Multimodal Assessment of Depression from Behavioral Signals

Jeffrey F. Cohn, Nicholas Cummins, Julien Epps, Roland Goecke, Jyoti Joshi and Stefan Scherer

Reliable, valid, and efficient assessment of depression is critical to identify individuals in need of treatment and to gauge treatment response. Current methods of assessment are limited to subjective measures of patient self-report and clinical interview. They fail to take into account observable measures of behavior that could better inform detection of the occurrence and severity of depression. Recent advances in computer vision, signal processing, and machine learning have potential to meet the need for improved depression screening, diagnosis, and ascertainment of severity (i.e., assessment). This chapter reviews these advances. We describe multimodal measures of behavior and physiology, how these measures can be processed to extract features sensitive to depression, and how classification or prediction may be used to provide automatic assessment of depression occurrence and severity.

Following an overview in Sections 2 of how behavioral and physiological sensor signals can be processed in a multimodal manner, Sections 3 to 6 discuss in detail how to extract features sensitive to depression, based on insights from earlier scientific investigations and from studies of existing clinical assessment. These latter Sections also touch on key machine learning methods that have been adopted to date. Of particular interest is the fusion of information from different modalities, and Section 7 discusses fusion in the context of both classification and prediction.

For researchers investigating automatic analysis of depression, high quality research data is a critical concern. Similarly important for the practical application of multimodal depression assessment systems is their likely context of use. Both concerns are discussed in Section 8, while Section 9 provides an overview of key challenges in this research area.

In this chapter, the focus is primarily on depression. This emphasis reflects the emphasis in the field. Most work to date on automated, multimodal assessment of psychopathology has focused on depression. Other disorders, such as post-traumatic stress disorder, generalized anxiety disorder, traumatic brain injury, suicidality, dementia, Alzheimer's disease, schizophrenia, Parkinson's disease, and autism spectrum disorder have received less attention and accordingly are not discussed in detail. The approaches and many of the methods we discuss, nevertheless, are applicable to disorders that share behavioral features, are comorbid with depression, or otherwise share variance with depression. Dimensional models of psychopathology [HiTOP, Undated; Kotov, Gamez, Schmidt, & Watson, 2010] in particular group unipolar depression (Major Depressive Disorder and Dysthymia), Generalized Anxiety Disorder, PTSD, and in some models Borderline Personality Disorder together as distress orders. From this perspective, the research we review on depression is relevant to these disorders as well.

1. Depression

Depression is one of the most common mental disorders [Kessler, Chiu, Demler, and Walters, 2005] and a leading cause of disease burden worldwide [Mathers and Loncar, 2011; World Health Organization, 2014]. It severely impacts quality of life, is costly both for affected individuals and society, and increases risk for suicide [Blair-West, Cantor, Mellsop, and Eyeson-Annan, 1999].

Depression may be *uni*polar or *bi*polar. Unipolar is the more common and is typically what is meant by "depression." Bipolar depression includes manic or hypomanic episodes in addition to depressive episodes over the course of the disorder. Bipolar depression has not been a focus of research in multimodal machine learning; it is a topic for future work.

Although there are a several variants of (unipolar) depression, the most frequent and severe form is Major Depressive Disorder (MDD), which has a 1-year prevalence of about 7% in adults [American Psychiatric Association, 2015]. MDD and all forms of depression are distinct from sadness, a normal emotion that has adaptive cognitive and interpersonal benefits [Forgas, 2014]. Symptoms of depression include markedly depressed mood or markedly diminished interest or pleasure most of the day, significant change in weight or appetite, insomnia or hypersomnia, psychomotor retardation or agitation that is observable by other persons, diminished ability to think or concentrate, increased indecisiveness, fatigue or loss of energy, feelings of worthlessness or excessive or inappropriate guilt, and recurrent thoughts of death and suicidal ideation, intentions, or actions [American-Psychiatric-Association, 2013]. To meet diagnostic criteria for MDD, markedly depressed mood or loss of interest and four or more of the symptoms are required over a period of two weeks or more, and these symptoms may not be due to another disorder, medical condition, or other factor [American-Psychiatric-Association, 2013]. Because there are many possible combinations of these symptoms, depression is a heterogeneous disorder that varies in presentation and severity. It varies as well in etiology, functions, prognosis, and course.

Depression often is comorbid with other disorders and may be a prominent component of them as well. Generalized Anxiety Disorder (GAD) is often comorbid with depression, and depression is a strong component of Post-Traumatic Stress Disorder (PTSD). Commonalities among depression and other disorders have motivated calls for dimensional models of psychopathology [HiTOP, Undated; Watson and Clark, 2006; Watson et al., 2007] in place of the more traditional categorical models. Because most efforts in computer vision and signal processing relevant to depression have been informed by categorical models of depression, we pursue that perspective in the following. We wish to acknowledge, however, that the close relation between depression and other disorders, especially GAD and PTSD, implies that the approaches discussed herein are likely relevant to other disorders or dimensions of psychopathology as well.

Assessment of Depression

While laboratory tests for depression have been proposed (e.g., dexamethasone suppression test) [Carroll, 1984], none had proven specific for depression. To date, screening and diagnosis depend primarily on subjective reports from patients, their families, or caregivers. Diagnostic interviews include the SCID-5 [First, Williams, Karg, & Spitzer, 2015] and the M.I.N.I. [Sheehan et al., 1997].

These instruments can require an hour or more to administer, are highly dependent on the expertise of the interviewer, and may require additional measures to assess severity. Based on clinical trials, the Hamilton Rating Scale for Depression (HRSD) [Hamilton, 1960] is the current standard for assessing severity and treatment response [Gotlib and Hammen, 2002; Zimmerman et al., 2004].

For screening purposes, the Beck Depression Index (BDI) [Beck et al., 1996], the Center for Epidemiological Studies Depression Scale [Eaton, Muntaner, Smith, Tien, & M., 2004; Radloff, 1977; WW, C, C, Tien, and M., 2004], and the Patient Health Questionnaire (PHQ-9) [Kroenke et al., 2001] are widely used. These are all self-report inventories and lack specificity for depression (e.g., [Campbell and Cohn, 1991]). Newer screening techniques, such as Computing Adaptive Testing, tailor a test paradigm to match a patient's impairment level [Gibbons et al, 2012], but have yet to achieve clinical impact.

None of these methods explicitly take into account observable measures of behavior. Depression has marked observable influence on psychomotor functioning (retardation or agitation), expression of affect (reductions in positive affect and increases in negative), and interpersonal communication. Modalities include facial expression, gaze, head and body motion, and vocalization. The approaches reviewed herein quantify objective changes in these modalities to inform our understanding of depression and to detect depression and depression severity. They answer the need for reliable, valid, and efficient assessment of depression for research and clinical use.

Since 2009 the National Institute of Mental Health (NIMH), through the Research Domain Criteria (RDoC) project, has been encouraging research into new diagnostic tools for mental illness based on neurobiology and measurable behavioral signals [Cuthbert and Insel, 2013]. RDoC and related dimensional paradigms [HiTOP, Undated] emphasize broad, hierarchical constructs, such as distress, fear, and substance use disorders. For instance, both depression and generalized anxiety disorder are considered stress disorders by the hierarchical taxonomy of psychopathology [HiTOP, Undated]. Multimodal assessment can inform the relation between disorders and contribute to new diagnostic understanding.

The goals for objective, multimodal assessment of depression include: 1) Screening for depression, i.e. identifying from behavioral signals individuals who warrant referral; 2) Aiding rather than substituting or replacing clinical diagnosis, since diagnostic criteria include information that is occult to multimodal assessment (e.g., duration of episode) and like self-report measures lack specificity for depression (i.e., cannot rule out other causes for symptoms); and 3) Measuring severity of depression in individuals who have been diagnosed; and 4) Contributing to research on the validity of categorical versus dimensional perspectives on psychopathology. Clinically, multimodal assessment may be conducted in diverse contexts over extended periods of time without incurring reactivity effects. The remainder of this chapter offers a concise overview of the use of behavioral signals for the objective assessment of depression using automatic methods of feature extraction and classification or prediction.

2. Multimodal Behavioral Signal Processing Systems

Behavioral signal processing research shows potential advantages for assessing depression [Girard and Cohn, 2015; Cummins et al., 2015] and related disorders by employing quantitative methods to capture and model key behavioral signals. The term *behaviomedics* has even been coined to describe this broader research area [Valstar, 2014a]. Within the psychopathology literature alterations in facial movements, speech activity and body movement, are well reported. Recently, significant advances have been made into the automatic detection of depression using one or more of these behavioral modalities [Cummins et al., 2013; Girard et al., 2013; Joshi et al., 2013a; Scherer et al., 2014; Williamson et al., 2014], building on the recent availability of larger (publicly available) datasets collected under controlled conditions (Section 8).

The general structure of an automatic multimodal system for assessing depression is shown in Fig. 1, and begins with the acquisition of raw signal data from sensors (e.g. microphones, cameras, eye trackers, accelerometers). From these signals, features are extracted over regular intervals of time (frames or windows). The features characterise some aspect of the raw signal in a compact format, and are designed or chosen to be ideally both sensitive and specific to depression. Possibly after temporal alignment or synchronization across different modes, the features are then input to a machine-learning algorithm, which maps them to a depression state or score, based on prior knowledge in the form of labelled features from diverse levels of depression severity. To date, researchers have focused on three types of mappings: (i) depression detection, i.e. classification between non-depressed and depressed categories; (ii) severity recognition, i.e. classification between two or more ordinal categories (or score groups) of depression severity; and (iii) prediction, i.e. regression on the features to estimate a numerical depression score.

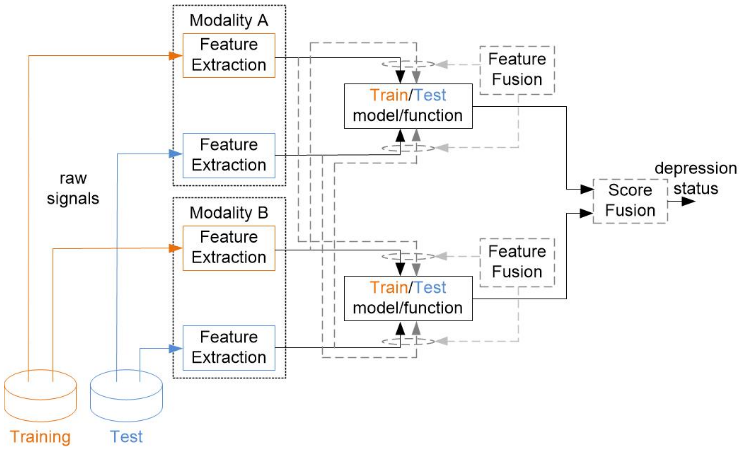


Figure 1. Overview of multimodal system for automatic assessment of depression, showing two modalities, with two of the more commonly employed different fusion possibilities shown dashed.

3. Facial Analysis

Since well before the automatic analysis of facial images began, clinicians and clinical scientists using manual analysis methods have observed marked changes in facial expression related to depression. Decreased smiling and laughter have been reported frequently (for citations, see [Girard, Cohn, Mahoor, et al., 2014]). Findings for negative expressions have been more mixed, with some studies reporting increases and others decreases, which may reflect the heterogeneity of depression. A key development toward automated measures was work by Cohn and colleagues [2009]. They applied both Ekman's manual *Facial Action Coding System* (FACS) [Ekman, Friesen, and Hager, 2002] and automated facial image analysis using *active appearance models* (AAM) to a large longitudinal database (more than 50 individuals) of depressed and no-longer depressed patients with the goal of comparing manual and automated methods. They found that automatic facial analysis closely approached manual coding in terms of two-class depressed/no-longer depressed classification accuracy. Together with follow-up studies using both automatic and manual FACS coding [Girard et al., 2013; Girard, Cohn, Mahoor, et al., 2014], they found high consistency both between manual and automatic coding of action units and in the pattern of depression effects that were found. This work opened the door to significant new opportunities for facial expression analysis using computer vision methods.

FACS describes facial expressions taxonomically in terms of *action units* (AU), which are the fundamental actions of individual muscles or muscle groups [Cohn, Ambadar, and Ekman, 2007]. An overview of literature on facial expression changes due to depression by Scherer et al. [2014] found that decreases in smiling and mouth movement and emotional expressivity and an increase in frowns are common across multiple studies. In the facial visual behavior descriptors selected for their own studies, Scherer et al. [2014] list smile intensity and smile duration, with the latter being important because some papers have reported depressed individuals smiling often but briefly. Smiling relates in part to context. Reed, Sayette and Cohn [2007] found that currently depressed subjects smiled as often as comparison subjects in response to comedy but more often used smile controls to dampen their smiles. Girard et al. [2013] found that among severely depressed individuals, both smiling and non-smiling expressions were more often associated with facial actions associated with contempt than for healthy controls. These kinds of changes are associated with lower AU 12 (lip corner puller / zygomaticus major) activity and higher AU 14 (dimpler / buccinator) activity. The influence of depression on negative affectivity also provides motivation for extracting facial information.

Most approaches to automated face analysis involve face detection, detection and tracking of facial landmarks, or fiduciary points, feature extraction, and action unit or expression detection. For nearly frontal face images (plus/minus about 15 to 20 degrees), face detection and rough head pose estimation can be accomplished using the Viola and Jones (VJ) face detector (OpenCV and MATLAB implementations available). For the VJ face detector, the required training time is longer and the false positive rate is higher than another recently proposed face detector available in the dlib C++ library, based on Histogram of Oriented Gradients (HOG) and a structural SVM based training algorithm.

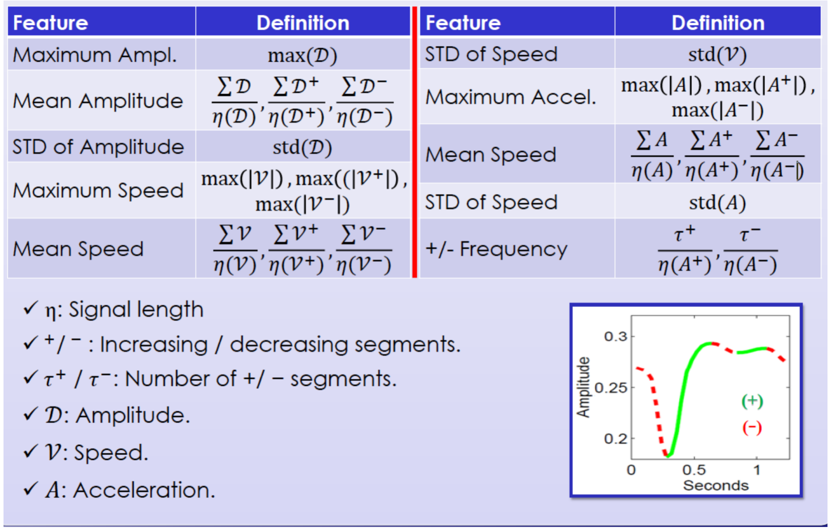
To quantify the shape and appearance of a face, deformable models have often been used. These include the active appearance model (AAM) [Baker, Matthews and Schneider, 2004; Cootes, Wheeler, Walker and Taylor, 2002] and constrained local model [Baltrusaitis, Robinson and Morency, 2012; Saragih, Lucey and Cohn, 2009]. The active appearance model is a two-dimensional triangular mesh with vertices arranged near fiduciary points in the facial image (e.g. the lower facial boundary, eyebrows, eyes and nose). In order to fit the model, some supervised learning is needed, which can be achieved by manually labeling a small proportion of video frames (5% in Cohn et al. [2009]) of a particular individual. Asthana et al. [2009, 2011] showed that a regression-based automatic face annotation and model building approach that only requires annotated frontal images, in the extreme case just a single annotated frontal image, drastically simplifies the deformable model building process. This is possible by posing the problem of learning the pattern of manual annotation between frontal face images, having arbitrary expressions, and corresponding face images at different poses in a data-driven regression framework.

In many cases, it is desirable that the initial active appearance model (before fitting) is a global model (e.g. across all sessions), so that a common basis for interpreting the model parameters for each individual or session can be established. Various features can be extracted from the positions, velocities and statistics of the vertices across a particular recording, and dimensionality reduction methods may be used to find a more compact representative feature set, given the typically large number of vertices.

More recently, supervised descent based models [Xiong and De la Torre, 2013] have made possible precise measurement of dense facial features that require no person-specific training. The approach of [Jeni et al., 2016] is capable of 3D alignment from 2D video over a radius of about plus/minus 60 degrees, which is more than sufficient for the range of head pose observed in spontaneous facial expression in interviews and small-group interactions.

A variety of approaches to action unit detection have been proposed (for reviews, see [Cohn and De la Torre, 2015; Corneanu, Oliu, Cohn, and Escalera, 2015, in press]). They typically involve face and facial feature detection, extraction of shape, appearance, or spatio-temporal features and training a classifier for action unit detection. Recent approaches eschew the use of hand-crafted features (e.g., SIFT or Gabor) in favour of empirical approaches to such as convolutional neural networks. Initial efforts suggest that such approaches are competitive with conventional approaches for action unit detection and may generalize better across databases [Ghosh et al., 2015]. A review is beyond the scope of the current chapter. Error analyses suggest that action unit detection is reliable within a range of about plus/minus 20 degrees from frontal and is comparable for men and women and robust to ethnic differences, although more work on this topic is needed [Girard, Cohn, Jeni, Sayette, and De la Torre, 2014].

These approaches to depression measurement or detection all involve learning a mapping from the occurrence or intensity of facial actions to depression. An alternative or adjunct approach is to learn a mapping directly from the timing of facial movement to depression. At least three appear promising. One [Dibeklioglu and Hammal et al., 2015] quantifies the dynamics of facial motion and head motion from dense 3D tracking [Jeni et al., 2016] irrespective of the configuration of facial expression. This approach samples a rich set of dynamic features (Figure 2). For the binary classification of depressed versus remitted, they obtained a correct classification rate of 86%. For the more challenging three-state classification of depressed, intermediate, and remitted, they obtained a correct classification rate of 78%.



Feature	Definition	Feature	Definition
Maximum Ampl.	$\max(\mathcal{D})$	STD of Speed	$\mathrm{std}(\mathcal{V})$
Mean Amplitude	$\frac{\sum\mathcal{D}}{\eta(\mathcal{D})}, \frac{\sum\mathcal{D}^+}{\eta(\mathcal{D}^+)}, \frac{\sum\mathcal{D}^-}{\eta(\mathcal{D}^-)}$	Maximum Accel.	$\max(\lvert A\rvert), \max(\lvert A^+\rvert),$ $\max(\lvert A^-\rvert)$
STD of Amplitude	$\mathrm{std}(\mathcal{D})$	Mean Speed	$\frac{\sum A}{\eta(A)}, \frac{\sum A^+}{\eta(A^+)}, \frac{\sum A^-}{\eta(A^-\rvert)}$
Maximum Speed	$\max(\lvert\mathcal{V}\rvert), \max((\lvert\mathcal{V}^+\rvert),$ $\max(\lvert\mathcal{V}^-\rvert)$	STD of Speed	$\mathrm{std}(A)$
Mean Speed	$\frac{\sum\mathcal{V}}{\eta(\mathcal{V})}, \frac{\sum\mathcal{V}^+}{\eta(\mathcal{V}^+)}, \frac{\sum\mathcal{V}^-}{\eta(\mathcal{V}^-)}$	+/- Frequency	$\frac{\tau^+}{\eta(A^+)}, \frac{\tau^-}{\eta(A^-)}$

✓ η: Signal length

✓ $^+/^-$: Increasing / decreasing segments.

✓ τ^+ / τ^-: Number of +/ − segments.

✓ $\mathcal{D}$: Amplitude.

✓ $\mathcal{V}$: Speed.

✓ A: Acceleration.

Figure 2. Dynamic features measured from pitch, yaw, and roll of head motion and motion of the first eigenvector of facial fiduciary points. From [Dibeklioglu, Hammal, Yang, & Cohn, 2015].

A second unsupervised approach is to detect depression from *spatio-temporal interest points* (STIP) [Joshi et al., 2013b]. The STIP method finds salient ‘interest points’ around the face where there is significant local variation both spatially and temporally. Around these points, histograms of gradient and flow can be calculated. The STIP features can easily grow to a massively high dimension, which can be handled through per-recording clustering (*k*-means) in the approach by Joshi et al. [2013b].

Last, a spatio-temporal descriptor, known as *local binary patterns in three orthogonal planes* (LBP-TOP) can be mapped to depression status. A spatio-temporal descriptor is a low-level representation of video/image data that describes basic characteristics such as texture and motion. LBP-TOP is computationally simple, relatively robust to variations in illumination, and has been shown to be effective for automatic facial analysis in depression without requiring any manual labeling. LBP operates at the pixel level, thresholding neighborhood pixels with the value of the central pixel to provide a binary pattern, whose histogram can be used as a texture descriptor. In facial analysis, spatial information can be retained by dividing the image up into local regions, and computing the LBP descriptors separately for each local region. LBP-TOP extends this spatial approach into the spatio-temporal domain, by concatenating LBP co-occurrence statistics in each of the three orthogonal directions. Like STIP features, LBP-TOP features are very high-dimensional and can be summarized using clustering methods, such as *Bag of Words* (BoW) approaches, before being used in depression classification [Joshi et al., 2013a]. STIP and LBP-TOP features can be combined using BoW for increased reliability (see Figure 3).

Both approaches, those that map depression from the configuration of facial expression and those that map depression from the dynamics of facial movement, have shown promising results for the task of discriminating between depressed and non-depressed subjects and change in severity of over the course of depression.

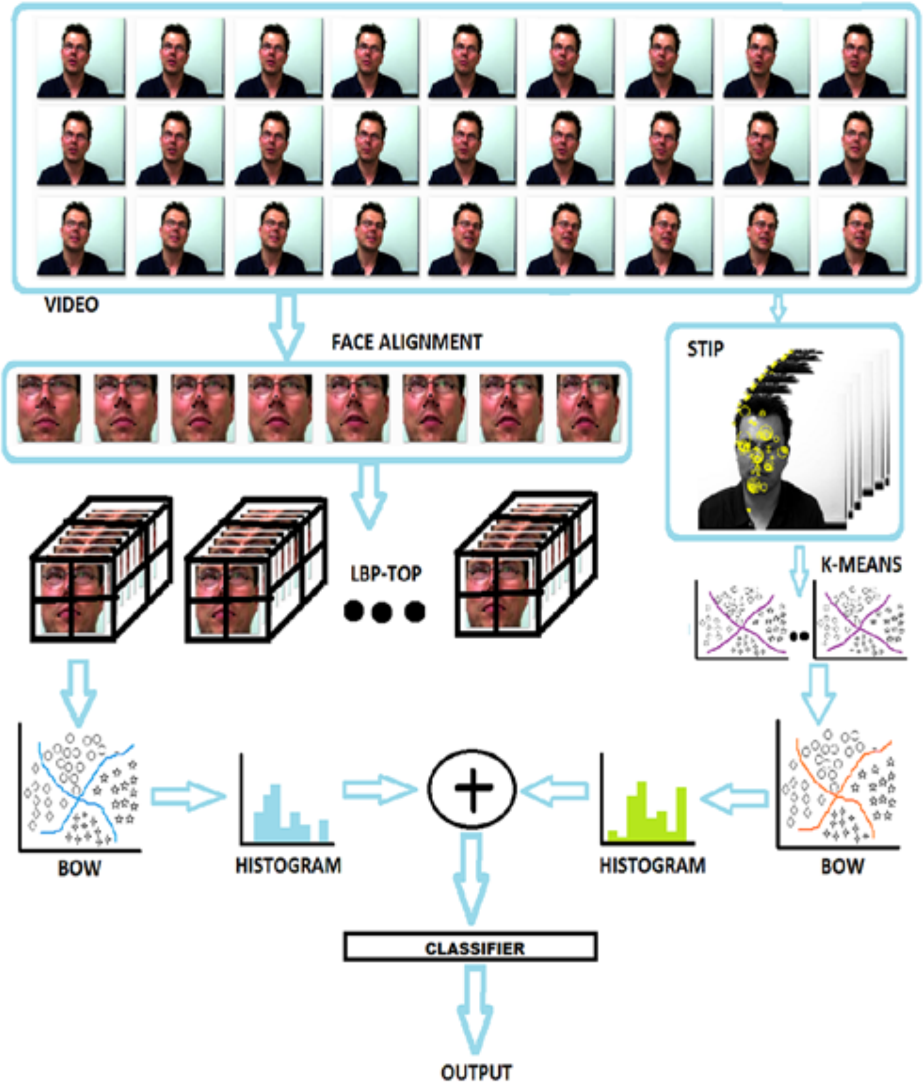


Figure 3. Combination of STIP and LBP-TOP features combined in a Bag of Words approach for action unit or depression detection. From [Joshi et al., 2012].

4. Speech Analysis

Speech is regarded as a key behavioral marker of depression [Cummins et al., 2015a]; cognitive and physiological alterations associated with depression influence speech production mechanisms and thereby the acoustic quality of speech. Indeed, prosodic abnormalities such as decreased verbal activity, slowed speech rate and monotonous pitch have long been associated with depression [Kraepelin, 1921]. Further, clinicians often use these acoustic alterations, in a subjective manner, when diagnosing a patient [Hall et al., 1995; Sobin and Sackeim, 1997]. Significant research has been undertaken to objectively identify potential relationships between changes in speech cues and depression [Cummins et al., 2015a]. However, many conflicting results, particularly for prosodic markers, are reported in the literature as to the exact nature of the observed effects.

The most commonly reported prosodic effects include: decreases in pitch variability [Nilsonne et al., 1987; Stassen et al., 1995]; decreases in energy variability [Quatieri and Malyska, 2012; Horwitz et al., 2013]. Whilst intuitively, these alterations match with subjective clinical descriptions of speech affected by depression as "flat" and "monotonous", there are also a substantial number of papers that do not support these findings [Alpert et al., 2001; Yang, Fairbairn, & Cohn, 2013; Stassen et al., 1991]. Alterations in speech rate measures are more consistently reported: both decreases in overall speech rate and increases in pause rate [Alghowinem et al., 2012; Mundt et al., 2012]. Recently, work by Trevino et al. [2011] reported that depression-induced changes in speech rate are potentially stronger when extracted at the phoneme level. Their analysis also indicates that depression has similar effects on phonemes that have been grouped together by manner of articulation.

Depression has also been linked to changes in a patient's *voice quality* [Cummins et al., 2015a]; a suprasegmental property of speech relating to phonation types. A range of voice quality measures/features such as increased spirantization [Flint et al., 1993], increased aspiration [Quatieri and Malyska, 2012], and decreased spectral harmonic and spectral tilt [Hönig et al., 2014] have been reported in speech affected by depression. These changes are all indicative of a more breathy phonation in depressed speech. A tense voice quality has also been linked with speech affected by depression: work led by Scherer demonstrates that Normalised Amplitude Quotient (NAQ) and the Quasi-Open-Quotient (QOQ) have each independently exhibited strong statistical significance for discerning between speech heavily affected by depression and speech not affected by depression [Scherer et al., 2013a, 2013b, 2013c]. NAQ is related to the speed of closure of the vocal fold periodic vibrations during voiced speech, while QOQ is related to the duration of vocal fold opening as a proportion of the pitch period.

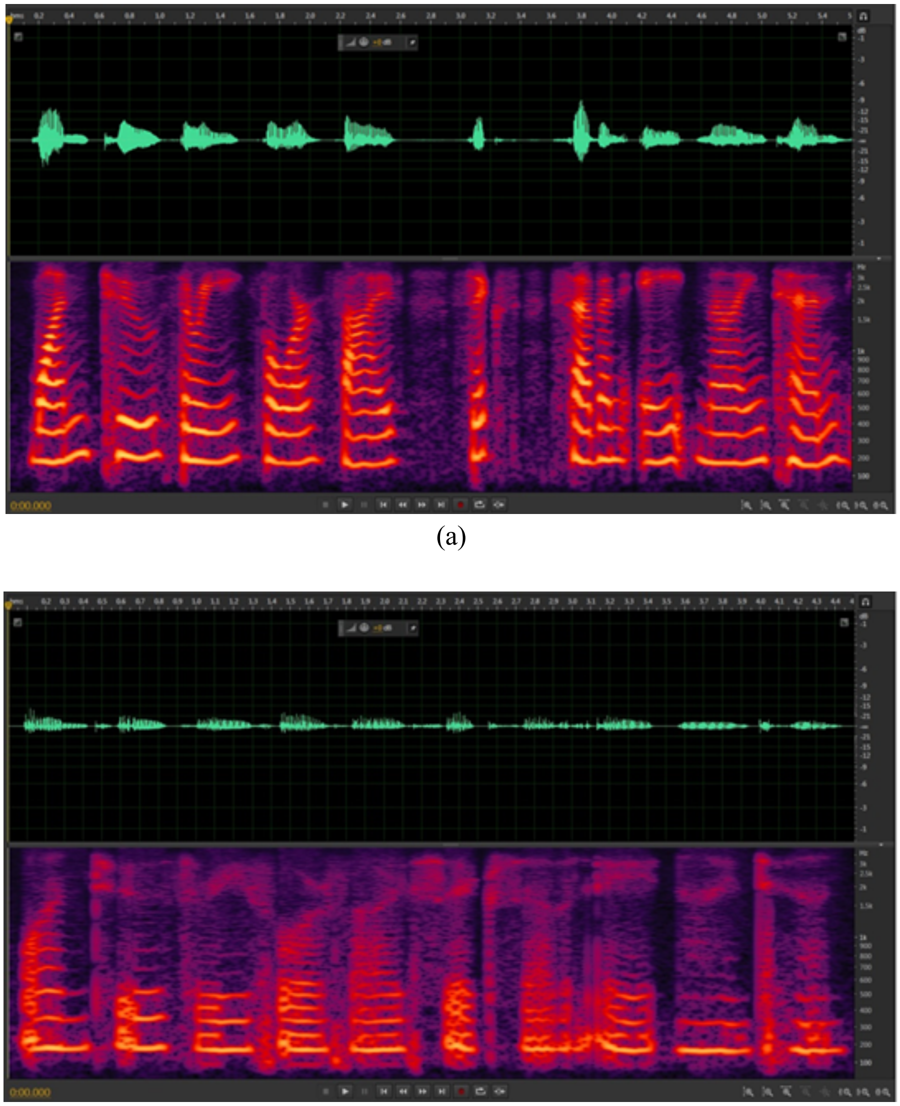
Changes in *formant features* - dominant components in the speech spectrum - also highlight the link between a tense voice quality and depression. Increases in muscle tension dampen the vocal tract resonances and limit articulator movements, altering formant behavior [Cummins et al., 2015a]. Alterations in formant behavior are well reported in the literature [Flint et al., 1993; France et al., 2000; Mundt et al., 2007; Scherer et al., 2016] and formant features have been widely employed in many speech-based depression detection systems [Helfer et al., 2013; Low et al., 2011; Williamson et al., 2013].

Depression has also been shown to manifest in the speech spectrum, a high-dimensional representation of the frequency distribution of a signal over a short time interval. The two most frequently observed effects are shifts in energy distribution and decreases the variance of sub-band energy measured at the utterance level [Cummins et al., 2015a]. Shifts in energy distribution is the more widely reported of the two effects; however, there are disagreements in the literature as to the exact nature of the effect. Whilst some papers report that increases in depression severity result in energy shifts from lower (below 500 Hz) to higher (500-1000 Hz) bands [France et al., 2000; Ozdas et al., 2004; Tolkmitt et al., 1982]; others report the opposite effect; more energy located in energy bands below 500 Hz in speech affected by depression [Yingthawornsuk et al., 2006]. More recently, it has been reported that speech affected by depression can potentially be characterised by a decrease in sub-band energy variability [Cummins et al., 2015c; Quatieri and Malyska, 2012]; this effect matches well with clinical descriptions of speech affected by depression as flat and monotonous. Work by Cummins et al. [2015c] indicates, over two commonly used speech-depression corpora, consistent trends of negative correlations in sub-band energy variability with increasing levels of depression.

Whilst there is a wide range of speech alterations attributed to depression, the clinical utility of many of the prosodic, source, formant and spectral features previously discussed is potentially limited. Most features were not specifically designed for the task of capturing depression in speech, and are hence also sensitive to the phonetic variability within an utterance as well as differences in speaker characteristics – age, gender, ethnicity – and emotional and social signals [Cummins et al., 2015a]. This issue is further confounded by the heterogeneous clinical profile of depression, highlighted by the many conflicting results seen in the literature [Cummins et al., 2015a]. There can be no guarantees as to the presence and strength of depression on any *one* of the speech features discussed. One approach that has been adopted to mitigate this is to use large or very large feature sets, for example those generated by the COVAREP or openSMILE toolkits for Audio/Visual Emotion Challenge (AVEC) [Valstar et al., 2014, 2013] baseline systems, however this approach needs to be employed with care to avoid overfitting.

Very recent research indicates that instead of attempting to analyse the effects of depression, there are advantages to analysing the effects of depression on overall speech motor control [Cummins et al., 2015c; Scherer et al., 2015, 2016; Williamson et al., 2013]. Speech motor control refers to the muscular systems and strategies needed to control speech production [Kent, 2000]; common speech motor control disorders include Dysarthria, Apraxia and Fluency Disorders such as stuttering. Speech alterations commonly associated with speech motor control such as intensity decay, prosodic abnormalities, articulatory and phonetic errors [Kent and Kim, 2003] can be visualised in the *spectrogram* of speech affected by depression (Figure 4).

Links between depression related effects and speech motor control can be inferred from the previously discussed prosodic, source, formant and spectral effects associated with depression. Prosodic effects relating to decreases in speech rate and phoneme rate observed in speech affected by depression provide evidence that depression can slow the movement of the articulatory muscles. Alterations to voice quality features provides evidence that depression can affect laryngeal coordination – also observable in the spectrogram of speech affected by depression (Figure 4). Finally, reductions in formant dynamics and sub-band energy variability indicate that depression can potentially increase articulatory effort.



(a)

(b)

Figure 4: Comparison of example spectrograms for the spoken digits 1 to 10 for (a) speech not affected by depression and (b) speech heavily affected by depression. The *x*-axes are time, the *y*-axes are amplitude (top plots) and frequency (lower plots), and the color intensity denotes regions of higher energy. Spectral features are well suited to capturing effects of depression: note the decreased intensity, flatter pitch contours, and decreases in harmonic to noise ratio comparing (b) with (a). From [Cummins 2016].

Williamson et al. [2013] proposed and introduced the *Vocal Tract Coordination* (VTC) feature space for depression prediction tasks, designed specifically to characterize reduced articulatory coordination as reflected in either formant frequencies or the mel-spectrum. Specifically, VTCs capture changes in correlation that occur at different time scales across either feature space. VTC feature spaces were used in both the AVEC-2013 and -2014 top-ranked systems; for full details on their extraction the reader is referred to Williamson et al. [2013] and Williamson et al. [2014].

Recently, results presented in Scherer et al. [2015, 2016] indicate that there are significant reductions in *Vowel Space Area* (VSA) in speech affected by depression. VSA is measured using the Euclidean distances between the F1 and F2 coordinates of the corner vowels /i/, /u/, and /a/ and is typically manually extracted from an utterance [Skodda et al., 2012]. Scherer et al. [2016] present a method for automatically measuring VSA and evaluate this on a range of commonly used depression-speech corpora to highlight its suitability for characterising speech affected by depression. Cummins et al. [2015c], using measures of acoustic variability, quantified the effects of depression on per-utterance Gaussian mixture models of *mel frequency cepstral coefficient* (MFCC) features. Results presented by the authors demonstrate that depression significantly reduces both the local and global spread of phonetic events. These findings indicate that speech affected by depression can be characterised by a reduction in the number of distinct speech sounds produced and a reduction in overall speaking effort and animation. Further, they provide strong evidence that depression should be considered a speech motor control disorder; similar characterizations are found in the appropriate literature [McRae et al., 2002; Orozco-Arroyave et al., 2014; Sapir et al., 2010].

A range of investigations have been undertaken into speech-based systems for the automatic classification of depression. Whilst a range of different classifiers have been explored, arguably *Support Vector Machines* (SVM) and *Gaussian mixture models* (GMMs) are the most typically used [Alghowinem et al., 2013, 2012; Cohn et al., 2009; Cummins et al., 2014, 2013a, 2013b, 2011; Helfer et al., 2013; Low et al., 2011, 2010; Ooi et al., 2013; Scherer et al., 2013a, 2013b; Sturim et al., 2011; Trevino et al., 2011; Valstar et al., 2013]. This is not surprising, as both techniques are suitable for robustly modelling small/sparse datasets and have a range of well-established software implementations (e.g. Weka, HTK, SVMTorch).

Typically, *detection* style classification papers have tended to adopt a more data-driven focus; a range of papers have investigated suitability of forming a classification systems from combinations of groups of prosodic, source, formant and spectral features with binary classification accuracy ranges between 60-80% reported [Alghowinem et al., 2012; Cummins et al., 2011; Low et al., 2010; Moore et al., 2008; Ooi et al., 2013]. On the hand *severity* style classification problems have tended to be more knowledge driven, for example, using single prosodic, voice quality or formant based features respectively [Cohn et al., 2011, Scherer et al., 2013a, 2013c and Helfner et al. 2013]; typical accuracies of 60-80% (3-5 classes) have been reported for depression severity classification.

Due to the public availability of the AVEC datasets (Section 8), speech based depression prediction approaches have recently gained greater attention. The AVEC data-centric multimodal depression severity prediction challenge (discussed in further detail in Section 7) audio baseline performances were set using the popular brute-forced approach which combines a high dimensional (2268 features) multivariate acoustic feature space and a *Support Vector Regressor* (SVR) back-end. This configuration of features and regression achieved results, in terms of root mean square error (RMSE) between predicted and clinician-assessed BDI scores, of 14.12 and 12.57 on the AVEC 2013 and AVEC 2014 test sets respectively [Valstar et al., 2014, 2013].

A range of regression approaches have been trialled on the AVEC data; as well as SVR [Cummins et al., 2014, 2013b; Mitra et al., 2014; Pérez et al., 2014; Sidorov and Minker, 2014; Valstar et al., 2014, 2013], Decision trees, [Kaya et al., 2014b], Relevance Vector Machines [Cummins et al., 2015b], Gaussian Processes [Pérez et al., 2014], Extreme Learning Machines [Kaya et al., 2014a] and Generalized Linear Models [Senoussaoui et al., 2014] have all shown reasonable performance – RMSEs between 9 and 10 – when predicting depression. However, underlying these approaches is the assumption that there is a regular relationship between changes in speech features and changes in depression scores; given the ordinal nature of depression scores such a relationship may not exist [Cummins, 2016].

To help account for the lack of regular relationship, less conventional regression approaches are starting to be explored for depression score prediction, the most promising of which are the 2-stage regression systems [Williamson et al., 2013; Cummins, 2016]. Williamson et al. [2013] proposed the *Gaussian Staircase Regression* (GSR) approach. The idea of this system was to capture the irregular relationship between speech features and depression scores by using a series of models – single Gaussians – along the depression score axis the test statistics from this pseudo GMM was then used as the basis for univariate depression prediction. As with VTCs, GSR prediction was a major system component in both the AVEC-2013 and 2014 winning entries, attaining RMSEs of 8.50 and 8.12 (not combined with video features, see Section 7) respectively [Williamson et al., 2014, 2013].

5. Body Movement and Other Behavior Analysis

A large body of research has examined the relationship between nonverbal behavior and clinical conditions. Most of this research resides in clinical and social psychology and communication science and, until very recently, the vast majority relied on manual annotation of gestures and facial expressions. Despite at least forty years of intensive research, there is still surprisingly little progress on identifying clear relationships between patient disorders and expressed behavior. In part, this is due to the difficulty in manually annotating data, inconsistencies in how clinical states and expressed behaviors are defined across studies, and the wide range of social contexts in which behavior is elicited and observed. Despite these complexities, there is general consensus on the relationship between some clinical conditions (especially depression and social anxiety) and associated nonverbal cues. In the following, we provide a review of key findings on non-verbal, non-facial behavior and its relationship to clinical conditions.

For computing body movement patterns, nine body parts relative to the torso centre have been considered. These body parts consist of the upper and lower parts of left and right arm, head, and upper and lower parts of left and right leg. A state-of-the-art Mixture of Parts based human body detector was employed to obtain an approximate location of various body parts in a video frame by Yang and Ramanan [2011]. The central location of nine parts can be transformed into polar coordinates. In order to compute the motion pattern in the body parts, the torso centre can be considered as the reference point. Polar coordinates for each part representing orientation and distance from the torso centre can be computed from the real pixel coordinate values. The motion patterns can be joined part-wise and a polar plot can be computed based on the values of distances and orientations depicting overall movement by all parts combined. In Figure 5, the plot represents the relative movement of the body parts with respect to the torso in a video.

The change in body movement patterns with changing severity of depression has been studied using within-subject longitudinal data [Joshi 2013c]. It was observed that there is a significant drop in the frequency of overall body movements when the participants were diagnosed with severe depression (Figure 5), and this is supported by other studies [e.g. Girard, 2014].

Emotional expressivity over time and emotional movements are also indicative of an underlying clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity [Perez and Riggio, 2003; Bylsma et al., 2008], less intense and shorter smiles [Scherer et al., 2013b, 2014], fewer mouth movements [Fairbanks et al., 1982; Schelde, 1998], more frowns [Fairbanks et al., 1982; Perez and Riggio, 2003] and fewer gestures [Hall et al., 1996; Perez and Riggio, 2003]. Some findings suggest it is not the total quantity of expressions that is important, but their dynamics. For example, depressed patients may frequently smile, but these are perceived as less genuine and often shorter in duration [Kirsch and Brunnhuber, 2007] than in non-clinical populations. Social anxiety and *post-traumatic stress disorder* (PTSD) share some of the features of depression [Scherer et al., 2013b, 2014, 2016] and also have a tendency toward heightened emotional sensitivity and more energetic responses including hypersensitivity to stimuli: e.g., more startle responses, and greater tendency to display anger [Kirsch and Brunnhuber 2007] or shame [Menke, 2011].

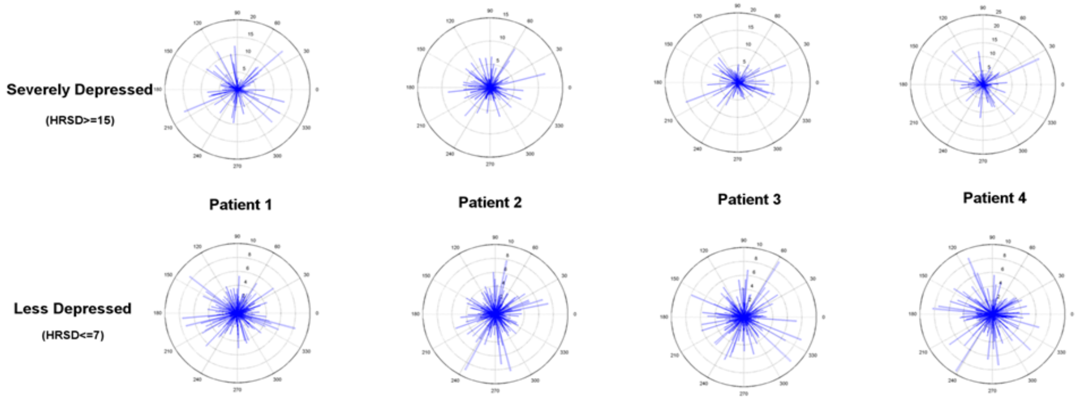


Figure 5: Body movement pattern for four individuals who were severely depressed at one point in time and have shown improvement over the course of treatment, from [Joshi 2013c]. Each column shows two plots depicting body motion patterns, which belong to the same participant in two different states. The blue lines in the polar plots represent the observed body movement pattern during the interview session. Note the denser blue region in the bottom row, depicting more movement under low severity of depression.

Certain gestures are seen with greater frequency in clinical populations. Fidgeting is often reported. This includes gestures such as tapping or rhythmically shaking hands or feet and is seen in both anxiety and depression [Fairbanks et al., 1982; Scherer et al., 2013b, 2014]. Similarly, “self-adaptors”, such as rhythmically touching, hugging or stroking parts of the body or self-grooming, e.g. repeatedly stroking the hair [Fairbanks et al., 1982] or touching the face [Scherer et al., 2013b, 2014], have been identified to be of interest in this field of research [Ekman and Friesen, 1969].

Gaze and mutual attention are critical behaviors for regulating conversations, so it is not surprising that a number of clinical conditions are associated with atypical patterns of gaze. Depressed patients have a tendency to maintain less mutual gaze [Waxer, 1974], show nonspecific gaze, such as staring off into space [Schelde, 1998] and avert their gaze, often together with a downward angling of the head [Perez and Riggio, 2003; Scherer et al., 2013b, 2014], which may be sustained for longer than usual [Alghowinem et al., 2013b]. The pattern for depression, PTSD, and perhaps other distress disorders is similar, with patients often avoiding direct eye contact with the clinician [Scherer et al., 2013c].

6. Analysis using Other Sensor Signals

A number of other signal types that can be practically acquired out of the clinical context have been investigated as indicators of depression. One of these is eye movement, with Abel et al. [1991] finding that smooth pursuit (ability of eye gaze to track a moving visual target) and saccade movements were impaired for depressed patients. These findings were confirmed using computer vision analysis, which provided depression detection accuracies of 70-75% for a gender-balanced database of 30 depressed and 30 control participants [Alghowinem et al., 2013], based on features extracted using ratios of horizontal and vertical eye position (relative to the eye corners and eyelid/eye fold) and eyelid opening ratio. It must be noted, however, that dysfunction in saccadic movement is likely to be non-specific to depression. They are reliably found in patients with schizophrenia and their first-degree relatives and are considered an endophenotype of schizophrenia [Calkins, Iacono, & Ones, 2008; Hong et al., 2008; Levy, Holzman, Matthysse, & Mendell, 1993].

Pupil diameter has also been investigated in studies by Siegle et al. [2001, 2003] and Burkhouse et al. [2015], who found that depressed individuals were slower to respond to emotional stimuli and spent more time processing those stimuli, as gauged by pupillometry. The measure of pupil diameter and reactivity to emotional stimuli has been shown to correlate with depression risk [Burkhouse et al., 2015] and the propensity for depressive rumination [Siegle et al., 2003]. While pupil diameter seems to be a strong behavioral predictor of depression it has not yet found its application in a large number of studies involving automatic behavior analyses. With improving computer vision techniques and facial tracking algorithms it is only a matter of time until the automatic and robust assessment of pupil diameter will become a commodity. As with saccades, indeed more so, pupil dilation lacks specificity for depression. Pupil dilation is a measure of sympathetic arousal, which is a primary dimension of emotion.

Other autonomic measures have been investigated, including phasic skin conductance, heart rate, and vagal tone [Dawson et al. 1985]. While these lack specificity for depression, they may prove useful especially in detecting change in severity over time.

In recent years, there has been considerable interest in signals that can be acquired from mobile device sensors [Glenn and Monteith, 2014], such as GPS and accelerometry [Schueller et al., 2014], real time self-monitoring of mood [Proudfoot et al., 2013] and text acquired from social media interaction [Nguyen et al., 2015; Milne et al., 2015]. Other studies have also revealed links between social connectivity and mental illnesses [Rosenquist et al, 2011, Pachucki et al, 2015]. It has been speculated that tracking changes in social networks using smart devices might reveal information on changes in at-risk individuals' mood state [Larsen et al, 2015].

There is also huge increase in the trend of using social networking sites, such as Facebook and Twitter, to engage and socialize with people. Recently, some studies have leveraged such linguistic information and created behavioral models to specifically predict depression [Park et al., 2012; Nguyen et al., 2014]. On observing Twitter feeds to analyse depression, it was deduced that there is a basic difference in the reason for using social media amongst depressed and non-depressed cohorts. Non-depressed participants used Twitter as an information consuming and sharing platform, whereas depressed participants perceived it as a medium for emotional interaction and social awareness [Park et al., 2012]. A detailed comparison of the characteristics of online depression communities and other popular and trending online communities was conducted by Nguyen et al. [2014], who found a significant difference between the written content, writing styles and affective information of the content of the two groups. Sentiment analysis showed that the textual content from depressed communities had lower valence (i.e. contained more negative words such as ‘hate’, ‘hard’ and ‘pain’) than that generated by people in the other online communities. Other studies have shown lowered online social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts [Choudhury et al., 2013]. Linguistic approaches have also been combined very successfully with other modalities (see e.g. [Valstar et al. 2016] and citing papers), and there are interesting prospects for future research in this area.

7. Multimodal Fusion

There is an extensive literature in psychology that points to various affective impairments associated with depressive disorders, as discussed in previous sections in detail. However, clearly defined markers of behavioral changes due to depression are still to be identified, which can be attributed to the different means of eliciting and observing expressed behaviors as well as the heterogeneous nature of depression symptoms. Clinicians study various facial movements, gestures, body movements and speech attributes to establish the differences in the expressivity in depressed patients. A multimodal automatic system incorporating inputs from various channels will likely prove to be better than unimodal techniques.

It has already been established that integration of different modalities yields a better performance for inferring affect, mainly in a classification context. One of the first works to integrate information from speech signals, facial dynamics and body movement to detect depression was presented by Joshi et al. [2013a], also in a classification paradigm. Intra-facial muscle movements and the movements of the upper body were analysed by computing spatio-temporal descriptors. In addition, various audio features (fundamental frequency, loudness, intensity and mel-frequency cepstral coefficients) were computed. A Bag of Word framework was then proposed to incorporate the audio and video signals separately. Two fusion paradigms were investigated: (i) Decision-level fusion (late fusion) classification was performed separately on each set of BoW features (audio and video), and then AND or OR decision-level fusion was performed; and (ii) Feature-level fusion (early fusion) classification was performed on the concatenated BoW features (audio and video). In their investigation, the feature-level (early) fusion proved more effective than decision-level (late) fusion in terms of two-class depression detection accuracy [Joshi et al. 2013a].

The BoW approach deserves some additional explanation. As originally proposed, BoW represents a document based on the unordered word frequency. In the context of the problem here, for example for vision based analysis [Joshi et al. 2013a], a video clip (set of video frames) is a document in the BoW sense. The computed spatio-temporal descriptors, which represent facial or body movement and provide discriminative information to detect depression, are “words”. During the training phase, a dictionary (video codebook) is learned from the computed words. One of the main advantages of the BoW approach is the resulting low-dimensional feature which is generated, in the form of a histogram. Another benefit of using a BoW framework is that it handles different feature lengths, since the recorded data to be processed will in general have different durations based on individuals' responses. The use of codebooks makes it simpler to deal with such samples of different length.

A further advantage is that if multiple modalities are represented using BoW, then (due to the histogram representation), they are well calibrated for fusion. Integration of audio and video codebooks in a multimodal framework has been shown to produce better performance than the corresponding unimodal audio or video systems, in a two-class depression classification context

[Joshi et al. 2013a]. The high performance of the integrated system indicates the importance of different channels, which, when fused, were found to provide an absolute gain of about 10% over individual channels in two-class classification [Joshi et al. 2013a].

In an attempt to expedite the advancement of multimodal depression assessment technologies, a depression sub-challenge was organized as part of the Audio/Visual Emotion Challenge (AVEC) [Valstar et al., 2013, 2014] with the aim of estimating the self-reported level of depression on a common database (see Section 8) and reference feature set. The AVEC challenge poses depression assessment as a prediction problem, and since its initiation, has seen multimodal depression assessment concentrate primarily on prediction problems and prediction fusion.

In one entry to the challenge, Cummins et al. [2013] proposed a multimodal framework. For the audio modality, acoustic supervectors were extracted using GMM of MFCC feature space. In the vision pipeline, clustering was performed on the video clips using spatio-temporal low level features. For each cluster, Pyramid of Histogram of Gradient (PHOG) was computed on the central frame and further embedded in a BoW framework. The authors employed feature fusion and trialled various different SVR kernels for prediction. Results presented show that their multimodal approach was able to outperform both the audio and visual challenge baselines.

A multimodal framework based on audio, text and video modalities was presented by Gupta et al. [2014]. In their system, the openSMILE toolkit based speech features shared by the AVEC 2014 organisers were augmented by video feature streams. Fiducial points and motion based features were computed and fused with the Local Binary Pattern features. For analysing the text, crowd-sourcing was performed to generate the transcripts. Furthermore, a lexicon was generated using the transcripts. The authors trailed a range of systems consisting of feature-level fusion followed by supervised feature dimensionality reduction and a SVR back-end to infer the depression intensity label.

A multimodal approach based on speech spectral information and facial action units was also proposed for AVEC [Williamson et al. 2014]. Their speech system consisted of Vocal Tract Coordination features (outlined in Section 4), phoneme rate features, phoneme pitch dynamics and an alternate vocal tract coordination feature; the correlation between formant frequencies and cepstral peak prominence. The authors complemented these features with facial features, derived from facial action units, designed to capture visual changes in muscle coordination. The training data were partitioned into eight depression score-dependent ranges and single Gaussian were estimated for each individual range, for each feature set. A Gaussian staircase regression (also outlined in Section 4) model was trained to map the input to a log-likelihood based score (between the low and high depression pseudo GMM's created by the aforementioned partitioning) for each feature. Their system trialled three different predictors comprising linear combinations of GSR log-likelihood scores from the different features spaces. Then their predictions were combined according to (i) the accuracy of each predictor and (ii) whether the subjects in the AVEC database were 'repeat' or 'non-repeat' (appeared in database more than once or once respectively).

Motivated by the highly promising GSR approach, Cummins [2016] proposed a generalized 2-stage framework for prediction against an irregular variable. In the first stage, a series of models are used to segment the depression score axis and extract a rank score per segment, in the second stage the rank scores are used as the basis for performing the final depression score prediction. Provided the segmentation of the depression score axis is meaningfully chosen, the first stage of the 2-stage regression thereby focuses on the relationships existing between speech features and depression only within localized regions along the depression axis. The second stage, which aggregates multiple rank scores, effectively allows for changes in these relationships along the axis to be accounted for in the prediction. Results gained from a comprehensive set of prediction tests presented in Cummins [2016] demonstrate the framework's suitability for depression score prediction: all 2-stage systems presented are shown to match or out-perform corresponding conventional regression systems. Work by Williamson et al. [2014] shows the benefits to using this kind of approach as part of a multimodal system. The continued exploration of two-stage prediction in all modalities is a promising area of future research.

8. Implementation-Related Considerations and Elicitation Approaches

Having reviewed recent developments in multimodal assessment of depression and related disorders based on behavioral signals, it is important to consider how data might be collected and what the deployment contexts might be for services or products based on this kind of system.

Example tasks that have been employed to elicit multimodal data from depressed individuals to date include speaking sustained vowels, diadochokinetic stimuli (e.g. "pa-ta-ka"), read speech, interview, virtual human interview (which can increase the comparability of behaviors between individuals) [DeVault et al., 2014], and mood induction. The choice of elicitation method to maximise for depression assessment sensitivity is in general an open research problem. Although research has shown that spontaneous speech (such as that elicited in an interview) is more characteristic of depression level [Alghowinem et al., 2012], there may be advantages for example in constraining the spoken content (since phonetic diversity is a source of nuisance variability), or in specifically assessing psychomotor skills, or in provoking facial responses to mood induction stimuli that are valuable for discrimination. It is also important to consider the context in which the data are collected, as affective expressions may differ as a result. Intra-personal tasks (e.g. responding to stimuli on a computer screen) provide a different context than inter-personal tasks (e.g. interviews, couples or family interactions). Work presented in Cummins et al. [2015b] suggests that the data-collection paradigm affects the strength of the manifestation of depression in speech samples; the authors speculate that tasks which invoke strong emotional or motivational effects in the reader will result in acoustically richer (in terms of depression information) samples.

Similarly, choosing or summarising salient segments from a recording is in general an open research problem: some studies have suggested that as little as the first 10s of response is just as effective for analysis as many minutes of data. Alternatively, methods such BoW, clustering or statistical functionals can be employed to deal with different recording lengths.

Some examples of multimodal corpora that have appeared in multiple publications to date include (most other similar depression datasets are not publicly available):

- The Distress Assessment Interview Corpus (DAIC) comprises audio and video recordings of spontaneous conversations from over 500 US English speakers during semi-structured interviews with a virtual human interviewer [Gratch, et al., 2014] designed to simulate standard protocols for identifying people at risk for major depression or PTSD and to elicit nonverbal and verbal behavior indicative of psychological distress. Each conversation lasts for around 20 minutes on average. Binary and severity labels were recorded for depression and PTSD using (self-assessed) questionnaires PHQ-9 and PCL-C. A subset of this data has been made publically available as part of the AVEC 2016 depression sub challenge [Valstar et al 2016][1].

- The Audio-Visual Emotion Challenges (AVEC) of 2013 and 2014[2] used subsets of the AViD corpus, which comprises 292 audio and video recordings of German speakers completing tasks including sustained vowels, counting, reading, and spontaneous story-telling. On average, each recording (of all tasks) lasts around 25 minutes. One (self-assessed) BDI-II scale depression rating was made for each speaker, and continuously-rated arousal and valence annotation was also developed by a team of annotators.
- The Black Dog database comprises recordings of audio and video from 40 speakers with a major depressive disorder, rated using the QIDS-SR scale by clinic psychiatrists at the Black Dog Institute, and 40 age-matched controls. Participants completed video viewing, image rating [McIntyre et al., 2009].
- The Pittsburgh database contains audio and video (head and shoulders, and full body view) recordings of semi-structured clinical interviews of 57 depressed patients, who were each evaluated four times at 7-week intervals using the HAM-D scale.

Some of the likely use-case contexts for systems providing multimodal assessment of depression and related disorders include:

- Clinic-based: This is a controlled environment in which more elaborate, sensitive and/or expensive sensor configurations can be employed. This environment is also more conducive to a human or virtual human interviewer.
- PC/laptop-based: This may be suitable for passive, ongoing monitoring, for example in a home or office environment.
- Smartphone-based and wearable-based: This is the least controlled environment in general, in terms of background audio, lighting, movement and other types of noise. However it is perhaps the richest in terms of assessing behavior, especially where the aim is to detect 'triggers' for depressive episodes, or to perform quick checks at regular periods after taking medication, for example.

The growth in the number of systems based on smartphones or wearable devices has been notable in recent years, and in turn this has spurred a good deal of commercial activity, as noted by Glenn and Monteith [2014]. Opportunities possibly afforded by these devices could include identifying situational, pharmacological, behavioral or cyclical triggers that deepen depression in individuals, or interventions where depression is detected [Donker et al., 2013].

As implied in Section 7, the type of output a system produced by the system could be categorical (e.g. "non-depressed", "depressed"), ordinal (e.g. "mild", "moderate", "severe", as per many scales), or continuous (e.g. predicting a numerical score). At present there is no consensus as to which may be preferred; it is likely that this will emerge as clinical trials of automatic systems of this kind are conducted. For systems designed to monitor patients who have already been diagnosed, the task design and system output type will need to be carefully selected according to the clinical objective, for example identifying responders to a particular treatment (e.g. 50% or more reduction in initial symptoms). It would also be of interest to find indicators of depression that are not sensitive to treatments, since these may act as trait-like indicators of susceptibility to depression. Similarly to how tests such as the Complete Blood Test, which is used by doctors to get an overview of a patient's overall physical health, a multimodal system could be used to review a patient's overall mental health.

[1] AVEC 2016 data available from http://sspnet.eu/avec2016/.

[2] AVEC 2013 and AVEC 2014 are available for research purposes http://sspnet.eu/avec2014/

Instead of providing a single score or diagnosis the objective analysis of behavioral changes could provide clinicians and medical practitioners with a rich battery of information on a patient's emotional, cognitive or mental state. The output of such a test could be used to support diagnosis of a range mental and neurological disorders.

9. Conclusion and Current Challenges

This chapter has given an overview of the role multimodal systems can play in the assessment of depression, with a particular emphasis on behavioral signals, together with an introduction to some of the key methods proposed and studies conducted to date. It seems clear that depression is in general very challenging to assess by any means using any one scale, due to its broad and heterogeneous clinical profile. This challenge, together with the compelling societal need for accessible and accurate assessment, seems a perfect motivation for research into automatic and multimodal systems.

Readers may note some similarities between this chapter and others that touch on affective computing, for example the small databases (relative to the ‘big data’ of multimedia research), the lack of definitive feature set (especially that is specific to the attribute of interest) and the significant signal variability posed by individual differences and environmental factors. However it is also worth noting some of the differences between depression assessment and much other affective computing research: depression is a long-term rather than short-term state; databases may often only contain one depression level example per individual (emotion databases typically have several emotions per individual); a particular depression score could possibly be due to many different combinations of symptoms; and depression databases will usually have only one labeller (inter-labeller agreement is unknown).

The approaches reviewed here all take a categorical view of depression. That is, that depression is a qualitatively different disorder from GAD, PTSD, and other clinical disorders, and that the goal of multimodal assessment is to contribute to diagnosis of depression. As noted previously, the assumption that depression is qualitatively unique among psychopathologies has been challenged. Ample evidence suggests that depression often is comorbid with GAD and depression symptoms often are components of other disorders, especially PTSD. These and similar findings have motivated dimensional conceptualizations of psychopathology. Theory and some data suggest that GAD, depression (MDD), and PTSD, as well as Dysthymia and Borderline Personality Disorder, are better understood as ‘distress’ disorders. Multimodal assessment can inform this debate by including a broader range of psychopathology in our research. Scherer and his colleagues [2014] have made initial advances in this direction. They include both depression and PTSD in their research. It is hoped that future work will continue to broaden patient characteristics in order to inform conceptual advances in psychopathology.

Future research challenges in this area abound, and some can be highlighted in particular:

- Since depression is measured on a numerical scale, prediction is an important problem. However to date there are few methods for fusing predictors that have been shown in this context to provide an improvement over the individual predictors - these will be essential to multimodal prediction.
- Relatively little is known about the generalizability of findings with respect to individual differences in participants or countries of origin. Initial work suggests that computer vision

based methods are robust to skin color [Girard, Cohn, Jeni, Sayette, and De la Torre, 2014], for instance, but replication is needed. Similarly, the extent to which findings generalize across different cultures [Alghowinem et al., 2016] is not yet well understood.

- Most work has focused on single modalities, such as face or voice. Greater attention to multimodal approaches is needed, for example the contribution of unique variance to detection of depression [Dibeklioglu, Hammal, and Cohn, 2017, in press] by different modalities. Further studies investigating the relative contributions of different features across different modalities are needed: these may be either statistical in nature (e.g. percentage of variance accounted for by major predictors) or machine learning-based empirical comparisons (e.g. classification accuracy of individual features vs. fused features/systems).
- Although features from various different modes have been shown to be *sensitive* to depression, there is virtually no research on whether they are *specific* to depression. This is an issue both because behavioral indicators tend to also be sensitive to other related disorders, and because other disorders (e.g. anxiety) often co-occur with depression. Alternatively, as argued above, lack of specificity may be inherent in any measure of depression to the extent that dimensional models of psychopathology prove more valid than traditional diagnostic categories.
- Because of the wide clinical profile of depression and the co-occurrence of other disorders with depression, there is interest in assessing sub-symptoms of depression rather depression directly. Work of this kind is already underway, for example assessment of psychomotor coordination [Trevino et al., 2011], of cognitive function, and of responses to negative-valence emotional stimuli [McIntyre et al., 2011]. However more work is needed, and consensus is yet to be reached even on which sub-symptoms should be assessed.
- Multimodal depression assessment to date has employed the most straightforward combination of modes (i.e. feature- or score-level fusion). There may be significant opportunities for features that exploit more complex multimodal information, for example temporal patterns across two or more modes, for example as captured by VTC features.
- Mitigation of confounding factors, such as acoustic variability due to speaker characteristics or phonetic content or variability in pose or illumination, is a key issue remaining to be satisfactorily resolved, although some advances have been made in this area [Williamson et al., 2013; Cummins et al., 2014].
- Approaches are needed for mitigating the irregular, non-linear relationships between changes in behavior features and changes in depression status or sub-symptom status.
- More study of longitudinal multimodal data in depressed individual is needed, to determine how effective within-user automatic analysis of depression may be as the extent of depression symptoms vary over time.
- Building the evidence base for using automatically-extracted behavioral signals as the basis for new diagnostic criteria.
- Bridging the ongoing need for close collaboration with clinicians

Table of Terms

Depression refers broadly to the persistence over an extended period of time of several of the following symptoms: lowered mood, interest or pleasure, psychomotor retardation, psychomotor agitation, diminished ability to think/concentrate, increased indecisiveness, fatigue or loss of energy, insomnia, hypersomnia, significant weight loss or weight gain, feelings of worthlessness or excessive guilt, recurrent thoughts of death or recurrent suicidal ideation. It is important to note that there are multiple definitions of depression (see references in Section 1).

Psychomotor retardation refers to a slowing of thought and reduction of physical movements, including cognitive, emotional and physical responses to stimuli. A symptom of depression, it is often explicitly measured as part of clinical depression assessment scales.

Hand-crafted features refer to features developed to extract a specific type of information, usually as part of a hypothesis-driven research study. By contrast, data-driven features are those extracted automatically from raw signal data by algorithms [e.g. neural networks], whose physical interpretation often cannot easily be described.

Spatio-temporal features are those which have both a spatial and a time dimension. For example, the intensity of pixels in a video vary both in terms of their position within a given frame (spatial dimension) and in terms of the frame number for a given pixel coordinate (temporal dimension).

Bag-of-words (BoW) is a data-driven algorithm for summarising large volumes of features. It can be thought of as a histogram whose bins are determined by partitions (or clusters) of the feature space.

Voice quality refers to the type of phonation during voiced speech. Depending on the physical movement of the vocal folds during phonation, the perceived quality of speech can change, even for the same speech sound uttered at the same pitch. Descriptors such as 'creaky' and 'breathy' are applied to specific modes of vocal fold vibration.

Formants (especially F1, F2) are broad peaks in the amplitude spectrum of speech due to the acoustic resonances of the vocal tract: F1 (in Hz) is the lowest frequency peak and F2 is the next highest frequency peak. Unrelated to the pitch of speech, formants are controlled by the vocal tract shape, and allow discrimination between different speech sounds (e.g. vowel /a/ has a higher F1 and lower F2 than /e/).

Overfitting is a problem that occurs when the training or estimation of a machine learning method is performed on data with too few training examples relative to the number of parameters to be estimated. The resulting problem is that the method becomes too closely tuned to the training data, and generalises poorly to unseen test data.

Longitudinal data refers to multiple recordings of the same type from the same individual at different points in time, between which it is likely that the individual's state (e.g. depression score) has changed.

Specificity is the ability of a test or indicator to correctly identify users without a particular condition (e.g. depression), in the context of a two-class detection problem. In affective computing research, many features sensitive to a particular attribute of interest (e.g. depression) have been reported, however these same features may also be sensitive to other attributes (e.g. fatigue or intoxication), and could hence be considered to have weak specificity.

Elicitation refers to the means by which individual users are encouraged to produce behavior which is then acquired and processed by a multimodal system. In the general use context of multimodal systems, there is not usually the opportunity to control the user's behavior to any great extent, however in a more clinical context such as depression assessment, users may implicitly or explicitly agree to a multimodal system dictating quite precisely what behavior is elicited from them.

Focus Questions

1. Why might analysis of behavior be a promising approach for assessing depression ? In particular, why can a multimodal approach be expected to be more successful than unimodal approaches ?
2. Discuss the benefits and limitations of three different methods for eliciting behavior for automatic assessment of depression, and the possible implications for interface design.
3. Choose five different kinds of features from this chapter and for each explain what aspect of depression they characterize, and the system requirements to implement them.
4. Choose one feature from this chapter, and based on further reading, present a detailed explanation of the mathematical/computational approach, explaining all symbols, why the feature takes the mathematical form it has, and what other features [if any] might capture similar information about depressed behavior.
5. Bag-of-words has proven useful in conjunction with a variety of different features in the depression assessment context. What are some reasons why ?
6. For some types of features, toolboxes are available that allow extraction of a wide variety of features. What are some important considerations when choosing which toolboxes and/or features to employ ? What is a limitation of large feature sets ?
7. For a particular modality of your choice, what are (at least) two sources of variability that are undesirable from the perspective of depression assessment ? Based on further reading, what kinds of approaches might help mitigate them ?
8. Compare and contrast two different approaches for fusing automatic depression assessment systems. Based on further reading, has one been found more effective than the other ?
9. What are the desirable attributes of research datasets comprising behavior of depressed individuals ? List as many as possible.
10. Considering use-cases of a multimodal automatic depression assessment system, what are some of the benefits and limitations of the different possible contexts ? Consider in particular (i) in a clinic, (ii) at home and (iii) in a mobile, everyday context.

References

Abel, L.A., Friedman, L., Jesberger, J., Malki, A., and Meltzer, H.Y., 1991. "Quantitative assessment of smooth pursuit gain and catch-up saccades in schizophrenia and affective disorders", *Biol. Psychiatr*y 29, 1063–1072.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., and Parker, G., 2012, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech", in *Proc. Int. FLAIRS Conf. AAAI*, Marco Island, FL, USA, pp. 141–146.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M., 2013a, "Eye movement analysis for depression detection", in *Proc. IEEE Int. Conf. on Image Processing,* pp. 4220-4224.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., Breakspear, M., 2013b, "Head pose and movement analysis as an indicator of depression", in *Proc. Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 283-288.

Alghowinem, S., Goecke, R., Wagner, M., and Epps, J., 2013c, "A Comparative of Different Classifiers for Detecting Depression from Spontaneous Speech", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 8022 – 8026.

Alghowinem, S., Goecke, R., Epps, J., Wagner, M., Cohn, J. F., 2016, "Cross-Cultural Depression Recognition from Vocal Biomarkers", in *Proc. INTERSPEECH*, pp. 1943-1947.

Alpert, M., Pouget, E.R., and Silva, R.R., 2001, "Reflections of depression in acoustic measures of the patient's speech", *J. Affect. Disord.*, 66, 59–69.

American-Psychiatric-Association, 2015. *Diagnostic and statistical manual of mental disorders : DSM-V.*, 5th ed. American Psychiatric Association, Washington, DC.

Asthana, A., Khwaja, A. and Goecke, R., "Automatic Frontal Face Annotation and AAM Building for Arbitrary Expressions from a Single Frontal Image Only", in *Proc. IEEE Int. Conf. on Image Processing*, pp. 2445-2448, Cairo, Egypt.

Asthana, A., Lucey, S., and Goecke, R., "Regression based automatic face annotation for deformable model building", *Pattern Recognition*, 44(10-11): 2598–2613, Oct.-Nov. 2011.

Baker, S., Matthews, I., and Schneider, J., 2004, "Automatic Construction of Active Appearance Models as an Image Coding Problems", *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Baltrusaitis, T., Robinson, P., and Morency, L.-P., 2012, "3D constrained local model for rigid and non-rigid facial tracking", in *Proc. IEEE Computer Vision and Pattern Recognition*, Providence, RI.

Beck, A.T., Steer, R.A., Ball, R., and Ranieri, W.F., 1996, "Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients", *J. Pers. Assess.*, 67, 588–597.

Blair-West, G. W., Cantor, C. H., Mellsop, G. W., and Eyeson-Annan, M. L., 1999, "Lifetime suicide risk in major depression: sex and age determinants", *J. Affective Disorders*, 55(2-3), 171-178.

Blais, M., and Baer, L., 2010. "Understanding Rating Scales and Assessment Instruments", in Baer, L., Blais, M.A. (Eds.), *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health, Current Clinical Psychiatry*. Humana Press, New York, NY.

Burkhouse, K. L., Siegle, G. J., Woody, M. L., Kudinova, A. Y., and Gibb, B. E., 2015, "Pupillary reactivity to sad stimuli as a biomarker of depression risk: Evidence from a prospective study of children", *Journal of Abnormal Psychology*, 124(3), 498.

Bylsma, L. M., Morris, B. H., and Rottenberg, J., 2008, "A meta-analysis of emotional reactivity in major depressive disorder", *Clinical Psychology Review*, 28(4), 676-691.

Calkins, M. E., Iacono, W. G., and Ones, D. S., 2008, “Eye movement dysfunction in first-degree relatives of patients with schizophrenia: A meta-analytic evaluation of candidate endophenotypes”, *Brain and Cognition*, 68, 436-461.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E., “Predicting Depression via Social Media”, in *Proc. AAAI Conf. on Web and Social Media*, 2013.

Campbell, S. B., & Cohn, J. F., 1991, “Prevalence and correlates of postpartum depression in first-time mothers”, *Journal of Abnormal Psychology, 100*(4), 594-599.

Carroll, B. J., 1984, “Dexamethasone suppression test”, In R. C. W. Hall & T. P. Beresford (Eds.), *Handbook of psychiatric diagnosis* (Vol. 1, pp. 3-28). Netherlands: Springer.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E., “Predicting Depression via Social Media”, in *Proc. AAAI Conf. on Web and Social Media*, 2013.

Chu, W.-S., De la Torre, F., and Cohn, J. F., 2016, “Selective transfer machine for personalized facial action unit detection”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 3515 – 3522.

Cohn, J. F., Ambadar, Z., and Ekman, P., 2007, “Observer-based measurement of facial expression with the Facial Action Coding System”, in Coan, J. A. and Allen, J. J. B. (Eds.), *Handbook of emotion elicitation and assessment*. Oxford University Press Series in Affective Science (pp. 203-221). New York, NY: Oxford University.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F., “Detecting depression from facial actions and vocal prosody”. In *Proc. Int. Conf. on Affective Computing and Intelligent Interaction,* 2009, pp. 1-7.

Cohn, J. F., and De la Torre, F., 2015, “Automated face analysis for affective computing”, in Calvo, R. A., D'Mello, S. K., Gratch, J., and Kappas, A. (Eds.), *Handbook of affective computing* [pp. 131-150]. New York, NY: Oxford.

Cootes, T. F., Wheeler, G. V., Walker, K. N., and Taylor, C. J., 2002, “View-based active appearance models”, *Image and Vision Computing*, 20[9-10], 657-664.

Corneanu, C., Oliu, M., Cohn, J. F., and Escalera, S., 2015, “Survey on RGB, thermal, and multimodal approaches for facial expression analysis: History, trends, and affect-related applications”, *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Cummins, N., 2016. *Automatic assessment of depression from speech: paralinguistic analysis, modelling and machine learning*. PhD Thesis, UNSW Australia.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R., 2011, “An Investigation of Depressed Speech Detection: Features and Normalization”, in *Proc. INTERSPEECH*, Florence, Italy, pp. 2997–3000.

Cummins, N., Epps, J., Sethu, V., Breakspear, M., and Goecke, R., 2013a, “Modeling Spectral Variability for the Classification of Depressed Speech”, in *Proc. INTERSPEECH*, Lyon, France, pp. 857–861.

Cummins, N., Epps, J., Sethu, V., and Krajewski, J., 2014, “Variability Compensation in Small Data: Oversampled Extraction of i-vectors for the Classification of Depressed Speech”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Florence, Italy, pp. 970–974.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R. and Epps, J., 2013, “Diagnosis of Depression by Behavioral Signals: A Multimodal Approach”, in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC’13)*, Barcelona, Spain, pp. 11–20.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F., ″A Review of Depression and Suicide Risk Assessment using Speech Analysis″, *Speech Communication*, vol. 71, July 2015a, pp. 10-49.

Cummins, N., Sethu, V., Epps, J., and Krajewski, J., 2015b, “Relevance Vector Machine for Depression Prediction”, in *Proc. INTERSPEECH*, Dresden, Germany, pp. 110–114.

Cummins, N., Sethu, V., Epps, J., Schnieder, S., and Krajewski, J., 2015c, “Analysis of acoustic space variability in speech affected by depression”, *Speech Commun.*, 75, 27–49.

Cuthbert, B.N. and Insel, T.R., 2013, “Toward the future of psychiatric diagnosis: the seven pillars of RDoC”, *BMC Med.*, 11, 126.

Dawson, Michael E., Schell, A. M., Braaten, J. R., and Catania, J. J., “Diagnostic utility of autonomic measures for major depressive disorders”, *Psychiatry Research*, 15.4, 1985, 261-270.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgilia, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.P., “Simsensei kiosk: A virtual human interviewer for healthcare decision support,” in *Proc. Autonomous Agents and Multiagent Systems*, 2014, pp. 1061–1068.

Dibeklioglu, H., Hammal, Z., Yang, Y., and Cohn, J. F., 2015, “Multimodal detection of depression in clinical interviews”, in *Proc. ACM Int. Conf. on Multimodal Interaction*, Seattle, WA.

Dibeklioglu, H., Hammal, Z., and Cohn, J. F., 2017, in press, “Multimodal measurement of depression severity in the context of clinical interviews”, *IEEE Journal of Biomedical Health Informatics*.

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., and Christensen, H., 2013, “Smartphones for smarter delivery of mental health programs: a systematic review”, *J. Medical Internet Research*, 15, no. 11.

Eaton, W., Muntaner, C., Smith, C., Tien, A., and M., Y., 2004, “Center for Epidemiologic Studies Depression Scale: Review and revision (CESD and CESD-R)”, In M. E. Maruish (Ed.), *The Use of*

Psychological Testing for Treatment Planning and Outcomes Assessment, Lawrence Erlbaum, pp. 363-377, Mahawah, NJ: Lawrence Erlbaum.

Ekman, P., and Friesen, W. V., 1969, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding" *Semiotica*, 1(1), 49-98.

Ekman, P., Friesen, W. V., and Hager, J. C., 2002, *Facial action coding system*: Research Nexus, Network Research Information, Salt Lake City, UT

Fairbanks, L. A., McGuire, M. T., and Harris, C. J., 1982, "Nonverbal interaction of patients and therapists during psychiatric interviews", *Journal of Abnormal Psychology*, 91(2), 109.

First, M. B., Williams, J. B. W., Karg, R. S., and Spitzer, R. L., 2015, "Structured Clinical Interview for DSM-5—Research Version", Arlington, VA: American Psychiatric Association.

Flint, A.J., Black, S.E., Campbell-Taylor, I., Gailey, G.F., and Levinton, C., 1993, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression", *J. Psychiatr. Res.*, 27, 309–319.

Forgas, J. P., 2014, "Four ways sadness may be good for you: The Greater Good Science Center", University of California Berkeley.

France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., and Wilkes, M., 2000, "Acoustical properties of speech as indicators of depression and suicidal risk", *IEEE Trans. Biomed. Eng.*, 47, 829–837.

Ghosh, S., Laksana, E., Scherer, S., and Morency, L.-P., 2015, "A multi-label convolutional neural network approach to cross-domain action unit detection", in *Proc Int. Conf. on Affective Computing and Intelligent Interaction*, 609-615.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., and Kupfer, D. J., 2012, "Development of a computerized adaptive test for depression", *Archives of General Psychiatry*, 69(11), 1104-1112.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., and Rosenwald, D. P., 2013, "Social risk and depression: Evidence from manual and automatic facial expression analysis", in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 1-8.

Girard, J. M., Cohn, J. F., Jeni, L. A., Sayette, M. A., and De la Torre, F., 2014, "Spontaneous facial expression can be measured automatically", *Behavior Research Methods*, 47(4), 1136-1147.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P., 2014, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses", *Image and Vision Computing*, 32[10], 641-647.

Girard, J. M., and Cohn, J. F., 2015, "Automated audiovisual depression analysis", *Current Opinion in Psychology*, 4, 75-79.

Glenn, T., and Monteith, S, 2014, “New measures of mental state and behavior based on data collected from sensors, smartphones, and the Internet”, *Current Psychiatry Reports*, 16(12), 1-10.

Gotlib, I. H., & Hammen, C. L., 2002, *Handbook of Depression*. New York: Guilford.

Gratch, J., Artstein, R., Lucas, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, A., and Morency, L.- P., “The distress analysis interview corpus of human and computer interviews,” in *Proc. Language Resources and Evaluation Conference*, 2014, pp. 3123–3128.

Guo, C.C., Hyett, M.P., Nguyen, V.T., Parker, G., Breakspear, M., 2016, “Distinct neurobiological signatures of brain connectivity in depression subtypes during natural viewing of emotionally salient films”, *Psychological Medicine*, 46[07]: 1535 – 1545.

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., and Narayanan, S., 2014, “Multimodal prediction of affective dimensions and depression in human-computer interactions”, *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge [AVEC’14]*, Orlando, FL.

Hall, J.A., Harrigan, J.A., and Rosenthal, R., 1995, “Nonverbal behavior in clinician—patient interaction”, *Appl. Prev. Psychol*., 4, 21–37.

Hamilton, H., 1960, “HAMD: A rating scale for depression”, *Neurosurg. Psychiat*., 23, 56–62.

Helfer, B.S., Quatieri, T.F., Williamson, J.R., Mehta, D.D., Horwitz, R., and Yu, B., 2013, “Classification of depression state based on articulatory precision”, in *Proc. INTERSPEECH*, Lyon, France, pp. 2172–2176.

HiTOP, (Undated), “The hierarchical taxonomy of psychopathology [HiTOP]”, from http://medicine.stonybrookmedicine.edu/HITOP

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., and Krajewski, J., 2014, “Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender”, in *Proc. INTERSPEECH*, Singapore, pp. 1248–1252.

Hong, L., Turano, K., O'Neil, H., Hao, L., Wonodi, I., and McMahon, R., 2008, “Refining the predictive pursuit endophenotype in schizophrenia”, *Biological Psychiatry*, 63, 458-464.

Horwitz, R., Quatieri, T.F., Helfer, B.S., Yu, B., Williamson, J.R., and Mundt, J., 2013, “On the relative importance of vocal source, system, and prosody in human depression”, in *Proc. IEEE Int. Conf. on Body Sensor Networks*, Cambridge, MA, USA, pp. 1–6.

Ikonen, T., Haataja, K., and Toivanen, P., 2013, “3D imaging of human brain in the diagnostics of insomnia and depression: A comparative analysis, a novel insomnia/depression diagnostics approach, and lessons learned”, in *Proc. Int. Conf. on Intelligent Systems Design and Applications*, pp. 230–235.

Jeni, L. A., Cohn, J. F., and Kanade, T., 2016, “Dense 3D face alignment from 2d videos in real-time”, *Computer Vision and Image Understanding*.

Joshi, J., Dhall, A., Goecke, R., Breakspear, M., and Parker, G., 2012, "Neural-Net Classification for Spatio-Temporal Descriptor Based Depression Analysis", in *Proc. Int. Conf. on Pattern Recognition*, pp. 2634-2638, Tsukuba, Japan.

Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., and Breakspear, M., 2013a, "Multimodal Assistive Technologies for Depression Diagnosis and Monitoring", *Journal on Multimodal User Interfaces* Special Issue on Multimodal Interfaces for Pervasive Assistance, vol. 7, no. 3, pp. 217-228.

Joshi, J., Goecke, R., Breakspear, M., and Parker, G., 2013b, "Can body expressions contribute to automatic depression analysis ?", in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Shanghai, China.

Joshi, J., Dhall, A., Goecke, R., and Cohn, J. F., 2013c, "Relative body parts movement for automatic depression analysis", in *Proc. Conf. on Affective Computing and Intelligent Interaction*, pp. 492–497.

Kaya, H., Çilli, F., and Salah, A., 2014a, "Ensemble CCA for Continuous Emotion Prediction", in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge [AVEC'14]*, Orlando, FL, USA, pp. 19–26.

Kaya, H., Eyben, F., and Salah, A. A., 2014b, "CCA based Feature Selection with Application to Continuous Depression Recognition from Acoustic Speech Features", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Florence, Italy, pp. 3757–3761.

Kent, R.D., and Kim, Y.-J., 2003, "Toward an acoustic typology of motor speech disorders", *Clin. Linguist. Phon.*, 17, 427–445.

Kessler, R. C., Chiu, W. T., Demler, O., and Walters, E. E., 2005, "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication", *Archives of General Psychiatry*, 62, 617-627.

Kirsch, A., and Brunnhuber, S., 2007, "Facial expression and experience of emotions in psychodynamic interviews with patients with PTSD in comparison to healthy subjects", *Psychopathology*, 40(5), 296-302.

Kotov, R., Gamez, W., Schmidt, F., and Watson, D., 2010, "Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis", *Psychological Bulletin*, 136(5), 768-821.

Kraepelin, E., 1921, "Manic Depressive Insanity and Paranoia", *J. Nerv. Ment. Dis.*, 53, 350.

Kroenke, K., Spitzer, R.L., and Williams, J.B.W., 2001, "The PHQ-9", *J. Gen. Intern. Med.*, 16, 606–613.

Larsen, M.E., Cummins, N., Boonstra, T.W., O' Dea, B., Tighe, J., Shand, F., Epps, J., Christensen, H., 2015, "The Use of Technology in Suicide Prevention", in *Proc. IEEE Int. Conf. Engineering in Medicine and Biology Society*, Milan, Italy, pp. 7316–7319.

Levy, D. L., Holzman, P. S., Matthysse, S., and Mendell, N. R., 1993, “Eye tracking dysfunction and schizophrenia: A critical perspective”, *Schizophrenia Bulletin*, 19, 461-536.

Low, L.S.A., Maddage, N.C., Lech, M., Sheeber, L., and Allen, N., 2010, “Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Dallas, Texas, U.S.A, pp. 5154–5157.

Low, L.S.A., Maddage, M.C., Lech, M., Sheeber, L.B., and Allen, N.B., 2011., “Detection of Clinical Depression in Adolescents; Speech During Family Interactions”, *IEEE Trans. Biomed. Eng.*, 58, 574–586.

Mathers, C. D., and Loncar, D., 2011, “Projections of global mortality and burden of disease from 2002 to 2030”, *PLoS Medicine*, 3(11).

McIntyre, G., Göcke, R., Hyett, M., Green, M., and Breakspear, M., 2009, “An approach for automatically measuring facial activity in depressed subjects”, in *Proc. Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 1-8.

McIntyre, G., Göcke, R., Breakspear, M., and Parker, G., 2011, “Facial Response to Video Content in Depression”, in *Proc. ICMI Workshop on Inferring Cognitive and Emotional States from Multimodal Measures*.

McRae, P.A., Tjaden, K., and Schoonings, B., 2002, “Acoustic and perceptual consequences of articulatory rate change in Parkinson disease”, *J. Speech. Lang. Hear. Res.*, 45, 35–50.

Milne, D., Paris, C., Christensen, H., Batterham, P. and O'Dea, B., 2015, “We Feel: Taking the emotional pulse of the world”, in *Proc. Cong. Int. Ergonomics Association*, Melbourne, Australia.

Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., and Graciarena, M., 2014, “The SRI AVEC-2014 Evaluation System”, in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC’14)*, Orlando, Florida, USA, pp. 93–101.

Menke, R., 2011, *Examining nonverbal shame markers among post-pregnancy women with maltreatment histories*, Doctoral dissertation, Wayne State University.

Moore, E., Clements, M.A., Peifer, J.W., and Weisser, L., 2008, “Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech”, *IEEE Trans. Biom. Eng.*, 55, 96–107.

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

Mundt, J.C., Vogel, A.P., Feltner, D.E., and Lenderking, W.R., 2012, “Vocal Acoustic Biomarkers of Depression Severity and Treatment Response”, *Biol. Psychiatry*, 72, 580 –587.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., and Berk, M., 2014, “Affective and content analysis of online depression communities”, *IEEE Trans. Affective Computing*, 5(3):217–226.

Nguyen, Thin, O’Dea, B., Larsen, M., Phung, D., Venkatesh, S., Christensen, H., 2015, “Using linguistic and topic analysis to classify sub-groups of online depression communities”, *Multimedia Tools and Applications*, 1-24.

Nilsonne, A., Sundberg, J., Ternstrom, S., and Askenfelt, A., 1987, “Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression”, *J. Acoust. Soc. Am.*, 83, 716–728.

O'Hara, M. W., Zekoski, E. M., Philipps, L. H., & Wright, E. J., 1990, “Controlled Prospective Study of Postpartum Mood Disorders: Comparison of Childbearing and Nonchildbearing Women”, *Journal of Abnormal Psychology, 99*(1), 3-15.

Ooi, K.E.B., Lech, M., and Allen, N.B., 2013, “Multichannel weighted speech classification system for prediction of major depression in adolescents”, *IEEE Trans. Biomed. Eng*., 60, 497–506.

Orozco-Arroyave, J.R., Belalcázar-Bolaños, E.A., Arias-Londoño, J.D., Vargas-Bonilla, J.F., Haderlein, T., and Nöth, E., 2014, “Phonation and Articulation Analysis of Spanish Vowels for Automatic Detection of Parkinson’s Disease”, in Sojka, P., Horák, A., Kopeček, I., and Pala, K. [Eds.], *Text, Speech and Dialogue*, Lecture Notes in Computer Science. Springer International Publishing, Cham, Switzerland, pp. 374–381.

Østergaard, S.D., Jensen, S.O.W., and Bech, P., 2011, “The heterogeneity of the depressive syndrome: when numbers get serious”, *Acta Psychiatr. Scand*., 124, 495–496.

Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., and Wilkes, D.M., 2004, “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk”, *IEEE Trans. Biomed. Eng*., 51, 1530–1540.

Pachucki, M. C., Ozer, E. J., Barrat, A., and Cattuto, C., 2015, “Mental health and social networks in early adolescence: a dynamic study of objectively-measured social interaction behaviors”, *Soc Sci Med*, vol. 125, pp. 40-50.

Park, M., Cha, C., and Cha, M., 2012, “Depressive moods of users portrayed in twitter”, in *Proc. ACM SIGKDD Workshop on Healthcare Informatics*, pp. 1–8.

Peng, H., Hu, B., Liu, Q., Dong, Q., Zhao, Q., and Moore, P., 2011, “User centered depression prevention: An EEG approach to pervasive healthcare”, in *Proc. IEEE PervasiveHealth*, pp. 325–330.

Perez, J. E., and Riggio, R. E., 2003, “Nonverbal social skills and psychopathology”, *Nonverbal Behavior in Clinical Settings*, 17-44.

Pérez, H., Escalante, H.J., Villaseñor-Pineda, L., Montes-y-Gómez, M., Pinto-Avedaño, D., and Reyes-Meza, V., 2014, “Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition”, in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC’14)*, Orlando, Florida, USA, pp. 49–55.

Proudfoot, Judith, Clarke, J., Birch, M.-R., Whitton, A. E., Parker, G., Manicavasagar, V., Harrison, V., Christensen, H., and Hadzi-Pavlovic, D., 2013, “Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial”, *BMC Psychiatry* 13.1, 1.

Quatieri, T.F., and Malyska, N., 2012, “Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity”, in *Proc. INTERSPEECH*, Portland, OR, pp. 1059–1062.

Radloff, L. S., 1977, “The CES-D scale: A self-report depression scale for research in the general population”, *Applied Psychological Measurement, 1*, 385-401.

Reed, L. I., Sayette, M. A., and Cohn, J. F., 2007, “Impact of depression on response to comedy: A dynamic facial coding analysis”, *Journal of Abnormal Psychology, 116*(4), 804-809.

Rosenquist, J. N., Fowler, J. H., and Christakis, N. A., “Social network determinants of depression”, *Mol Psychiatry*, vol. 16, pp. 273-81, Mar 2011

Sapir, S., Ramig, L.O., Spielman, J.L., and Fox, C., 2010, “Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech”, *J. Speech. Lang. Hear. Res*., 53, 114–25.

Saragih, J. M., Lucey, S., and Cohn, J. F., 2009, “Subspace constrained mean-shift”, in *Proc. IEEE Int. Conf. on Computer Vision*, Kyoto, Japan.

Schelde, J. T. M., 1998, “Major depression: Behavioral markers of depression and recovery”, *The Journal of Nervous and Mental Disease*, 186(3), 133-140.

Scherer, S., Stratou, G., Gratch, J., and Morency, L.-P., 2013a, “Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD”, in *Proc. INTERSPEECH*, Lyon, France, pp. 847–851.

Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., and Gratch, J., 2013b, “Automatic Behavior Descriptors for Psychological Disorder Analysis”, in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Shanghai, China, pp. 1 – 8.

Scherer, S., Stratou, G., and Morency, L.P., 2013c, “Audiovisual Behavior Descriptors for Depression Assessment”, in *Proc. ACM Int. Conf. on Multimodal Interaction*, Sydney, Australia, pp. 135–140.

Scherer, S., Stratou, G., Lucus, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A. (Skip), Morency, L.-P., and Lucas, G., 2014, “Automatic Audiovisual Behavior Descriptors for Psychological Disorder Analysis”, *Image Vis. Comput*. 32, 1–21.

Scherer, S., Morency, L.P., Gratch, J., and Pestian, J., 2015, “Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Brisbane, Australia, pp. 4789–4793.

Scherer, S., Lucas, G., Gratch, J., Rizzo, A., and Morency, L.-P., “Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews”, *IEEE Trans. Affective Computing*, vol. 7, no. 1, 2016, pp. 59-73.

Scherer, S., G. M. Lucas, J. Gratch, A. Skip Rizzo and L. P. Morency, 2016, “Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews”, in *IEEE Trans. Affective Computing*, vol. 7, no. 1, pp. 59-73, 2016.

Schueller, Stephen M., Begale, M., Penedo, F. J., and Mohr, D. C., 2014, “Purple: A modular system for developing and deploying behavioral intervention technologies”, *Journal of Medical Internet Research* 16.7, e181.

Senoussaoui, M., Sarria-Paja, M., Santos, J.F., and Falk, T.H., 2014, “Model Fusion for Multimodal Depression Classification and Level Detection”, in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC’14)*, Orlando, Florida, USA, pp. 57–63.

Sheehan, D. V., Lecrubier, Y., Harnett-Sheehan, K., Janavs, J., Weiller, E., Bonara, L., . . . Dunbar, G. C., 1997, “Reliability and validity of the M.I.N.I. International Neuropsychiatric Interview (M.I.N.I.)”, According to the SCID-P, *European Psychiatry*, 12, 232-241.

Sidorov, M., and Minker, W., 2014, “Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features: A Multimodal Approach”, in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC’14)*, Orlando, Florida, USA, pp. 81–86.

Siegle, G. J., Granholm, E., Ingram, R. E., and Matt, G. E., 2001, “Pupillary and reaction time measures of sustained processing of negative information in depression”, *Biological Psychiatry*, 49.7: 624-636.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., and Carter, C. S., 2002, “Can’t shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals”, *Biological Psychiatry*, 51[9]:693–707.

Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., and Thase, M. E., 2003, “Do the seconds turn into hours? Relationships between sustained pupil dilation in response to emotional information and self-reported rumination”, *Cognitive Therapy and Research*, 27(3), 365-382.

Skodda, S., Grönheit, W., Schlegel, U., 2012, “Impairment of vowel articulation as a possible marker of disease progression in Parkinson’s disease”, *PLoS One*, 7, e32132.

Sobin, C., Sackeim, H.A., 1997, “Psychomotor symptoms of depression”, *Am. J. Psychiatry*, 154, 4–17.

Stassen, H. H., Albers, M., Püschel, J., Scharfetter, C., Tewesmeier, M., and Woggon, B., 1995, “Speaking behavior and voice sound characteristics associated with negative schizophrenia”, *J. Psychiatr. Res.*, 29, 277–296.

Stassen, H. H., Bomben, G., and Gunther, E., 1991, “Speech characteristics in depression”, *Psychopathology*, 24, 88–105.

Sturim, D., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., McCree, A., 2011, "Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis", in *Proc. INTERSPEECH*, Florence, Italy, pp. 2983–2986.

Tolkmitt, F., Helfrich, H., Standke, R., and Scherer, K. R., 1982, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics", *J. Commun. Disord.*, 15, 209–222.

Trevino, A., Quatieri, T., and Malyska, N., 2011, "Phonologically-based biomarkers for major depressive disorder", *EURASIP J. Adv. Signal Process*., 1–18.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M., 2013, "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge", in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC'13)*, pp. 3–10.

Valstar, M., 2014a, "Automatic behavior understanding in medicine", in *Proc. ACM Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, 57-60.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M., 2014b, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge", *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC'14)*, 3–10.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M., 2016, "AVEC 2016-Depression, Mood, and Emotion Recognition Workshop and Challenge", in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge [AVEC'16]*, 3–10.

Watson, D., and Clark, L. A., 2006, "Distress and fear disorders: An alternative empirically based taxonomy of the 'mood'and 'anxiety'disorders", *British Journal of Psychiatry*, 189(6), 481-483. doi: 10.1192/bjp.bp.106.03825

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. A., . . . Stuart, S., 2007, "Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS)", *Psychological Assesment*, 9(3), 253-268.

Waxer, P., 1974, "Nonverbal cues for depression", *Journal of Abnormal Psychology*, 83[3], 319.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., and Mehta, D. D., 2013, "Vocal Biomarkers of Depression Based on Motor Incoordination", in *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC'13)*. Barcelona, Spain, pp. 41–48.

Williamson, J., Quatieri, T., Helfer, B., Ciccarelli, G., and Mehta, D. D., 2014, "Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing", *Proc. ACM Int. Workshop on Audio/Visual Emotion Challenge (AVEC'14)*, Orlando, FL, pp. 65–72.

World Health Organization, 1993, "The ICD-10 classification of mental and behavioral disorders: Diagnostic criteria for research", Geneva, Switzerland: World Health Organization.

World Health Organization, 2014, “Global estimates 2014 summary tables”, Geneva, Switzerland: World Health Organization.

WW, E., C, M., C, S., Tien, A., & M., Y., 2004, “Center for Epidemiologic Studies Depression Scale: Review and revision (CESD and CESD-R)”, In M. E. Maruish (Ed.), *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment. 3rd ed. Mahwah, NJ: Lawrence Erlbaum; 2004:363-377.*

Xiong, X., and De la Torre, F., 2013, “Supervised descent method and its applications to face alignment”, in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Portland, OR.

Yang, Y., and Ramanan, D., “Articulated pose estimation with flexible mixtures-of-parts”, in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1385–1392.

Yang, Y., Fairbairn, C., and Cohn, J., 2012, “Detecting Depression Severity from Vocal Prosody”, *IEEE Trans. Affect. Comput.*, 4, 142–150.

Yingthawornsuk, T., Keskinpala, H. K., France, D., Wilkes, D. M., Shiavi, R. G., Salomon, R.M., 2006, “Objective Estimation of Suicidal Risk using Vocal Output Characteristics”, in *Proc. INTERSPEECH*, Pittsburgh, PA, pp. 649–652.

Zhao, Q., Sun, J., Cong, F., Chen, S., Tang, Y., and Tong, S., 2013, “Multi-domain feature analysis for depression: A study of N170 in time, frequency and spatial domains simultaneously”, in *Proc IEEE Int. Conf. Engineering in Medicine and Biology Society*, pp. 5986–5989.

Zimmerman, M., Chelminski, I., & Posternak, M., 2004, “A review of studies of the Hamilton depression rating scale in healthy controls: Implications for the definition of remission in treatment studies of depression”, *Journal of Nervous & Mental Disease, 192*(9), 595-601.