends = np.where(np.diff(is_silent.astype(int)) == -1)[0]

    if len(pause_starts) == 0 or len(pause_ends) == 0:
        return 0

    if pause_starts[0] > pause_ends[0]:
        pause_starts = np.concatenate(([0], pause_starts))

    if pause_starts[-1] > pause_ends[-1]:
        pause_ends = np.concatenate((pause_ends, [len(y_filtered)]))

    pause_lengths = (pause_ends - pause_starts) / sr
    pause_lengths = pause_lengths[pause_lengths > min_pause_length]

    return np.mean(pause_lengths)

audio_file_path = "voice data path.wav"
avg_pause_length = average_pause_length(audio_file_path)
print("Average pause length:", avg_pause_length)
```

University of Huddersfield
School of Music Humanities and Media

**POSTGRADATE STUDENT / STAFF RESEARCH
ETHICAL REVIEW**

SECTION A: TO BE COMPLETED BY THE APPLICANT

Before completing this section please refer to the School Research Ethics web pages which can be found at <https://research.hud.ac.uk/strategy/concordat-research-integrity/mhmresearchgovernanceandethics> . Applicants should consult the appropriate ethical guidelines.

Please ensure that the statements in Section C are completed by the applicant (and supervisor for PGR students) prior to submission.

Project Title	Investigating how vocal parameters can be used in order to detect depression
Applicant	Joel Cooke
Supervisor (where applicable)	Professor Monty Adkins
Award (where applicable)	Masters of the Arts Research Degree - University of Huddersfield
Project start / end date	13 th February – 26 th April 2023

Mark 'X' in one or more of the following boxes if your research involves:

- ☒ direct contact with human/animal participants
- ☒ access to identifiable personal data for living individuals not already in the public domain
- ☐ increased danger of physical or psychological harm for researcher(s) or subject(s)
- ☒ research into potentially sensitive areas
- ☐ use of students as research assistants
- ☐ covert information gathering or deception
- ☐ children under 18 or subjects who may be unable to give fully informed consent
- ☐ prisoners or others in custodial care (e.g. young offenders)
- ☐ significantly increased danger of physical or psychological harm for researcher(s) or subject(s), either from the research process or from publication of research findings
- ☐ joint responsibility for the project with researchers external to the University.

Please note that if you provide sufficient information about the research (what you intend to do, how it will be carried out and how you intend to minimise any risks), this will help the

ethics reviewers to make an informed judgement quickly without having to ask for further details.

SECTION B: PROJECT OUTLINE (TO BE COMPLETED IN FULL BY THE APPLICANT)

Issue	Please provide sufficient detail to allow appropriate consideration of any ethical issues. Forms with insufficient detail will need to be resubmitted.
Aims and objectives of the study. Please state the aims and objectives of the study.	<p>The study's objective is to explore the effectiveness of analysing vocal parameters in diagnosing depression among young female adults. I aim to identify the most useful vocal parameter for this purpose and determine the most suitable vocal tasks for participants to perform to obtain accurate and reliable results. Our goal is to assess the accuracy of these findings and determine whether they warrant a larger-scale study involving a greater number of participants.</p>
Brief overview of research methodology The methodology only needs to be explained in sufficient detail to show the approach used (e.g. survey) and explain the research methods to be used during the study.	<p>Initially the participants will fill out a survey in which they will disclose information such as their age, gender and if they have/have not been diagnosed with depression. Other information will also be gathered such as their ethnicity, native language, and if they smoke or drink. This data will be gathered before participants are confirmed to participate in the study. I aim to recruit approximately 50% from Group A (sample group) and 50% from Group B (control group). Participants will be chosen based on if they fall within the required age range of 18-26, if they are female, and if there is an equal number of participants in the control and sample group. Participants will perform vocal tasks, from which quantitative observations will be made from their vocal parameters. These parameters include voicing onset time and fundamental frequency.</p>
Does your study require any permissions for study? If so, please give details	<p>Participants registering their interest in the study is permission in itself. The last question on the questionnaire confirms they accept to having their voice analyzed, as well as having their vocal recordings stored for 6 months after the master's research has been completed. This is to ensure we have access to this data if any evidence of the study needs to be shown to examiners during the grading phase. Once they have accepted this and if they meet the requirements of the demographic we are searching for, they will be accepted to participate in the study.</p>

<p>Participants Please outline who will participate in your research. Might any of the participants be considered 'vulnerable' (e.g. children)</p>	<p>I aim to recruit a minimum of 40 participants between the ages of 18 and 26 years old from the University of Huddersfield for this study. There will be 2 groups of participants that we will be using in this study, a control and sample group. Participants will be categorized in group A if they have previously been diagnosed with depression by a doctor, this will be the sample group. If participants have never been diagnosed with depression and if they currently do not believe they suffer from depression, they will be selected for the healthy group (Group B), which will be the control group.</p>
<p>Access to participants Please give details about how participants will be identified and contacted.</p>	<p>Participants will be recruited via the University of Huddersfield. An email will be sent out to all Undergraduate and Postgraduate students with information about the study and a link to fill out a webform to register their interest. Students will be recruited based on if they fall within the age bracket of 18-26 years old, and if they are female. There will be a certain number of students chosen based on if they have been diagnosed with depression. I aim for 50% of the participants selected being from Group A (having previously been diagnosed with depression), and 50% of the participants being selected from Group B (those who have never suffered from depression). Once they have been selected, a message/email will be sent to the participants with instructions on how to record the vocal tasks, and general information about the study.</p>
<p>How will your data be recorded and stored?</p>	<p>The recording of the vocal tasks will take participants no more than 5 minutes and will be sent to a secured number/email where the data will then be transferred and stored on an encrypted hard drive. If participants wish, the data will be deleted at the end of the data analysis stage. If patients are comfortable with the data being stored for a longer period, then it will be held until 6 months after the masters is completed, and then deleted.</p>
<p>Informed consent. Please outline how you will obtain informed consent. If informed consent or consent is <u>NOT</u> to be obtained please explain why.</p>	<p>Informed consent will be obtained in the survey sent out to participants. Only once participants have confirmed they want to take part will they be recruited for the study. The questionnaire in which informed consent is requested can be found here https://forms.office.com/r/uHsa37gcJy. The participants will be made aware that the study is to understand differences in the human voice. The main purpose to of the study (to detect depression) will not be confirmed in order to avoid any bias, in the case that participants may speak differently to how they usually would upon knowing the studies objectives.</p>

<p>Confidentiality Please outline the level of confidentiality you will offer respondents and how this will be respected. You should also outline about who will have access to the data and how it will be stored. (This information should be included on Information your information sheet.)</p>	<p>Full confidentiality will be offered to participants. Participants are given the option to write the name they wish to go by for the duration of the study. It is at the discretion of the participants whether they would like to give their real name or not. The names of participants and personal details such as mobile numbers and emails are not important in the study. If the participant provides these details, then they will be deleted immediately after the study, as only information related to their voice recordings, depression status, gender and age hold any importance in the study. Participants will be made aware that their names and contact details will be deleted from the database after the data has been gathered.</p> <p>When participants submit their vocal tasks, they have 3 options to do this. Either via email, WhatsApp, or Telegram. If submitted via email or WhatsApp, then their personal number and email will be visible to myself for the duration of the data collection stage. After this those details will be deleted from the database and device used to receive the voice recordings. The 3rd option to submit vocal tasks is via Telegram, which offers complete anonymity as Telegram allows you to hide your number and name. Having these 3 options allows participants to have a choice on which way suits them most for the submission of their vocal tasks. Websites to collect data were investigated, however these are all run by 3rd parties, meaning the data would not be secured. As a result, WhatsApp, Telegram and E-mail were chosen as the ways to submit the voice recordings.</p> <p>No personal data will be submitted via WhatsApp, Telegram, or E-mail other than their voice recordings for the purpose of the study. The only other information that will be gathered from participants will be that gathered from the questionnaire (https://forms.office.com/r/uHsa37gcJy).</p>
---	--

<p>Recorded Media</p> <p>Will the research involve the production of recorded media such as audio and/or video recordings? If so how will you ensure that there is a clear agreement with participants as to how these recorded media may be stored, used and (if appropriate) destroyed?</p>	<p>Before the study starts, participants will be asked whether they want to opt out of their voice data being stored on a secured hard drive after the study is completed. If they opt out, then their voice data will be deleted from the hard drive once the data analysis stage has completed. If they do not opt out, then their voice data is stored on a secured hard drive after the master's study completes for a period of 6 months. The data however will not be shared with any third parties at any point in the future. The question as to whether the participants want to opt in or out of this will be asked on a questionnaire before the study commences. All answers will be recorded, and if the participant chooses to opt out at a later date also, then their voice data will be immediately deleted.</p>
--	---

<p>Anonymity If you offer your participants anonymity, please indicate how this will be achieved.</p>	<p>Data will be anonymized by ensuring only basic details of the participants are stored, including age, gender, nationality, first language, previous mental health conditions (if any). Voice recordings will be gathered via either WhatsApp, Telegram, or email. If participants are not comfortable sending their voice recordings via their personal number, then email can be used. If participants are not comfortable with using their email, then they can provide their Telegram name where their number and personal details such as their real name can be hidden completely. The name is only used during the data collection stage. After this stage the data will be anonymized, leaving only basic details (excluding the participants name) to be stored for the data analysis stage. Participants names will not be stored in the database after the data collection stage is over. Each participant will be assigned a number, to avoid linking voice data to participants names.</p>
<p>Harm Please outline your assessment of the extent to which your research might induce psychological stress, anxiety, cause harm or negative consequences for the participants (beyond the risks encountered in normal life). If more than minimal risk, you should outline what support there will be for participants. If you believe that there is minimal likely harm, please articulate why you believe this to be so.</p>	<p>Even though the study is dealing with potentially vulnerable participants, there is no harm believed to be caused from the study. Participants are told exactly what to expect and what is required from them, they are consenting adults that are asked to perform vocal tasks for no more than 5 minutes.</p> <p>Participants with depression are not asked to relive or talk about any negative or traumatic experiences in their past. They are never asked to open wounds in which I would not be qualified to close. All vocal recordings are simple and contain clear instructions on how to record them. The study is non-intrusive and has been designed to ensure participants are at ease. Clear instructions are given on the participation information sheet, and all vocal tasks can be completed from the comfort of their own home.</p>
<p>Does the project include any security sensitive information? Please explain how processing of all security sensitive information will be in full compliance with the "Oversight of security - sensitive research material in UK universities: guidance (October 2012)" (Universities UK, recommended by the Association of Chief Police Officers)</p>	<p>This study does not include any security sensitive information. It is related to depression only. All information gathered is used for the betterment of the human society through investigating whether we can use voice as a diagnosis tool to assist doctors around the world with mental health conditions.</p>

Retrospective applications. If your application for ethics approval is retrospective, please explain why this has arisen.

This application for ethics approval is partly retrospective, as the recruitment of participants and the first round of data collection took place on 13th February 2023. My submission date for the masters is August 2023, meaning time is limited. Several participants were ready to submit their voice recordings on 13th February, and waiting for ethics approval before gathering this data would have meant missing out on vital data needed in to complete the masters on time. All processes followed so far have been in line with what has been stated above. Another round of recruitment of participants and collection of voice data is planned for the 26th April 2023.

SECTION C – SUMMARY OF ETHICAL ISSUES (TO BE COMPLETED BY THE APPLICANT)

Please give a summary of the ethical issues and any action that will be taken to address the issue(s).

Ethical issues for this study would include the storage of voice data, discrimination when selecting participants, the privacy of participants personal details and dealing with potentially vulnerable participants. It is important to store voice data safely to ensure no 3rd parties get access to the data. If participants allow me to hold any of their voice data for 6 months after the study is completed, then it is important I take on the responsibility of holding this, only using it for the purpose mentioned and ensuring it is stored on an encrypted hard drive with a password to access the data.

The next ethical issue would be being discriminative when selecting participants for the study. Participants are chosen based on their age, whether they are female, and if they have or have not been diagnosed with depression in the past. Participants must be aged between 18 and 26 and approximately half of the participants should have been diagnosed with depression. I am aiming to recruit females for this study, due to the vocal tract length being different in males and females, therefore effecting the fundamental frequency, which is one of the vocal parameters we are investigating in connection with depression.

These are the important aspects which I am looking for when selecting participants. Ethnicity and native language for example are not important in our analysis, therefore everyone who submits a request to join the study will get a fair chance to participate if they fall within the fall within Group A or B, are female, and if their age is between 18 and 26.

Another ethical issue would be the privacy of the participants personal details and dealing with potentially vulnerable participants. Privacy would be maintained by deleting all data after the data collection stage is completed. This includes email addresses, mobile phone numbers and participants names. Beyond the data collection phase, the voice data is only identified via basic information such as age, gender, and depression status. Dealing with potentially vulnerable participants suffering from depression comes with extra reasonability, however the study is not intrusive and should not cause any distress to participants, as they are not asked to talk about their problems or their past. Participants will only be recording basic vocal tasks which takes no more than 5 minutes and can be done from their own home.

SECTION D – ADDITIONAL DOCUMENTS CHECKLIST (TO BE COMPLETED BY THE APPLICANT)

Please supply copies of all relevant supporting documentation electronically. If this is not available electronically, please provide explanation and supply hard copy.


I have included the following documents

Information sheet	Yes <input checked="" type="checkbox"/>	Not applicable <input type="checkbox"/>
Consent form	Yes <input checked="" type="checkbox"/>	Not applicable <input type="checkbox"/>
Letters	Yes <input type="checkbox"/>	Not applicable <input checked="" type="checkbox"/>
Questionnaire	Yes <input checked="" type="checkbox"/>	Not applicable <input type="checkbox"/>
Interview schedule	Yes <input type="checkbox"/>	Not applicable <input checked="" type="checkbox"/>

SECTION E – STATEMENT BY APPLICANT

I confirm that the information I have given in this form on ethical issues is correct.
(Electronic confirmation is sufficient).

Applicant name: JOEL COOKE


Applicant Signature: 

Date: 28th March 2023

Affirmation by Supervisor (where applicable)

I can confirm that, to the best of my understanding, the information presented by the applicant is correct and appropriate to allow an informed judgement on whether further ethical approval is required

Supervisor name: PROF MONTY ADKINS

Supervisor Signature: 

Date: 31 March 2023

Affirmation of DDoGE Dr
Claire Barber
Date:
25/04/2023

Participant Information Sheet

Research Project Title:

Investigating the differences in the human voice

Name of Researcher: Joel Cooke

Contact Details of Researcher: joel.cooke@hud.ac.uk

You are being invited to take part in a research project. It is important for you to understand why this research is being done and what it will involve. Please take time to read the following information. Ask if there is anything that is not clear or if you would like more information. May I take this opportunity to thank you for taking time to read this.

1. What is the purpose of the project?

The research intends to discover differences in the human voice. In this study we will be analyzing the voice through observing various microfeatures which are present in all our voices, such as 'voicing onset time', 'jitter', 'shimmer', and the change in 'fundamental frequencies'. The study will be for a maximum of 5 minutes on Monday 13th February 2023 and can be completed by sending over your voice recordings via email/WhatsApp/Telegram. The study will form part of my master's thesis at the University of Huddersfield.

2. Why have I been chosen?

You have fallen in the age bracket of the participants we are recruiting to join the study.

3. Do I have to take part?

Participation on this study is entirely voluntary, so please do not feel obliged to take part. Refusal will involve no penalty whatsoever and you may withdraw from the study at any stage without giving an explanation. If you do not record and submit the recordings between the hours of 6-10pm on the 13th February, then submissions after this time will be void and not count towards the study.

4. What do I have to do?

The vocal tasks will be completed on Monday 13th February and will take no more than 5 minutes of your time to complete. The definition of a vocal task in this instance is recording a short snippet of your own voice. There will be a total of 3 vocal tasks to complete, of which the clear instructions for each vocal task are below. 2 reminders will be given on the day of the submission. The first reminder at 5pm (before the study commences at 6pm) and the second reminder at 8pm (during the study).

An example of each vocal task will be sent to you before the study. Vocal tasks can be recorded as a voice note via your phone, however, please ensure that the microphone is not too close or too far away from your mouth and ensure that you are recording the vocal tasks in a place where the background noise is low.

Vocal Task 1:

You will extend the sounds A, U and M separately.

Aaaaaaa (for 5-8 seconds)

Uuuuuu (for 5-8 seconds)

Mmmm (for 5-8 seconds)

You will extend these sounds for 5-8 seconds each. An example will be sent of this before the study to ensure you are comfortable with performing the vocal task. Please perform this vocal task 2 times for each sound. 2x for Aaa, 2x for Uuu and 2x for Mmm. You can leave a few seconds silence between each extended sound.

Vocal Task 2:

The second vocal task will involve recording your voice reading back the following passage. Please try to practice this passage a few times before recording to ensure you are comfortable.

“When sunlight strikes raindrops in the air, they act like a prism and form a rainbow.

The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look but no one ever finds it.

When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow

Vocal Task 3:

Speak about your day for 10-20 seconds. This can include anything you did, what you enjoyed, what you achieved, what you ate, even the weather! Please try to be as relaxed as possible, as it is not the content of your speech that is important in the study, rather the analysis of microfeatures within the voice.

5. Are there any disadvantages to taking part?

There should be no foreseeable disadvantages to your participation. If you are unhappy or have further questions at any stage in the process, please address your concerns initially to the researcher if this is appropriate. Alternatively, please contact Professor M. Adkins (m.adkins@hud.ac.uk) at the School of Music, Humanities and Media, University of Huddersfield.

6. Will all my details be kept confidential?

All information which is collected will be strictly confidential and anonymized before the data is presented in any work, in compliance with the Data Protection Act and ethical research guidelines and principles. If any names have been given, once the data has been collected your voice data will be anonymized, meaning we will only use the basic details such as age and gender to identify your data. Your name or contact information will no longer be stored in the database after we have finished collecting the voice data.

8. What happens to the data collected?

The voice data will be collected via your contact number or email. After the data collection stage is over your contact details will be deleted from the database and not shared with anyone. Any information which links the data back to you will also be deleted from the system. Voice data will be stored on a secured hard drive until the study is over, of which it will then be deleted. Data will be used purely for the purpose of analyzing changes in vocal characteristics and no data will be shared with any 3rd parties.

7. What will happen to the results of the research study?

The results of this research will be written up in the final thesis paper by August 2023. If you would like a copy, please contact me at joel.cooke@hud.ac.uk.

9. Where will the research be conducted?

The research will be conducted entirely remotely. You will submit all vocal tasks via WhatsApp, Email or Telegram (whichever mode of submission you have chosen in the questionnaire). There is no need to attend any meetings in person, everything will be online.

12. Who has reviewed and approved the study, and who can be contacted for further information?

The study has been reviewed and approved by my supervisor Professor Monty Adkins. If you have any questions or doubts, please contact me at joel.cooke@hud.ac.uk.