Frontiers in ICT

Research Article
5 December 2014

Objective Methods for Reliable Detection of Concealed Depression

Cynthia Solomon, Michel F. Valstar, Richard Morriss, and John Crowe

[1] *University of Nottingham, Faculty of Engineering, UK*
[2] *University of Nottingham, School of Computer Science, UK*
[3] *University of Nottingham, Institute of Mental Health, UK*

Correspondence*:
Michel Valstar
University of Nottingham, School of Computer Science, UK ,
michel.valstar@nottingham.ac.uk

ABSTRACT

Recent research has shown that it is possible to automatically detect clinical depression from audio-visual recordings. Before considering integration in a clinical pathway, a key question that must be asked is whether such systems can be easily fooled. This work explores the potential of acoustic features to detect clinical depression in adults both when acting normally and when asked to conceal their depression. Nine adults diagnosed with mild to moderate depression as per the Beck Depression Inventory (BDI-II) and Patient Health Questionnaire (PHQ-9) were asked a series of questions and to read a excerpt from a novel aloud under two different experimental conditions. In one, participants were asked to act naturally and in the other, to suppress anything that they felt would be indicative of their depression. Acoustic features were then extracted from this data and analysed using paired t-tests to determine any statistically significant differences between healthy and depressed participants. Most features that were found to be significantly different during normal behaviour remained so during concealed behaviour. In leave-one-subject-out automatic classification studies of the 9 depressed subjects and 8 matched healthy controls, an 88% classification accuracy and 89% sensitivity was achieved. Results remained relatively robust during concealed behaviour, with classifiers trained on only non-concealed data achieving 81% detection accuracy and 75% sensitivity when tested on concealed data. These results indicate there is good potential to build deception-proof automatic depression monitoring systems.

Keywords: Behaviomedics, Depression, Affective Computing, Social Signal Processing, Automatic Audio Analysis

1 INTRODUCTION

Mental health disorders have a devastating impact on an individual's health and happiness. Worldwide, it is estimated that four of the ten leading causes of disability for persons aged five and older are mental disorders [78]. Among developed nations, major depression is the leading cause of disability: according to European Union Green Papers dating from 2005 [26] and 2008 [27], mental health problems affect one in four citizens at some point during their lives. As opposed to many other illnesses, mental ill health often affects people of working age, causing significant losses and burdens to the economic system, as well as the social, educational, and justice systems. The economic burden of these illnesses exceeds $300 billion in the US alone [33]. Despite these facts, the societal and self-stigma surrounding mental health disorders

has remained pervasive, and the assessment, diagnosis, and management of these starkly contrasts with the numerous technological innovations in other fields of healthcare.

Objective methods are necessary to improve current diagnostic practice since clinical standards for diagnosis are subjective, inconsistent, and imprecise. To overcome this, researchers have started to focus on known physical cues (biomarkers) that correlate with depression, such as stress levels [68], head movements [2, 47], psychomotor symptoms [48], and facial expressions [80]. Recent advances in Affective Computing and Social Signal Processing promise to deliver some of these objective measurements.

Affective Computing is the science of creating emotionally aware technology, including automatically analysing affect and expressive behaviour [65]. By their very definition, mood disorders are directly related to affective state and therefore affective computing promises to be a good approach to depression analysis. Social Signal Processing addresses all verbal and non-verbal communicative signalling during social interactions, be they of an affective nature or not [82]. Depression has been shown to correlate with the breakdown of normal social interaction, resulting in observations such as dampened facial expressive responses, avoiding eye contact, and using short sentences with flat intonation.

Although the assessment of behaviour is a central component of mental health practice it is severely constrained by individual subjective observation and lack of any real-time naturalistic measurements. It is thus only logical that researchers in affective computing and social signal processing, which aim to quantify aspects of expressive behaviour such as facial muscle activations and speech rate, have started looking at ways in which their communities can help mental health practitioners. This is the fundamental promise of the newly defined research field of Behaviomedics [79], which aims to apply automatic analysis and synthesis of affective and social signals to aid objective diagnosis, monitoring, and treatment of medical conditions that alter one's affective and socially expressive behaviour.

For depression, recent challenges organised to measure severity of depression on a benchmark database have shown relatively impressive success in automatically assessing the severity of depression [80, 81]. The winner of the 2014 challenge, a team from the MIT Lincoln Lab [84], attained an average error of 6.31 on a severity of depression score ranging between 0 and 43, indicating that even the first approaches in this direction have significant predictive value.

However, previous research has also indicated that identifying reliable indicators of depression is non-trivial. Symptoms of depression can vary greatly both within and between individuals. Moreover, people naturally modify their behaviour to adapt to their social environment. This may involve hiding the true extent of someone's feelings. While altering the social presentation of emotion may be a part of everyday life, this can be especially problematic for people with depression, particularly since people are often hesitant to ask for help given the societal stigma of mental illness, which further decreases the probability of accurate diagnosis. With the promise of Behaviomedical tools to automatically screen for or even diagnose depression, a serious question that needs to be addressed is: how easy is it to fool such automatic systems?

We conducted an experiment where participants were asked to perform two tasks: read a section of a popular book, and answer a question regarding their current emotional state. This experiment was repeated by participants who were known to suffer from major depressive disorder. After the first time, participants were given a brief explanation of how an automated depression analysis system might detect depression from their voice, and participants were asked to modify their behaviour so to avoid being detected as depressed. However, it turns out that while the participants did try to conceal their depression, this was not successful and our automatic depression recognition system performed almost as well as on the non-concealed data.

The research we report on in this work contains two major contributions: firstly we show that with as little as two audio features and a simple Naive Bayes classifier we can accurately discriminate between depressed and non-depressed people with an accuracy of 82.35%. We also explore more generally which auditory features differ significantly between healthy and depressed individuals. Secondly, and perhaps

more saliently, we show how these differences are impacted by an individual's attempt to conceal their depression, and reveal for the first time experimental evidence that it may not be possible for people to conceal the cues of depression in their voice.

2 DEPRESSION

Depression is the most prevalent mental health disorder and is estimated to affect one in ten adults. Traditionally, scientific and clinical approaches classify depression based on observable changes in patient affect that are not expected reactions to loss or trauma. Although there is a wide range in both the symptoms and severity of depression, it is generally agreed upon as per the Diagnostic and Statistical Manual 4th ed. (DSM-IV)[1] that to be diagnosed with major depressive disorder, a patient must exhibit five or more of the following symptoms [3]:

1. Depressed mood most of the day or nearly every day
2. Markedly diminished interest or pleasure in all or almost all activities most of the day or nearly every day
3. Significant unintentional weight loss or gain or increase/decrease in appetite
4. Insomnia or hypersomnia nearly every day
5. Noticeable psychomotor agitation or retardation nearly every day
6. Fatigue or loss of energy nearly every day
7. Feelings of worthlessness or either excessive or inappropriate guilt nearly every day
8. Diminished ability to think, concentrate, or make decisions nearly every day
9. Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt and specific plan

However, depression often has a much more significant impact than just these enumerated symptoms and can affect or be affected by a variety of biological, environmental, social, or cognitive factors. Depression itself cannot be understood without taking into account the social context and environment, as nationality, ethnicity, and socio-economic status all influence the prevalence and presentation of depression [37]. There is also significant variation between individual experiences of depression [49].

This symptom-based definition of depression makes accurate diagnosis problematic, as it is difficult to objectively measure psychological rather than physiological phenomena. Although diagnostic criteria are to some extent arbitrary, the classification itself can have a significant impact upon the recommended treatment. Additionally, depression cannot always be categorically distinguished from other mental health disorders. Depression and anxiety in particular often co-exist and exhibit similar effects on patient affect [11]. Diagnosis requires experienced clinicians and an understanding of an individual's history, psychological testing records, self-reporting, and assessment during clinical interviews [87]. This is often a lengthy procedure, and relevant data or experts may not always be accessible.

2.1 SELF-ASSESSMENT OF DEPRESSION

Due to the difficulty of consistent, efficient, and accurate diagnosis, self-assessments are often used as a quick way to diagnose and monitor depression. It should be noted that whilst these methods are inherently flawed by their very nature in requiring a patient to critically and honestly assess their own behaviour, they nonetheless serve as a reasonable quantifiable standard to be measured against. The two most commonly

[1] We adhere here to the widely accepted DSM-IV rather than DSM-V, which has been met with severe criticism to the point where the National Institute of Mental Health has decided not to adopt it.

used assessments are the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9). The BDI test was created in 1961 and has been updated several times since then. The most recent version was created in 1996 and modified for better adherence to the DSM-IV criteria [5, 4]. Conversely, the PHQ-9 was created in the mid-1990s as an improvement to the lengthier Primary Care Evaluation of Mental Disorders (PRIME-MD) and expressly scores the DSM-IV criteria through self-report [70].

A comparison between these two assessments can be found in Table **??** [43, 42, 44]. Numerous studies have investigated the relationship between the two tests for a range of patients with different mood disorders, backgrounds, and conditions, and have reported correlations ranging from 0.67 - 0.87 [17, 44, 14, 15, 28]. These tools have also been shown to correlate highly with clinician-rated depression measurements, such as the 17-item Hamilton Rating Scale for Depression (HRSD-17) [44].

Although a variety of other depression diagnostic tests exist, these two were chosen for our research, for reasons of availability and comparibility. The BDI-II is used as the gold standard for measuring depression severity in the recent Audio/Visual Emotion Challenges (AVEC 2013/2014, [80, 81]). On the other hand, the PHQ-9 is a simple, efficient, and free test that is often used interchangeably with the BDI-II. Because the PHQ-9 only takes about a minute to complete, it was deemed advantageous to add as a check for reliability and to allow future researchers to freely compare their results against ours.

2.2 EMOTION REGULATION AND DECEPTION

In everything from normal social interactions to police investigations, people are constantly trying to discern the veracity of other's behaviour. Consequently, scientists have tried to ascertain behavioural cues that could indicate deception [13]. However, these cues are not necessarily indicative of everyday attempts to suppress or regulate emotions and their expression. Over one's lifetime, people learn which emotions they should feel and express in a given social context [25, 55]. Regulating emotion is necessary for social functioning, although the extent of regulation required varies between cultures [54, 21].

These implicit rules that define what is socially acceptable not only influence what people feel, but also how their feelings are perceived both personally and by others [40]. Culture can thus contribute to the pressure to deny or understate one's feelings to be more "socially acceptable," making diagnosis difficult. Conversely, somatic complaints have no stigma attached to them and are therefore more readily presented. A study by Kirmayer et. al in 1993 demonstrated how commonly this phenomenon occurs with depression and anxiety. A majority of patients presented with exclusively somatic symptoms to their primary care physicians and only acknowledged a psychological aspect when prompted for further information. However, psychological complaints are more readily recognised and accurately diagnosed by primary care physicians. Moreover, these results were replicated in several countries, demonstrating that this suppression is persistent across numerous ethno-cultural groups [40, 41, 18, 69].

3 SPEECH-BASED AUTOMATIC DEPRESSION DETECTION

In previous work on automatic depression recognition both the audio and video modalities have been used (e.g. [85] for audio, and [19] for video). Of the two, the audio modality has so far been the most successful, with an audio-based approach by MIT-Lincoln Lab winning the AVEC 2013 depression recognition challenge [81, 85]. While it is expected that ultimately a combination of audio and video modalities will gain the highest possible recognition rates, for simplicity we focus on audio features only in this study of concealment of depression.

[1] As measured in Cameron's study for a relevant population in these experiments [8].

Measure	BDI-II	PHQ-9
Depression Criteria	DSM-IV	DSM-IV
Last Updated	1996	1999
Questions	21	9
Time Requirement	3 – 5 minutes	$<$ 1 minute
Rating Scale	Intensity (0 – 3)	Frequency (0 – 3)
Cost	£6/copy	Free
Time Scale	Over the last 2 weeks	Over the last 2 weeks
Designed to Measure	Severity	Severity
Ability to Measure Symptom Directionality	Yes	No
Correlation with HRSD-17 [1]	0.74	0.78
Score Range	0 – 13: Minimal 14 – 19: Mild 20 – 28: Moderate 29 – 63: Severe	0 – 4: No depression 5 – 9: Mild 10 – 14: Moderate 15 – 19: Moderately severe 20 – 27: Severe

3.1 THEORY OF SPEECH PRODUCTION

When first trying to understand how speech is produced, it is helpful to view the human vocal apparatus as a source and filter. In voiced speech, the source is the pressure wave created from the interaction of air pushed through the larynx and the vibrating vocal cords, whereas in unvoiced speech, by definition air does not interact with the vocal cords. The vocal cords can be adjusted by controlling muscles in the larynx, although the specific geometry of the cords is speaker-dependent. When speaking, air from the lungs flows quickly enough that the cords self-oscillate, which varies the size of the glottal opening and in turn, the amount of air allowed through. The resulting glottal volume velocity ultimately defines the periodicity of speech, or fundamental frequency [73]. However, this frequency changes naturally when speaking through further modifications of the jaw, lips, tongue, etc. It is important to note that the actual sound produced by the larynx during phonation is not created by the vibrations themselves, but rather, by

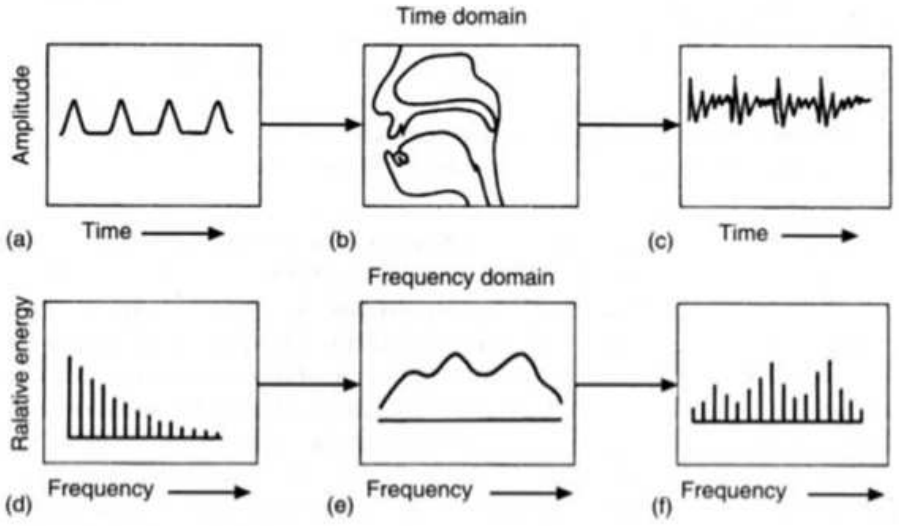


Figure 1. The source-filter theory of speech production can be traced to the glottal volume velocity, which produces (a) a glottal wave. This is then altered by (b) the shape of the vocal tract before (c) an emitted sound wave is audible. The (d) equivalent glottal spectrum can also be viewed as being affected by (e) the vocal tract transfer function to create (f) the resulting acoustic spectrum at the mouth opening. Figure taken from [10].

the modulated stream of air moving through the vibrating folds (see Fig. 1). Understanding how speech is produced is pertinent when one considers that depression affects different aspects of motor control, which is thus reflected in articulatory changes.

3.2 RELATED WORK

Qualitative observations of depression have been well documented, such as slower body movements, cognitive processing, and speech production, with depressed speech often described as "dead" and "listless" [57, 62, 59]. Speech content itself changes, which can be quantified by the amount of personal references, negators, direct references, or expressions of feelings in conversation [29, 83].

In contrast, parameters derived from the recorded speech signal rather than its subject matter content have only been explored within the last 40 years, with studies often producing conflicting results. For example, some studies investigating spectral energy distributions have concluded that depression is associated with increased energy at lower frequencies [64, 77] while others have found the opposite [16].

Other speech parameters have offered more consistent results. Fundamental frequency (F_0) is one of the most widely studied parameters and has demonstrated moderate predictability of depression severity, with decreasing amplitude and variability generally indicative of higher severity [86, 61, 52, 12]. Low et. al evaluated the resultant classification accuracy of mel frequency cepstral coefficients (MFCCs), energy, zero-crossing rate (ZCR), and Teager energy operators (TEO), and found a combination of MFCCs and other features most effective [52, 53, 51]. Formant patterns have also been shown to reflect reduced articulatory precision due to depression [23, 45].

Several promising prosodic features are measures of known clinical observations. Most commonly, depressed speech is reported as quantifiably quieter, less inflected, and slower, with fewer words uttered and a lower word rate [83]. Reliable measures of this include, but are not limited to, average speech duration, total speaking time, pause duration both within speech segments and before responses, variability in pause duration, average voice level (loudness), variance of voice level across all peaks (emphasis), and variance in pitch (inflection) [61, 20, 72, 9]. Changes in these parameters can reflect

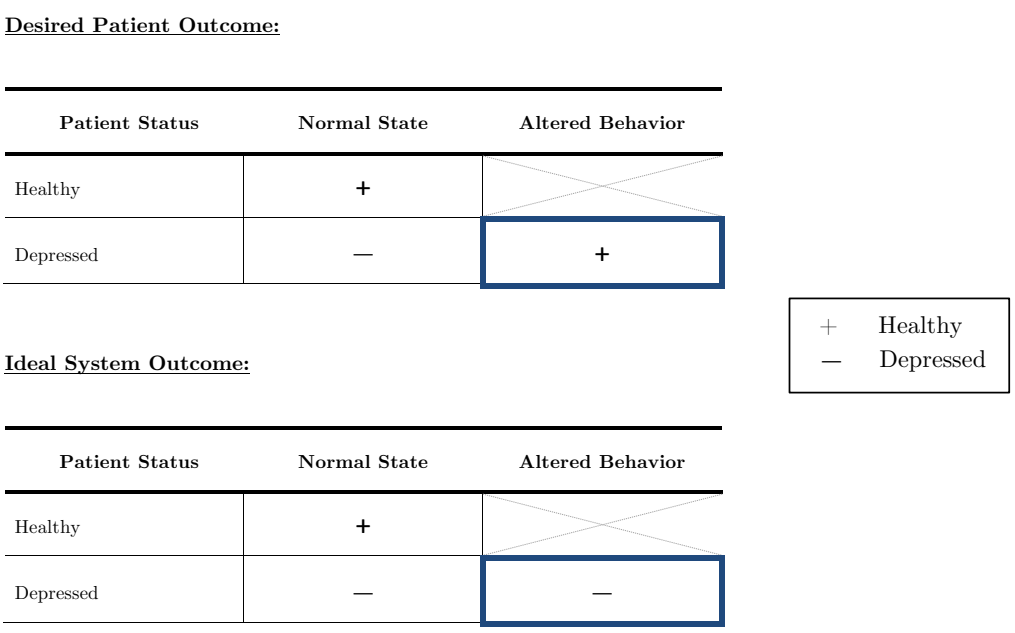


Figure 2. The depressed population can be subdivided into those who act as they feel, and those who try to outwardly suppress their symptoms. Although it is imperative to focus on both in diagnostic systems, truly robust systems should be able to accurately diagnose an individual with depression no matter the circumstance.

temporal changes in depression severity due to treatment [60]. These results have been replicated with non-English speakers [35].

It is important to note that experimental results generally differ if the depressed patient shows "agitated" or "retarded" symptoms. Retarded depression exhibits similar characteristics to sadness, whereas agitated depression involves a level of fear or restlessness. There is also a noticeable difference if data originates from automatic speech (counting or reading) or free speech [1]. The cognitive demand of free speech generally emphasizes speech abnormalities, particularly in pause time, moreover, different regions of the brain are activated during automatic and free speech [71, 31]. Consequently, both types were used in these experiments.

4 METHODOLOGY

This work aims to not only determine audio features that differ between healthy and depressed people, but also to investigate how they change when people with depression try to conceal their true emotions. The population of interest is outlined in blue in Fig. 2 below. Based on a set of optimised features, our goal was automatic depression recognition which will still be able to correctly classify a person as depressed even if they are trying to hide their depression. Healthy individuals who alter their behaviour to appear depressed were not of interest in this study.

4.1 DATA ACQUISITION

Participants were recruited primarily from postgraduate students at the University of Nottingham, as they were the most accessible. Advertisements were posted on social media websites, Call for Participants, and sent to a variety of different list serves. Participants self-identified as "depressed" or "healthy," but

these classifications were confirmed via PHQ-9 and BDI-II self-questionnaires. The purpose of the study was explained in full to both depressed and control participants. However, in order not to influence the participants' behaviour, the explanation did not include exactly what audio cues we were investigating as "objective measures" (e.g. vocal prosody, volume, etc.).

Ethics approval was obtained through the Ethics review board of the School of Computer Science at the University of Nottingham. The submission contained a consent form, information sheet, and a detailed checklist that described the experimental protocol and appropriate safeguard methods. 13 females (mean age 24.5 $\pm$ 3.1) and 4 males (mean age 25.5 $\pm$ 4.5) were recruited for this study. Of these, approximately half of both genders were classified as "healthy controls" and the other half classified as "depressed" by an initial PHQ-9 assessment and confirmed by the BDI-II. Following the questionnaire results, one participant's classification was altered, resulting in a distribution of nine depressed and eight healthy individuals. Because this participant was either trying to conceal her depression already or simply considered herself healthy, she did not complete the same concealment task. She is referred to as the "reclassified participant" in all further discussion.

Inclusion criteria required participants to be over 18, willing to provide informed consent, and, for depressed participants, meeting the DSM-IV criteria for mild to moderate depression by either a PHQ-9 score between 5 and 15 or BDI-II score between 14 and 28. Participants were excluded if they had a pre-existing psychotic mental health disorder (bipolar disorder, depression with delusions and hallucinations or paranoid ideation or schizophrenia or delusional disorder according to their own account), a high score on items regarding suicidal thoughts on either diagnostic test, or depression scores above the range listed previously, which would indicate moderately severe depression or higher. Medication was noted, but did not render a participant ineligible. In addition to the 17 that met the appropriate criteria, three people were deemed ineligible due to the severity of their depression.

Ideally participants would have been ethnically and culturally uniform so as to eliminate any effects on emotion regulation, speech content, or facial expressions unrelated to depression. However, given that the available population was almost exclusively limited to postgraduate university students, of which over half are international, this was unrealistic. As a second-best alternative to a uniform population, healthy participants were recruited to match the age, gender, cultural background, and native language of each of the depressed participants. Age matching was done to the closest possible age. Some matches were exact, but for others we applied a minimum mean square error approach. Cultural background was defined by whether a participant's first language was English or not (in practice this meant international students vs non-international students). In some cases exact nationality matches could be made, but this was not possible for every participant. Smoking habits were also noted due to the damage that smoking causes to the vocal cords and larynx [30]. The knowledge of who was who's matched partner was not used in any of the statistical and machine learning analyses performed in this work.

The collection protocol of these experiments is illustrated in Figure 3. The same acquisition system, location, and interviewer were used throughout experiments to ensure a consistent, controlled environment. Eligible participants were asked two separate sets of questions. Depressed participants were instructed to act naturally when answering one set and to alter their behaviour so as to hide any physical indicators of depression in the other. Healthy participants were instructed to act as naturally as possible throughout the course of the experiments. After each set of questions, participants read a one-page excerpt from the third "Harry Potter" book aloud so as to have a standard phonetic content for comparison. The order of the question sets was randomised.

Over the course of the experiments, participants were recorded with a webcam and microphone. Audio and video data were recorded using a Logitech C910 HD Pro Webcam and a Blue Snowball Microphone. Video was collected solely for potential use in future research in creating multimodal systems. Speech samples were digitised with a sampling rate of 44.1 kHz and 16-bits. A new AVI file was created for each set of questions and excerpt readings. The audio was then extracted into a WAV file for further processing and to ensure compatibility with a variety of software packages.

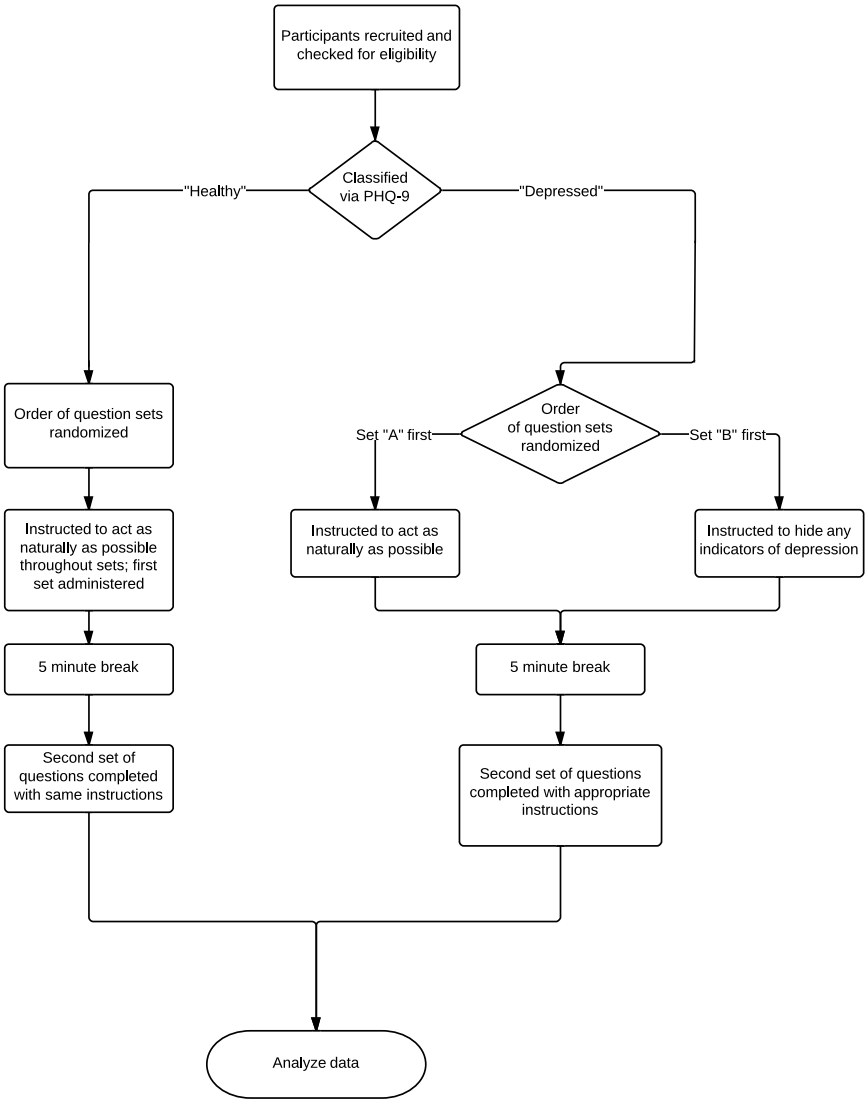


Figure 3. Overview of experimental protocol and testing process. Eligible participants were asked two separate question sets and were asked to read an excerpt from a popular book aloud. Depressed participants were instructed to act naturally when answering one set and to alter their behaviour so as to hide any physical indicators of depression in the other. Healthy participants were instructed to act as naturally as possible throughout the course of the experiments. The order of the question sets was randomised.

In order to provide standard conversation topics during the experiment, the interviewer asked a series of pre-set questions, taking care not to react to any of the subject's responses. The script was designed to maximise the amount of depressive cues collected in a short period of time. Participants were given time before the interview to familiarise themselves with their surroundings and ask any final clarification questions. In both scripts, initial questions were simple and positive in tone to establish rapport between interviewer and interviewee, and were followed by two reflective, potentially negative questions. When acting naturally, depressed participants were asked what they felt were physical indicators of depression, as well as how they have tried to conceal their depression previously, the rationale being that it is easier to be honest when not consciously trying to deceive.

However, when participants were asked to conceal their depression, they were asked about a deeply emotional topic, namely to describe their experiences with depression. In turn, this simulated the corresponding difficulty of hiding strong emotions in everyday life. The last questions were more open-ended, which allowed the participants to choose the topic and thus feel more in control of the situation. This was recommended as good practice for experiments involving social psychology [66, 24].

Control subjects were given similar questions – the only exceptions being that they were asked to describe a time when they had felt the need to conceal their emotions and how they thought it would feel to have depression. In addition to these questions, participants were also instructed to read aloud the first three paragraphs from the third "Harry Potter" novel [67][2], as it is both readable and relatable across a range of cultures, and also allows for more direct comparison of speech characteristics across groups and question sets.

4.2 FEATURE EXTRACTION

Due to any potential effects of equipment or environment on later analysis, signals were pre-processed before feature extraction by first manually removing voices other than the participant's (i.e. the interviewer's) and parsing the resultant signal into a new file for every question. Speech was then enhanced through spectral subtraction. The signal was split into frames of data approximately 25 ms long, as it was assumed that speech properties were stationary within this period, and a Hamming window was applied to each frame to remove signal discontinuity at the ends of each block. Each frame was normalised and a power spectrum was extracted to estimate the noise using a minimum mean-square error (MMSE) estimator. The noise spectra were averaged over several frames of "silence," or segments when only noise was present, and an estimate of the noise was then subtracted from the signal but prevented from going below a minimum threshold. In turn, this helped prevent over dampening of spectral peaks. Furthermore, because this threshold was set as a SNR, it could also vary between frames. This was implemented as a modified version of the spectral subtraction function in the MATLAB toolbox VOICEBOX [7].

Next, this enhanced signal was passed through a first-order high pass FIR filter for pre-emphasis. This filter was defined as:

$$H(z) = 1 + \alpha z^{-1} \tag{1}$$

where α was set as -0.95, which presumes that 95% of any sample originated from the prior one. Pre-emphasis serves to spectrally flatten the signal to amplify higher frequency components and offset the naturally negative spectral slope of voiced speech [38]. As human hearing is more sensitive above 1000 Hz, any further analysis is then also made more sensitive to perceptually significant aspects of speech that would otherwise be obscured by lower frequencies.

The selection of features significantly influences the accuracy of machine learning classifiers. As described in [50, 58, 56], acoustic features are often split into categories and subcategories to determine optimal feature sets. In this study, similar groupings are used and split into prosodic, spectral, cepstral, and TEO-based. A statistical analysis is then used to whittle down the number of features to only include those that are statistically significant.

4.2.1 Prosodic Features: Pitch and Fundamental Frequency Pitch is commonly quantified by and considered equivalent to fundamental frequency (F_0). F_0 is a basic and readily measurable property of periodic signals that is highly correlated with perceived pitch. F_0 approximates the periodic rate of glottis opening and closing in voice speech [56]. However, it is difficult to measure, as it changes over time and depends on the voicing state, which is often unclear. In these experiments, a slightly modified version of Talkin's pitch tracking algorithm in the MATLAB toolbox VOICEBOX was implemented. This algorithm is known for its relative robustness [73].

[2] The first two novels had introductions that expressed negative views of abnormality, which could potentially have been upsetting for participants.

4.2.2 Prosodic Features: Log Energy The logarithm of short-term energy is representative of signal loudness and is calculated on a per frame basis via Equation 2 below [50].

$$E_s(m) = \log \sum_{n=m-N+1}^{m} s^2(n) \tag{2}$$

where m is the frame number with N samples per frame, and $s(n)$ is the speech signal. Stress or emotion often affect measured energy.

4.2.3 Prosodic Features: Timing Measures Although speech is often segmented before analysis, prosodic analysis of the segment as a whole can also be useful. An automated script was written in the software package Praat [6] to extract various timing measures, calculated via Equations 3.3–3.6. The total number of syllables in the excerpt reading was considered constant for all participants, as the content was unchanged.

These features quantified some symptoms of psychomotor retardation in depressed patients, such as difficulty in thinking, concentrating, and choosing words. It was determined that performing these tests on spontaneous speech would be an inaccurate assessment of prosody due to the extent to which some participants in both groups connected or did not connect to the question. For example, some participants responded in single sentences, which did not offer enough data for fair comparison.

$$\text{Speech Rate} = \frac{\text{Number of Syllables}}{\text{Total Time}} \tag{3}$$

$$\text{Phonation Time} = \text{Duration of Voiced Speech} \tag{4}$$

$$\text{Articulation Rate} = \frac{\text{Number of Syllables}}{\text{Phonation Time}} \tag{5}$$

$$\text{Avg. Syllable Duration} = \frac{\text{Phonation Time}}{\text{Number of Syllables}} \tag{6}$$

4.2.4 Spectral Features: Spectral Centroid The spectral centroid is derived from the weighted mean of frequencies present in a signal and represents the center of the power distribution. It is calculated by Equation 7 below:

$$SC = \frac{\sum_{n=0}^{N-1} f(n)\, S(n)}{\sum_{n=0}^{N-1} S(n)} \tag{7}$$

where $S(n)$ is the magnitude of the power spectrum for bin number n, bin center $f(n)$, and N total bins [50].

4.2.5 Spectral Features: Spectral Flux Spectral flux measures how fast power changes in a signal by comparing adjacent power spectra (Equation 8). In theory, depressed speech should waver more than

the steady voice of a healthy individual. To calculate it, the Euclidean norm of the difference in power between adjacent frames is measured:

$$SF(k) = \left\| |S(k)| - |S(k+1)| \right\| \tag{8}$$

where $S(k)$ is the power at frequency band with corresponding index k [50]. The spectral spread of each participant is normalized 0-1.

4.2.6 Spectral Features: Spectral Roll-Off Spectral roll-off is defined [50] as the frequency point at which 80% of the power spectrum lies beneath it, or as in Equation 9:

$$SR = 0.80 \sum_{n=0}^{k-1} S(n) \tag{9}$$

4.2.7 Cepstral Features Optimal representation of speech characterises an individual's unique "filter," or vocal tract, whilst removing any influence of the source. This is problematic, as per the source-filter model the two are inherently linked by convolution or multiplication in the time and frequency domains respectively. However, it is possible to use logarithms to separate the two by transforming the multiplications into summations:

$$C(z) = \log[X(z) * H(z)] = \log X(z) + \log H(z) \tag{10}$$

where X(z) and H(z) are the source and filter frequency responses [63]. If the filter primarily contains low frequencies and the source mainly high frequencies, an additional filter can theoretically separate the two. The Z-inverse of C(z), measured in units of frequency, is called the cepstrum.

Mel-frequency cepstral coefficients (MFCCs) are features commonly used in speaker recognition. A mel is simply a unit of measurement of perceived pitch, and takes into account the fact that humans have decreased sensitivity at higher frequencies. As with any short-term acoustic feature, the audio signal is assumed stationary over a small time scale (25 ms). If frames are shorter than this, not enough samples are present to adequately calculate speech properties, but if much longer, the signal changes too much throughout the frame. Frames are shifted by 10 ms to reflect signal continuity.

Once the FFT is computed over each frame, a mel filter bank is defined using a set number of triangular filters uniformly spaced in the mel-domain, and the log of the energy within the passband of each filter is calculated. Thirty filters were used based on results of prior optimisation for depression classification [50].

The discrete cosine transform (DCT) is then calculated on these logarithmic energies, and the MFCCs are the resulting coefficients. In doing so, energy is better represented according to human perception, and correlations between features are removed. Furthermore, by selecting only the first 12 coefficients, it is possible to isolate slower changes in filter bank energies, as higher frequency changes degrade recognition accuracy.

MFCC values provide information on the power spectral envelope of a sequence of frames. However, to obtain dynamic information on coefficient trajectories over time, Δ (differential) and $\Delta - \Delta$ (acceleration) coefficients can be calculated by Equation 11 below:

$$d_t = \frac{\sum_{\theta=1}^{\phi} \theta \left(c_{t+\theta} - c_{t-\theta}\right)}{2\sum_{\theta=1}^{\phi} \theta^2} \tag{11}$$

where d_t is the Δ coefficient at time t calculated from the static coefficients c over the window size ϕ of 9 frames [53]. $\Delta - \Delta$ coefficients were calculated in the same manner. In theory, depression should result in decreased articulatory precision that is then reflected in these values.

4.2.8 TEO-Based Features Teager Energy Operator (TEO)-based features are useful tools for analysing a signal's energy profile and the energy required to generate that signal [36]. When applied to speech production, they are capable of taking into account nonlinear airflow, and are thus particularly significant in stress recognition due to the turbulent (and thus nonlinear) airflow at more emotional states. Two main types of vortices contribute to voice quality– the first of which results from normal flow separation due to the opening and closing of the glottis and is responsible for speaker loudness and high frequency harmonics. The second type is caused by fast air flow in emotional or stressed states, which creates vortices around the vestibular folds and consequently produces additional excitation signals unrelated to the measured fundamental frequency generated by glottal closure [39, 74, 76, 75]. The operator used to generate this TEO energy profile is mathematically calculated via Equation 12 below:

$$\psi(x[n]) = (x[n])^2 - x[n+1] * x[n-1] \tag{12}$$

where ψ is the Teager Energy Operator and $x[n]$ is the corresponding n^{th} sample of speech [50]. Some studies have reported strong performance of these features in classifying depression [53, 52], which prompted their use in these experiments.

4.2.9 Statistical Analysis The aforementioned features were tested for their ability to discriminate between pairwise comparisons of healthy and depressed participants using t-tests with each WAV file used as a data point. Any features that were not statistically significant at the 0.05 alpha-level were not used in later modelling. One-tailed t-tests were used if the relationship between that feature and depression was known. For example, depressed participants should exhibit lower energy levels than their healthy counterparts. If the relationship was not known, two-tailed tests were performed.

5 RESULTS

Our analysis of relevant vocal cues of depression was done in three steps. Firstly we performed a brief visual inspection of features that the clinical literature suggests are strong indicators of depression. Secondly, we took inspiration from a study in audio-based emotion recognition and find which of the features that are valuable for emotion recognition also are statistically significant in detecting depression. Finally, we performed a Machine Learning analysis, in which we trained two simple classifiers to do subject-independent depression recognition. In the last study, we also experimented incremental greedy feature selection.

In general, the visual inspection of what are supposed to be relevant features for depression recognition did not reveal any strong patterns. Fig. 4 shows what were perhaps the most salient results, based on the Log Energy of the speech signal. Although most depressed participants seemed to generally have lower energy levels than their healthy counterparts as in Fig. 4(a), significant subject variations and an outlier obscured this relationship. Comparing participants directly in matched pairs proved much more

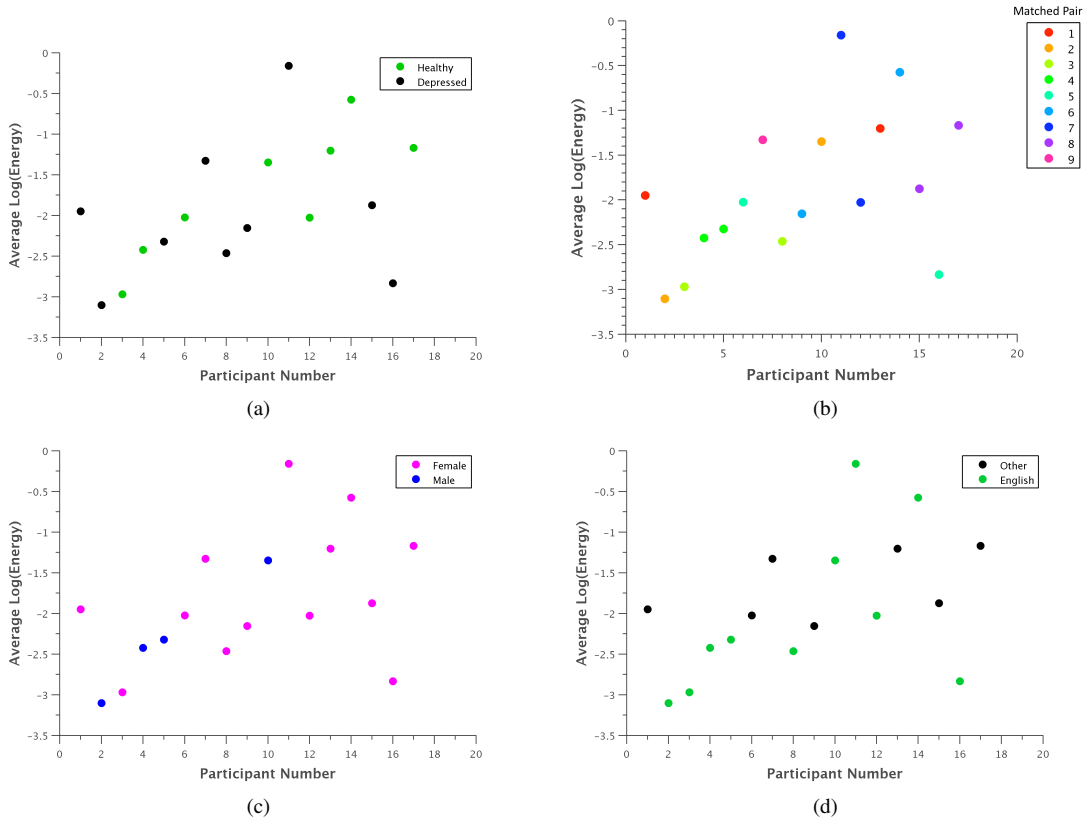


Figure 4. Average log energy was plotted for each participant grouped by (a) classification, (b) matched pairs, (c) gender, and (d) native language. Although patterns may be difficult to ascertain at first glance, in general each depressed participant exhibited significantly lower energy than their corresponding healthy control. Due to the uneven number of total participants, one point had no matched pair.

revealing, as almost all of the matched pairs demonstrated lower energy levels for depressed participants when compared with their healthy counterpart (Fig. 4(b)). Males generally had less energy than females (c), but with a sample size of four, this trend is not statistically significant. Native English speakers seemed to have wider variations in their average energy levels than non-native speakers (d), but again the trend is not statistically significant.

A statistical analysis was performed on select features to capture patterns that are not apparent from visual inspection. As the full space of features is very large, we focused our study on features that have previously been found to be of significant value in emotion recognition [32]. Because depression is inherently a mood disorder, the voice should exhibit similar cues as found in some negative emotions. Therefore we hypothesise that some of these features might also be significant for depression classification.

Each feature was compared within a matched pair of a healthy and a depressed individual, and the total number of statistically significant matched pairs is listed within each corresponding cell in Tables 1 and 2. If six or more (out of eight) pairs exhibited significant differences, the feature was deemed a potential indicator of depression and shaded in blue or teal. It was noted that this process does not necessarily imply that the difference is due to depression alone, but rather, that depression is possibly correlated with that particular feature. Both normal and concealed behaviour were tested. Features that were found to be significant for both normal behaviour and to detect emotion (as found by Iliou's study [32]) are highlighted with thicker black borders in Table 1. 24 features were noted as significantly different during normal behaviour.

Table 1. Features determined to be significantly different between healthy and depressed matched pairs during normal behaviour are shaded in blue. The number in each cell represents the number of matched pairs that were statistically different at the 0.05 α level. Features that were significant both during normal and concealed behaviour as well as in [32] are outlined in thick black lines for emphasis.

Normal Behavior							
Prosodic Feature	**Mean**	**STD**	**Mean of Derivative**	**STD of Derivative**	**Max**	**Min**	**Range**
Pitch	5	2	3	3	4	6	2
MFCC1	3	6	6	5	5	3	4
MFCC2	4	3	6	6	5	3	3
MFCC3	6	5	5	4	5	6	4
MFCC4	4	4	5	6	6	2	4
MFCC5	7	3	3	4	6	6	6
MFCC6	4	5	3	3	5	2	4
MFCC7	5	5	3	4	4	4	5
MFCC8	6	6	3	5	6	4	6
MFCC9	4	1	3	3	5	4	1
MFCC10	6	3	4	4	6	5	3
MFCC11	5	4	4	4	4	3	3
MFCC12	5	4	4	4	5	4	6
Energy	5	4	7	7	6	4	6

Table 2. The above table indicates speech features determined to be significantly different between health and depressed matched pairs during concealed behaviour. The number of differing matched pairs is specified in the cell, and features with six or more significant pairs are shaded in green. Features that were significant for both normal and concealed behaviour are again indicated by thick black lines.

Concealed Behavior							
Prosodic Feature	**Mean**	**STD**	**Mean of Derivative**	**STD of Derivative**	**Max**	**Min**	**Range**
Pitch	1	3	4	5	1	6	3
MFCC1	4	4	7	6	4	2	4
MFCC2	4	3	6	6	3	3	4
MFCC3	3	3	5	6	1	2	4
MFCC4	3	1	5	4	7	2	1
MFCC5	6	4	4	4	6	6	3
MFCC6	2	1	1	1	3	2	2
MFCC7	3	2	4	5	5	2	2
MFCC8	5	3	5	3	6	5	3
MFCC9	4	1	4	4	5	2	3
MFCC10	5	1	4	3	6	4	2
MFCC11	5	2	3	2	7	5	2
MFCC12	3	1	4	2	4	3	1
Energy	5	5	6	7	6	5	6

The significance of features between normal and concealed behaviour was also examined. Of the 17 features that exhibited significant differences between healthy and depressed individuals for concealed behaviour, 14 were also significant during normal behaviour (see Table 2). Although these features are not necessarily indicators of depression, it is nonetheless interesting that features that were considered significant for concealed behaviour were almost always significant for normal behaviour as well. The MFCCs were tested with a two-sided t-test whereas pitch and energy were tested with a one-sided test, as it was hypothesised that depressed participants would have lower pitch and energy values.

Log energy and its analogous statistics were some of the most distinguishing features between groups during spontaneous speech. Given that most depressed participants were generally softer-spoken than their healthy counterparts, this was somewhat expected. On the other hand, most features related to pitch and statistical functions thereof were surprisingly poor differentiators. This is perhaps due to the fact that pitch itself is extremely person dependent and might require normalisation for direct comparison. Additionally, participants who were non-native English speakers generally had wider variations in pitch irrespective of their classification. It was further noted that "significance" itself was tested differently in this study than in [32], which may account for some of the discrepancy.

5.1 MACHINE LEARNING EVALUATION

We performed Machine Learning analysis with two goals: to determine whether it is possible to detect depression even if a participant tried to conceal their depressive behaviour, and to to determine the minimum set of features needed to robustly discriminate between depressive and non-depressive behaviour.

For the first goal, machine learning models were trained on normal behaviour data to find an optimal classifier. Four different classifiers were assessed not only based on their subject-based classification accuracy, but also on their sensitivity. As these techniques would ideally be implemented in an automatic diagnostic device, it was more important to have high sensitivity (percentage of people correctly diagnosed with depression) than high specificity (percentage of people correctly identified as healthy). The trained model was then tested on the concealed data and its performance noted.

Naive Bayes, k-Nearest Neighbour, Random Forest and Neural Network classifiers were selected because they are known to perform well on such problems, and are very well understood. In training the models, leave-one-subject-out cross-validation was used to avoid one of the many common pitfalls in using machine learning techniques: overfitting, which can often lead to mistakenly overoptimistic results [34, 22]. Thus, we trained 16 separate models on the normal behaviour data, each time leaving out the data of one subject. Each model would then be tested on either the normal or concealed data of the left-out subject only. Note that in our approach, the concealed data was never used to train any of the models.

Of the four classifiers, both the Naive Bayes and kNN classifiers demonstrated remarkable classification accuracy for both normal and concealed behaviour, as shown in Table 3. Although both achieved classification accuracies (CA) of 88.24% on a per-subject basis, the Naive Bayes classifier exhibited superior sensitivities of 88.89% and 75% for normal and concealed behaviour respectively compared to 77.78% CA and 75% sensitivity for the kNN classifier. Results indicated that addition of the cepstral features did not improve results. Applying Occam's razor, it was found that best results are obtained using prosodic features in this setting.

To further refine a minimal set of robust indicators of depression, features that were found to be significant in previous sections were combined stepwise by category. For example, a model based solely on timing measures (TM) was created first, and other prosodic features of pitch and energy were incrementally added in and tested for improvement. Similarly, MFCCs were included in later iterations. The effects of feature selection are clearly shown in Table 4. The Naive Bayes model achieved a high level of accuracy using only two features: total time and average absolute deviation of pitch, whereas the kNN model required three: total time, average absolute deviation of pitch, and speaking rate. It is important

Table 3. Comparison of optimal classifier performances in terms of classification accuracy (CA) and sensitivity with the addition of prosodic (P), cepstral (C) or both categories of features. The results of the best classifiers are highlighted in grey. TM only means only timing measure features were used. If performances were numerically equivalent, minimum feature sets were considered superior.

Classifier	Feature Set	Normal Behavior		Concealed Behavior	
		CA	Sensitivity	CA	Sensitivity
Naïve Bayes	(P)– TM Only	76.47	66.67	56.25	50.00
	(P) –All	88.24	88.89	81.25	75.00
	(P) + (C)	88.24	88.89	81.25	75.00
kNN	(P)– TM Only	76.47	77.78	75.00	75.00
	(P) – All	88.24	77.78	75.00	75.00
	(P) + (C)	88.24	77.78	75.00	75.00
Random Forest	(P) – TM Only	76.47	77.78	68.75	75.00
	(P) – All	76.47	77.78	75.00	87.50
	(P) + (C)	76.47	77.78	75.00	87.50
Neural Network	(P) – TM Only	70.59	66.67	62.50	62.50
	(P) – All	76.47	77.78	62.50	62.50
	(P) + (C)	76.47	77.78	62.50	62.50

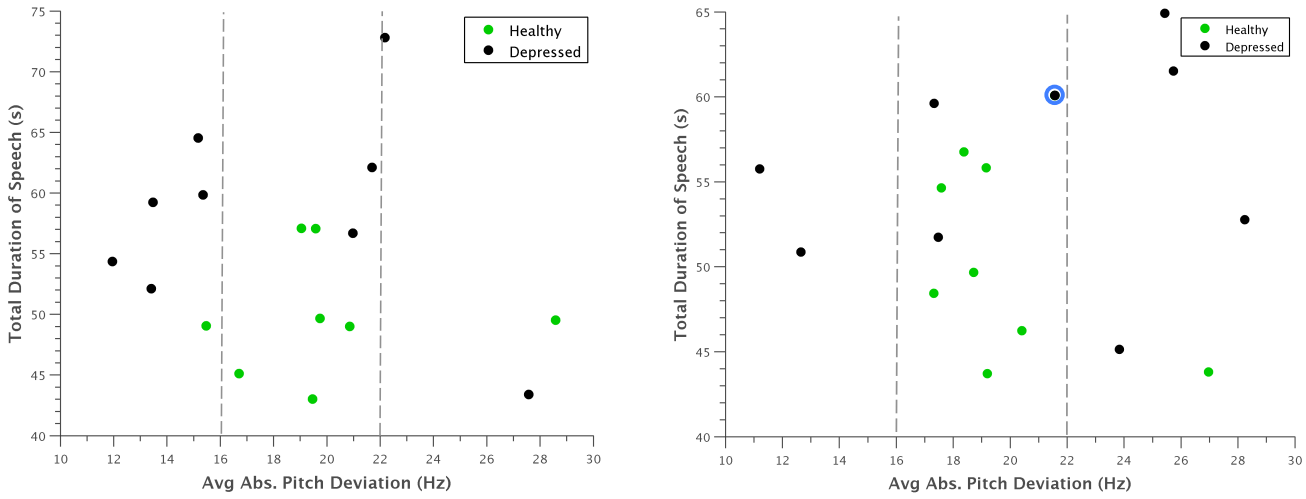
to note that the reclassified participant was not included in calculations performed on the concealed task data, so results were calculated out of 16 participants instead of 17.

Examining the data further revealed some interesting patterns. For example, when average absolute deviation of pitch was plotted against total time, there appeared to be a general range from 16-22 Hz that frame-to-frame pitch deviations for healthy individuals fell within (Fig. 5). Furthermore, during (a) normal behaviour, many of the depressed participants had noticeably lower pitch deviations than the control group, which logically corroborated with clinical observations. This pattern somewhat inverted during (b) concealed behaviour, although to such an extent that many depressed participants varied their pitch too much that the deviation still was not within the "normal" range.

A potential problem with the interpretation of Fig. 5 is that our experimental design only considers concealing voice control by depressed participants. The reason for this is that there is no need for non-depressed participants to appear non-depressed. Nevertheless, we want to judge whether the increased deviation in pitch of depressed also occurs in healthy controls if they conceal something from an

Table 4. Sequential feature selection ranked by information gain during normal behaviour. Optimal results were obtained with very few features.

Feature	Naïve Bayes Classifier		kNN Classifier		Inf. Gain
	CA	Sensitivity	CA	Sensitivity	
Average Absolute Deviation (Pitch)	64.71	66.67	58.82	66.67	0.594
Articulation Rate	82.35	77.78	70.59	77.78	0.521
MFCC8	70.59	66.67	70.59	77.78	0.403
Total Time	64.71	66.67	88.24	77.78	0.380
ASD	64.71	66.67	88.24	77.78	0.380
Phonation Time	70.59	77.78	70.59	66.67	0.380
MFCC1	70.59	77.78	76.47	77.78	0.359
Δ-MFCC7	76.47	77.78	76.47	77.78	0.330
STD of Derivative (Pitch)	70.59	77.78	76.47	77.78	0.330
MFCC7	64.71	77.78	76.47	77.78	0.330
Speaking Rate	76.47	77.78	76.47	77.78	0.315
Average Absolute Deviation (Energy)	76.47	77.78	76.47	77.78	0.315
MFCC10	70.59	77.78	76.47	77.78	0.286
Δ-MFCC6	70.59	77.78	76.47	77.78	0.168
Δ-MFCC9	70.59	77.78	76.47	77.78	0.095
Δ-MFCC8	70.59	77.78	76.47	77.78	0.095
STD (Pitch)	70.59	77.78	64.71	66.67	0.095
Δ-MFCC10	70.59	77.78	64.71	66.67	0.050



(a) Normal Behaviour

(b) Concealed Behaviour

Figure 5. Scatterplot of average absolute deviation in pitch against total duration of speech for (a) normal behaviour and (b) concealed behaviour. Most healthy participants appeared to have pitch deviations between 16-22 Hz, whereas depressed participants had markedly smaller deviations during normal behaviour. This trend was somewhat reversed when depressed participants were asked to conceal their depression, with many falling above this range. The reclassified participant is outlined in blue.

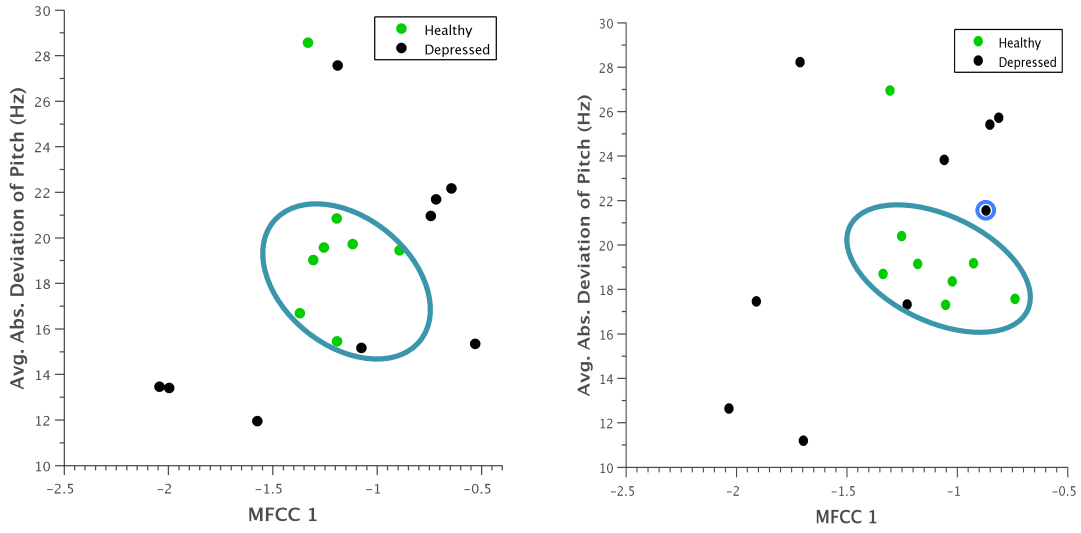
interviewer in similar experiments. For this we turned to the literature on lie detection. Agnelli & Cicery reported that in healthy subjects lying resulted in a greater number of pauses and words and either over-controlled reduced variation in tone or lacking control of tone (so more variable) [46]. The changes in tone within depressed subjects therefore follow the general pattern shown in healthy subjects in terms of tone that a few people do not change their tone under deception. However the overall conclusion that the machine is not fooled still stands.

A similar clustering occurred when average absolute deviation of pitch was plotted against the first mel-cepstral frequency coefficient (Figure 6). Although it is difficult to pinpoint a specific physical quantity that the first MFCC represents, the coefficients as a whole are used to uniquely characterise the vocal tract. This trend was thus noted as an interesting observation that could be investigated in future experiments.

Several expected clinical observations were confirmed visually by plotting some of these features against each other. For example, in Figure 7, articulation rate was plotted against total duration of speech. As indicative of the psychomotor retardation characteristic in many people with depression, depressed participants generally spoke at a slower rate and took more time to say the same phonetic content.

6 CONCLUSION

We presented our results of a study that looked into the automatic detection of depression using audio features in a human-computer interaction setting. In particular, we set out to discover how hard it would be to fool or cheat such an automated system. In our study on 17 matched healthy and depressed participants, we found that depressed participants seemed to follow the predicted pattern of lower energy levels in speech. Many of the prosodic and cepstral features that have before been used in emotion recognition were also found to be significant in depression recognition. However, not all features that were significant in differentiating depressed and healthy participants were the same as with those used in emotion recognition. These inconsistencies may suggest some dependency on the data collected and the methods used to acquire it, or perhaps on more fundamental differences between emotion and depression. One important finding of our study was that almost all features found to be significant during *concealed behaviour* were also



(a) Normal Behaviour (b) Concealed Behaviour

Figure 6. Scatterplot of average absolute deviation in pitch against the first mel-frequency cepstral coefficient for (a) normal behaviour and (b) concealed behaviour. A clustering of healthy participants within a particular range is again evident, and during concealed behaviour, depressed participants seem to vary their behaviour more substantially. The reclassified participant is again outlined in blue.

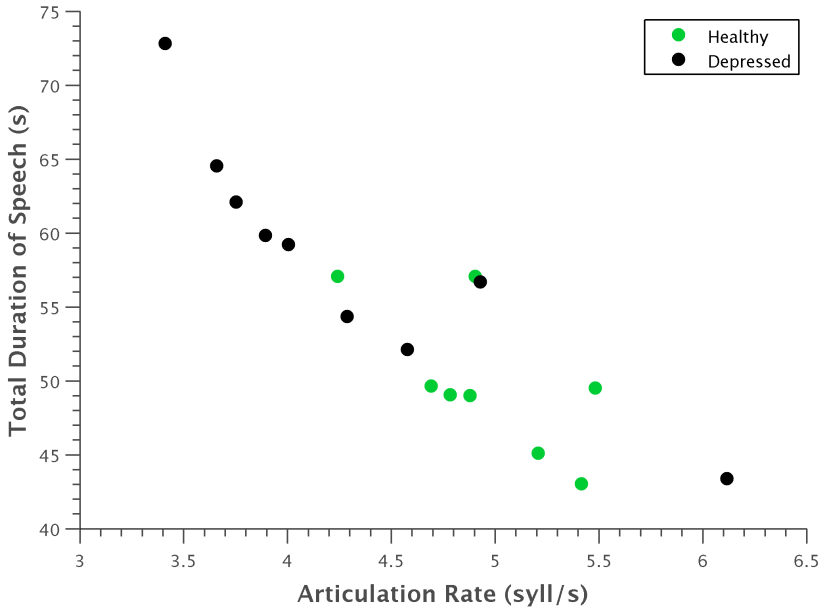


Figure 7. Scatterplot of articulation rate against total time. Depressed participants generally had a slower articulation rate and took more time to utter the same phonetic content. The strong linear correlation of these features is also evident.

significant during *normal behaviour*. This indicates that it may be hard to fool an automated system for depression screening. If supported by further evidence, this finding should have major implications for the development of reliable depression screening or monitoring systems. The second important finding from our study was that we attained high classification accuracy and depression recognition precision

using only simple Machine Learning techniques. Both k-Nearest Neighbours and Naive Bayes attained classification rates over 80% when using only 3 or 2 most salient features, respectively.

Classifications were surprisingly accurate given that only so few features were used, and remained high for concealed behaviour. This may indicate that the selected features are robust indicators of depression. However, our findings are presented in full knowledge that given such a small, restricted sample size ($n = 17$), findings from this study do not necessarily generalise to the population as a whole. A study on a larger population will form part of our future work. In addition, we will incorporate visual cues of depression to improve the accuracy of our predictions.

ACKNOWLEDGMENTS

The work of Michel Valstar, Richard Morriss, and John Crowe is partly funded by the NIHR-HTC 'MindTech'. In addition, part of Michel Valstar's work is funded by Horizon Digital Economy Research, RCUK grant EP/G065802/1. Part of Richard Morriss' work is funded by the National Institute of Healthcare Research Collaboration for Leadership in Applied Health Research and Care (NIHR CLAHRC0 East Midlands.

REFERENCES

[1]S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker. Detecting depression: A comparison between spontaneous and read speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7547–7551. IEEE, 2013.

[2]A. Altorfer, S. Jossen, O. Wurmle, M. L. Kasermann, K. Foppa, and H. Zimmermann. Measurement and meaning of head movements in everyday face-to-face communicative interaction. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 32(1):17–32, Feb. 2000.

[3]A. P. Association et al. *Diagnostic And Statistical Manual Of Mental Disorders DSM-IV-TR Fourth Edition*. American Psychiatric Publishing, Inc, 2000.

[4]A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597, 1996.

[5]A. T. Beck, C. Ward, M. Mendelson, et al. Beck depression inventory (bdi). *Arch Gen Psychiatry*, 4(6):561–571, 1961.

[6]P. Boersma and D. Weenink. Praat: doing phonetics by computer, version 5.4, [computer program]. http://www.praat.org/, 2014.

[7]M. Brookes et al. Voicebox: Speech processing toolbox for matlab. Technical report, University of London, 1997.

[8]I. M. Cameron, A. Cardy, J. R. Crawford, S. W. du Toit, S. Hay, K. Lawton, K. Mitchell, S. Sharma, S. Shivaprasad, S. Winning, et al. Measuring depression severity in general practice: discriminatory performance of the phq-9, hads-d, and bdi-ii. *British Journal of General Practice*, 61(588):e419–e426, 2011.

[9]M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder. Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35, 2004.

[10]K.-h. Chang. *Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring*. PhD thesis, EECS Department, University of California, Berkeley, May 2012.

[11]L. A. Clark and D. Watson. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of abnormal psychology*, 100(3):316, 1991.

[12]N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke. Modeling spectral variability for the classification of depressed speech. In *Interspeech*, pages 857–861, 2013.

[13]B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

[14]C. Diez-Quevedo, T. Rangil, L. Sanchez-Planell, K. Kroenke, and R. L. Spitzer. Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital spanish inpatients. *Psychosomatic Medicine*, 63(4):679–686, 2001.

[15]M. Dum, J. Pickren, L. C. Sobell, and M. B. Sobell. Comparing the bdi-ii and the phq-9 with outpatient substance abusers. *Addictive behaviors*, 33(2):381–387, 2008.

[16]D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on bio-medical engineering*, 47(7):829–37, July 2000.

[17]T. A. Furukawa. Assessment of mood: guides for clinicians. *Journal of psychosomatic research*, 68(6):581–589, 2010.

[18]J. Garcia-Campayo, A. Lobo, M. J. Perez-Echeverria, and R. Campos. Three forms of somatization presenting in primary care settings in spain. *The Journal of nervous and mental disease*, 186(9):554–560, 1998.

[19]J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing*, Dec. 2013.

[20]J. F. Greden, A. Albala, I. Smokler, R. Gardner, and B. Carroll. Speech pause time: A marker of psychomotor retardation among endogenous depressives. *Biological Psychiatry*, 1981.

[21]J. J. Gross and R. F. Muñoz. Emotion regulation and mental health. *Clinical psychology: Science and practice*, 2(2):151–164, 1995.

[22]I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[23]W. A. Hargreaves and J. A. Starkweather. Voice quality changes in depression. *Language and Speech*, 7(2):84–88, 1964.

[24]E. Harmon-Jones, D. M. Amodio, and L. R. Zinner. Social psychological methods of emotion elicitation. *Handbook of emotion elicitation and assessment*, pages 91–105, 2007.

[25]P. L. Harris, C. N. Johnson, D. Hutton, G. Andrews, and T. Cooke. Young children's theory of mind and emotion. *Cognition & Emotion*, 3(4):379–400, 1989.

[26]Health & Consumer Protection Directorate General. Improving the mental health of the population: Towards a strategy on mental health for the european union. Technical report, European Union, 2005.

[27]Health & Consumer Protection Directorate General. Mental health in the eu. Technical report, European Union, 2008.

[28]K. A. Hepner, S. B. Hunter, M. O. Edelen, A. J. Zhou, and K. Watkins. A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. *Journal of substance abuse treatment*, 37(3):318–325, 2009.

[29]M. K. Hinchliffe, M. Lancashire, and F. Roberts. Depression: Defence mechanisms in speech. *The British Journal of Psychiatry*, 118(545):471–472, 1971.

[30]H. Hirabayashi, K. Koshii, K. Uno, H. Ohgaki, Y. Nakasone, T. Fujisawa, N. Shono, T. Hinohara, and K. Hirabayashi. Laryngeal epithelial changes on effects of smoking and drinking. *Auris Nasus Larynx*, 17(2):105–114, 1990.

[31]R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt. On the relative importance of vocal source, system, and prosody in human depression. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[32]T. Iliou and C.-N. Anagnostopoulos. Classification on speech emotion recognition-a comparative study. *International Journal on Advances in Life Sciences*, 2(1 and 2):18–28, 2010.

[33]T. R. Insel. Assessing the economic costs of serious mental illness. *The American journal of psychiatry*, 165(6):663–5, June 2008.

[34]A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.

[35]R. Jouvent, D. Widlöcher, et al. Speech pause time and the retardation rating scale for depression (erd): Towards a reciprocal validation. *Journal of affective disorders*, 6(1):123–127, 1984.

590 [36]J. F. Kaiser. On a simple algorithm to calculate the energy of a signal. In *Acoustics, Speech, and Signal*
591 *Processing, 1990. ICASSP-90., 1990 International Conference on*, volume 381. Bell Communication
592 Research, Inc., April 1990.
593 [37]A. Karasz. Cultural differences in conceptual models of depression. *Social science & medicine*,
594 60(7):1625–1635, 2005.
595 [38]M. P. Kesarkar. Feature extraction for speech recognition. In *Tech. Credit Seminar Report, Electronic*
596 *Systems Group, EE. Dept, IIT Bombay*, 2003.
597 [39]S. Khosla, S. Murugappan, and E. Gutmark. What can vortices tell us about vocal fold vibration and
598 voice production. *Current opinion in otolaryngology & head and neck surgery*, 16(3):183–187, 2008.
599 [40]L. J. Kirmayer. Cultural variations in the clinical presentation of depression and anxiety: implications
600 for diagnosis and treatment. *Journal of Clinical Psychiatry*, 62:22–30, 2001.
601 [41]L. J. Kirmayer, J. M. Robbins, M. Dworkind, and M. J. Yaffe. Somatization and the recognition of
602 depression and anxiety in primary care. *The American journal of psychiatry*, 1993.
603 [42]K. Kroenke and R. L. Spitzer. The phq-9: a new depression diagnostic and severity measure.
604 *Psychiatric Annals*, 32(9):509–515, 2002.
605 [43]K. Kroenke, R. L. Spitzer, and J. B. Williams. The phq-9. *Journal of general internal medicine*,
606 16(9):606–613, 2001.
607 [44]S. Kung, R. D. Alarcon, M. D. Williams, K. A. Poppe, M. Jo Moore, and M. A. Frye. Comparing the
608 beck depression inventory-ii (bdi-ii) and patient health questionnaire (phq-9) depression measures in
609 an integrated mood disorders practice. *Journal of affective disorders*, 145(3):341–343, 2013.
610 [45]S. Kuny and H. Stassen. Speaking behavior and voice sound characteristics in depressive patients
611 during recovery. *Journal of Psychiatric Research*, 27(3):289–307, 1993.
612 [46]A. L. and C. R. The voice of deception:vocal strategies of naive and able liars. *J Nonverbal Behaviour*,
613 4(21):259–84, 1997.
614 [47]S. J. Leask, B. Park, P. Khana, and B. Dimambro. Head movements during conversational speech in
615 patients with schizophrenia. *Therapeutic advances in psychopharmacology*, 3(1):29–31, Feb. 2013.
616 [48]M. R. Lemke and a. C. Hesse. Psychomotor symptoms in depression. *The American journal of*
617 *psychiatry*, 155(5):709–10, May 1998.
618 [49]S. E. Lewis. *The social construction of depression: experience, discourse and subjectivity.* PhD
619 thesis, University of Sheffield, 1996.
620 [50]L. Low. *Detection of clinical depression in adolescents' using acoustic speech analysis*. PhD thesis,
621 RMIT University, 2011.
622 [51]L.-S. Low, M. Maddage, M. Lech, and N. Allen. Mel frequency cepstral feature and gaussian mixtures
623 for modeling clinical depression in adolescents. In *Cognitive Informatics, 2009. ICCI'09. 8th IEEE*
624 *International Conference on*, pages 346–350. IEEE, 2009.
625 [52]L.-S. Low, M. Maddage, M. Lech, L. Sheeber, and N. Allen. Influence of acoustic low-level
626 descriptors in the detection of clinical depression in adolescents. In *Acoustics Speech and Signal*
627 *Processing (ICASSP), 2010 IEEE International Conference on*, pages 5154–5157. IEEE, 2010.
628 [53]L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen. Content based clinical depression
629 detection in adolescents. In *Proc. Eur. Signal Process. Conf*, pages 2362–2365, 2009.
630 [54]J. D. Mayer and P. Salovey. Emotional intelligence and the construction and regulation of feelings.
631 *Applied and preventive psychology*, 4(3):197–208, 1995.
632 [55]P. Miller and L. L. Sperry. The socialization of anger and aggression. *Merrill-Palmer Quarterly*
633 *(1982-)*, pages 1–31, 1987.
634 [56]E. Moore, M. Clements, J. Peifer, and L. Weisser. Analysis of prosodic variation in speech for clinical
635 depression. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual*
636 *International Conference of the IEEE*, volume 3, pages 2925–2928. IEEE, 2003.
637 [57]E. Moore, M. Clements, J. Peifer, and L. Weisser. Comparing objective feature statistics of speech for
638 classifying clinical depression. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04.*
639 *26th Annual International Conference of the IEEE*, volume 1, pages 17–20. IEEE, 2004.
640 [58]E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser. Critical analysis of the impact of
641 glottal features in the classification of clinical depression in speech. *Biomedical Engineering, IEEE*
642 *Transactions on*, 55(1):96–107, 2008.

[59]P. J. Moses. The study of personality from records of the voice. *Journal of Consulting Psychology*, 6(5):257, 1942.
[60]J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 20(1):50–64, 2007.
[61]J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587, 2012.
[62]S. Newman and V. G. Mather. Analysis of spoken language of patients with affective disorders. *American journal of psychiatry*, 94(4):913–942, 1938.
[63]A. V. Oppenheim and R. W. Schafer. From frequency to quefrency: A history of the cepstrum. *Signal Processing Magazine, IEEE*, 21(5):95–106, 2004.
[64]P. F. Ostwald. *Soundmaking. The acoustic communication of emotion.* Charles C Thomas, 1963.
[65]R. Picard. *Affective Computing*. MIT Press, 1997.
[66]C. K. S. Quigley, K. A. Lindquist, and L. F. Barrett. Inducing and measuring emotion and affect: Tips, tricks, and secrets. *Cambridge University Press*, 2014.
[67]J. K. Rowling. *Harry Potter and the prisoner of Azkaban*. New York: Arthur A. Levine Books, 1999.
[68]A. Sano and R. W. Picard. Stress Recognition Using Wearable Sensors and Mobile Phones. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 671–676, Sept. 2013.
[69]G. E. Simon, M. VonKorff, M. Piccinelli, C. Fullerton, and J. Ormel. An international study of the relation between somatic symptoms and depression. *New England Journal of Medicine*, 341(18):1329–1335, 1999.
[70]R. L. Spitzer, K. Kroenke, J. B. Williams, P. H. Q. P. C. S. Group, et al. Validation and utility of a self-report version of prime-md: the phq primary care study. *Jama*, 282(18):1737–1744, 1999.
[71]D. E. Sturim, P. A. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Interspeech*, pages 2981–2984, 2011.
[72]E. Szabadi and C. Bradshaw. Speech pause time: behavioral correlate of mood. *American journal of psychiatry*, 140(2):265–265, 1983.
[73]D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518, 1995.
[74]H. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):599–601, October 1980.
[75]H. Teager and S. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.
[76]H. M. Teager and S. M. Teager. A phenomenological model for vowel production in the vocal tract. *Speech Science: Recent Advances*, pages 73–109, 1983.
[77]F. Tolkmitt, H. Helfrich, R. Standke, and K. R. Scherer. Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of communication disorders*, 15(3):209–222, 1982.
[78]US Department of Health and Human Services. Mental Health: A Report of the Surgeon General. Technical report, The Center of Mental Health Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health, Bethesda, 1999.
[79]M. Valstar. Automatic behaviour understanding in medicine. In *Proceedings ACM Int'l Conf. Multimodal Interaction*, 2014.
[80]M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014–3d dimensional affect and depression recognition challenge. *4th ACM international workshop on audio/visual emotion challenge*, 2014.
[81]M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013 - the continuous audio / visual emotion and depression recognition challenge. In *Proc. 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.
[82]A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'ericco, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Trans. Affective Computing*, 3:69–87, April 2012. Issue 1.

[83]W. WEINTRAUB and H. Aronson. The application of verbal behavior analysis to the study of psychological defense mechanisms. iv: Speech pattern associated with depressive behavior. *The Journal of nervous and mental disease*, 144(1):22–28, 1967.

[84]J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th ACM international workshop on Audio/visual emotion challenge*. ACM, 2014.

[85]J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 41–48, New York, NY, USA, 2013. ACM.

[86]Y. Yang, C. Fairbairn, and J. F. Cohn. Detecting Depression Severity from Vocal Prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150, Apr. 2013.

[87]T. Yingthawornsuk. *Acoustic analysis of vocal output characteristics for suicidal risk assessment*. PhD thesis, Vanderbilt University, 2007.