Image and Vision Computing 32 (2014) 648–658

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Automatic audiovisual behavior descriptors for psychological disorder analysis☆

Stefan Scherer [a,*], Giota Stratou [a], Gale Lucas [a], Marwa Mahmoud [a,b], Jill Boberg [a], Jonathan Gratch [a], Albert (Skip) Rizzo [a], Louis-Philippe Morency [a]

[a] University of Southern California Institute for Creative Technologies, Playa Vista, CA 90094, USA
[b] University of Cambridge, UK

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form 25 April 2014
Accepted 13 June 2014
Available online 20 June 2014

Keywords:
Psychological distress
Depression
Post-traumatic stress disorder
Anxiety
Nonverbal behavior
Automatic assessment
Audiovisual

ABSTRACT

We investigate the capabilities of automatic audiovisual nonverbal behavior descriptors to identify indicators of psychological disorders such as depression, anxiety, and post-traumatic stress disorder. Due to strong correlations between these disordersas measured with standard self-assessment questionnaires in this study, we focus our investigations in particular on a generic distress measure as identified using factor analysis. Within this work, we seek to confirm and enrich present state of the art, predominantly based on qualitative manual annotations, with automatic quantitative behavior descriptors. We propose a number of nonverbal behavior descriptors that can be automatically estimated from audiovisual signals. Such automatic behavior descriptors could be used to support healthcare providers with quantified and objective observations that could ultimately improve clinical assessment. We evaluate our work on the dataset called the Distress Assessment Interview Corpus (DAIC) which comprises dyadic interactions between a confederate interviewer and a paid participant. Our evaluation on this dataset shows correlation of our automatic behavior descriptors with the derived general distress measure. Our analysis also includes a deeper study of self-adaptor and fidgeting behaviors based on detailed annotations of where these behaviors occur.

1. Introduction

The recent progress in facial feature tracking and articulated body tracking [1–3] has opened the door to new applications for automatic nonverbal behavior analysis.[1]One promising direction for this technology is the medical domain where computer vision algorithms can assist clinicians and health care providers in their daily activities. For example, these new perceptual softwares can assist doctors during remote telemedicine sessions that lack the communication cues provided in face-to-face interactions. Automatic behavior descriptors can further add quantitative information to the interactions such as behavior dynamics and intensities. These quantitative data can improve both post-session and online analysis. Proper sensing of nonverbal cues can also provide support for an interactive virtual coach able to offer advice based on perceived indicators of distress or anxiety.

A key challenge when building such nonverbal perception technology is to develop and validate robust descriptors of human behaviors that are correlated with psychological disorders such as depression, anxiety, or post-traumatic stress disorder (PTSD). These descriptors should be designed to support the diagnosis or treatment performed by a clinician; no descriptor is diagnostic by itself, but they show tendencies in people's behaviors. A promising result in this direction is the recent work of Cohn and colleagues who studied facial expressions and vocal patterns related to depression [5,6].

In this paper, we present and validate automatic behavior descriptors related to depression, anxiety and/or PTSD and in particular to a more generic distress measure introduced in Section 4.2. We introduce a new dataset, called the Distress Assessment Interview Corpus, which consists of 70+ h of dyadic interviews designed to study the verbal and nonverbal behaviors correlated with psychological disorders. We describe our approach in automatically assessing indicators of psychological disorders from head pose, eye gaze, facial expressions (smiles), and acoustic measures capturing the voice quality and monotonicity of the speech. We also investigate fidgeting and self-adaptor gestures occurring during these interviews.

The next section presents a previous work studying the relationship between nonverbal behaviors and psychological disorders. Section 3 introduces the research goals of this work. In Section 4 we describe the procedure for data acquisition, the used psychological measures, as well

☆ This paper has been recommended for acceptance by Qiang Ji.

* Corresponding author at: University of Southern California Institute for Creative Technologies 12015 Waterfront Dr. 90094, Playa Vista, CA.
E-mail addresses: scherer@ict.usc.edu (S. Scherer), Stratou@ict.usc.edu (G. Stratou), Lucas@ict.usc.edu (G. Lucas), emmam3@cam.ac.uk (M. Mahmoud), Boberg@ict.usc.edu (J. Boberg), gratch@ict.usc.edu (J. Gratch), rizzo@ict.usc.edu (A.(S.) Rizzo), morency@ict.usc.edu (L.-P. Morency).

[1] This work is an extension of the work in [4] originally published in the proceedings of the IEEE Automatic Face and Gesture Recognition Conference (FG) 2013.

http://dx.doi.org/10.1016/j.imavis.2014.06.001
0262-8856/

as the recorded population. Section 5 presents the multimodal behavior analysis platform MultiSense. The manual annotation scheme is introduced in Section 6, and the observed results of the automatic and manual analysis are presented and discussed in Section 7. Finally, Section 8 concludes the paper and introduces future directions of our work.

2. Related work

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, 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 on how both 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. Several research programs around the globe study these relationships, including a project funded by the Australian Research Council [7–12], a Department of Defense funded project [13–15], and the DARPA funded project that the present work is funded on [4,16–23]. General findings from these research programs and other investigations inform our search for automatic nonverbal behavior descriptors, so we first review these key findings. Some nonverbal behaviors associated with psychological disorders are summarized in Table 1.

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 significantly less mutual gaze [24], show nonspecific gaze, such as staring off into space [25] and avert their gaze, often together with a downward angling of the head [26]. The pattern for depression and PTSD is similar, with patients often avoiding direct eye contact with the clinician.

Table 1
Summary of nonverbal behaviors found in the literature. Nonverbal behaviors written in italics are part of the analysis in the present work.

Authors	Nonverbal behavior	Disorder
Fairbanks et al., 1982	↓ Mouth movements	Depression
	↓ *Smiling*	
	↑ *Self-grooming*	
	↑ *Turning head away*	
	↑ *Fidgeting*	Anxiety
Girard et al., 2013	↓ Smiles	Depression
	↑ Smile controls	
Hall et al., 1995	↓ Gestures	Depression
	↓ Speech	
	↑ Long pauses	
Kirsch and Brunnhuber 2007	↑ Anger	PTSD
	↓ *Genuine joy*	
Perez and Riggio 2003	↑ *Gaze down*	Depression
	↑ *Gaze aversion*	
	↓ Emotional expressivity	
	↓ Gestures	
	↑ Frowns	
Schelde 1998	↑ *Nonspecific gaze*	Depression
	↓ Mouth movements	
	↓ Interaction	
Waxer 1974	↓ *Mutual gaze*	Depression
Darby et al., 1984	↓ *Pitch variability*	Depression
	↓ *Loudness variability*	
	↑ *Harsh voice*	
	↑ *Speech monotonicity*	
Flint et al., 1993	↑ *Vocal tension*	Depression

Emotional expressivity, such as the frequency or duration of smiles, is also indicative of an underlying clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity [26,27], fewer mouth movements [28,25], more frowns [28,26] and fewer gestures [29,26]. Some findings suggest that 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 [30] than what is found in non-clinical populations. Social anxiety and PTSD share some of the features of depression and also have a tendency for heightened emotional sensitivity and more energetic responses including hypersensitivity to stimuli: e.g., more startle responses, and greater tendency to display anger [30], or shame [31].

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 [28]. Similarly, “self-adaptors”, such as rhythmically touching, hugging or stroking parts of the body or self-grooming, e.g. repeatedly stroking the hair [28], have been identified to be of interest in this field of research [32].

Also acoustic indicators for depression were investigated and reduced speech variability and monotonicity in loudness and pitch were found [33,34]. Further, depressed speech was found to show increased tension in the vocal tract and the vocal folds [35,19]. In particular, increased tense voice quality characteristics within a comparable recording setup to the one investigated in the present work were found [19]. In this work a virtual human interview setup was investigated [19], whereas here we investigate human to human interviews.

These findings of indicative speech parameters have led to additional classification experiments [36]; the analysis involved glottal flow features as well as prosodic features for the discrimination of depressed read speech. The authors identified glottal flow features to be chosen by the feature selection algorithm for the majority of the classifiers as well as energy-based features for female speakers. Several spectral and energy based features were investigated for their discriminative capabilities of read speech using Gaussian mixture models, with Mel frequency cepstral coefficients and the first three formants yielding promising results [8]. Also, acoustic spectral measures associated with psychomotor retardation at different time resolutions are investigated in an international challenge to identify depression severity in the subject's voice characteristics [13].

More recently vocal and facial expressions were found as indicators of depression severity using within-subject analysis of longitudinal data. Both participants' and interviewers' vocal timing and fundamental frequency were found to correlate with Hamilton Rating Scale for Depression scores [37]. For facial expressions within the same longitudinal dataset, both manually and automatically coded facial action coding scheme action units (AU) varied markedly with depressive symptom severity. Significantly lower overall AU 12 activity (i.e. fewer smiles), significantly higher overall AU 14 activity (associated with contempt), and significantly more AU 14 activity during smiling, i.e. smile controls were observed [38].

Few multimodal studies are found in the literature with [6] being one of the exceptions. Facial action units and variability of fundamental frequency (f_0) as well as latency to respond to questions have been investigated [6]. Both approaches, yield promising discriminative power with about 80% accuracy for each modality.

One recent brewing controversy within the clinical literature is whether the specific categories of mental illness (e.g., depression, PTSD, anxiety, and schizophrenia) reflect discrete and clearly separable conditions or, rather, continuous differences along some more general underlying dimensions [39]. This parallels controversies in emotion research as to whether emotions reflect discrete and neurologically distinct systems in the brain, or if they are simply labels we apply todifferences along broad dimensions such as valence and arousal. Indeed, when it comes to emotion recognition, some meta-reviews

suggest that dimensional approaches may lead to better recognition rates than automatic recognition techniques based on discrete labels.

The broad dimension receiving the most support in clinical studies is the concept of general distress. For example, one study examined a large number of clinical diagnostic interviews and found that diagnoses of major depression and PTSD were bettercharacterized by considering only a single dimension of general distress [40]. Several other researchers have statistically re-examined the standard scales and interview protocols used to diagnose depression, anxiety, and PTSD and found that they highly correlate and are better seen as measuring general distress [41–43]. For this reason, we will investigate if general distress may be a more appropriate concept for recognizing clinical illness in addition to the more conventional discrete categories.

3. Research goals

We seek to investigate the following research goals:

1. Automatic gaze descriptors: As discussed in [25,24,26], subjects with psychological disorders show increased averted gaze and nonspecific gazing behavior based on manual annotations. Within our analysis we both seek to confirm these findings with automatic descriptors and investigate quantitatively the dynamics of both the head as well as eye gaze during dyadic conversations. In particular, we study the downward angling of the head and the eye gaze for subjects with psychological distress.
2. Automatic smile descriptors: Additionally, findings in [28] support that a reduced number of smiles can be observed in subjects with psychological disorders. However, this could not be confirmed for the number of smiles and laughter of depressed subjects in [5], but an increased amount of masking was observed. Further, [30] found less genuine smiles in PTSD patients. Again, we seek to further analyze these findings by analyzing smiling behaviors quantitatively and dynamically. In particular, we analyze if a reduced average duration of smiles as well as a reduced intensity of smiles can be observed for subjects with psychological distress, due to increased amount of masking and a reduced amount of genuine smiles.
3. Automatic voice assessment: Several research papers have discussed the indicative power of speech characteristics with respect to psychological disorders [34,33,35,8]. Within thiswork we investigate if findings of the literature can be confirmed or extended to the analyzed dataset. In particular, we investigate measurements of voice quality on a breathy to tense dimension as well as measures related to the monotonicity of the speech. We compare findings in this work with those of a recent publication investigating the speech characteristics of distressed subjects in a virtual human interview scenario [19].
4. Manual self-adaptor annotation: An additional research goal of this work is to better study the typical regions of self-adaptors (i.e. self-touches) for people with psychological distress. These were observed for people with depressionand anxiety [28]. Through manual annotations we seek to better understand the type of fidgeting and self-adaptors displayed by people with psychological disorders.

4. Distress assessment interview corpus

In this section we discuss the procedure for data acquisition of the Distress Assessment Interview Corpus (DAIC). We further introduce the employed psychological measures, and the overall size and characteristics of the corpus.

4.1. Population

The DAIC was recorded on two sites, comprising three conditions. At a US Vets center in California, 57 subjects were interviewed face-to-face. At the USC Institute for Creative Technologies, 110 subjects were interviewed, 54 face-to-face and 56 over a teleconferencing set-up. The interviews were conducted by one of two female interviewers. Both have basic clinical or psychological experience.

The population of subjects who were interviewed at the USC Institute for Creative Technologies was recruited off of Craigslist. One ad asked for participants who had been previously diagnosed with depression, PTSD, or traumatic brain injury, while another asked for any subjects between the ages of 18 and 65. All subjects who met the requirements (age, adequate eyesight) were accepted. Each subject was randomly assigned to either the teleconferencing or face-to-face condition. Some also were connected to a BIOPAC to measure psychophysiological signals.

The population at the US Vets site was recruited from among the inpatient and out-patient populations there. The in-patient population consists entirely of veterans. Some spouses and veterans who had completed one or more out-patient programs or were in a non-resident program were among the subjects.

For this paper, only the participants that were assigned to the face-to-face, non-BIOPAC condition were considered, due to possible impact of cables to the behavior. Of those, 54 were those recruited from Craigslist, and 57 were recruited from the US Vets population.

When participants were asked about their history of particular psychological disorders, 59.4% reported depression, and 29.5% PTSD. This information was not independently confirmed and only self-reported. Following the self-assessment using the inventories introduced in Section 4.2, 29% scored positive for depression, 32% for PTSD, and 62% for anxiety. For the categories distress vs. no-distress we split the population into tertiles allowing for balanced group sizes after calculating the overall distress score.

4.2. Measures

Standard clinical screening measures were used to assess PTSD, anxiety, and depression. Further, we introduce and motivate a measure of general distress based on the observed correlation between these three measures.

4.2.1. Post-traumatic stress disorder checklist-civilian (PCL-C)

The PTSD Checklist-Civilian version (PCL-C) [44] is a self-report measure that evaluates all 17 PTSD criteria using a 5-point Likert scale and is widely used in PTSDresearch [45,46]. It is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994). Scores range from 17 to 85, and PTSD severity is reflected in the size of the score, with larger scores indicating greater severity. Sensitivity and specificity are reportedly 0.82 and 0.83, respectively for detecting DSM PTSD diagnoses. The PCL-C is scored based on the DSM-IV schema, with symptomatic responses (moderatelyor above) to at least six items from three categories. The scores are added to assess the severity of symptoms.

4.2.2. State/Trait Anxiety Inventory (STAI)

The State/Trait Anxiety Inventory (STAI) [47,48] is another commonly used self-report questionnaire that can be used in the formulation of a clinical diagnosis, to help differentiate anxiety from depression, for psychological and health research, and for the assessment of clinical anxiety in patients. The STAI is a validated 20-item self-report assessment scale which includes separate measures of transient (state) and enduring (trait) levels of anxiety. Many reliability and validity tests have provided evidence that the STAI is an appropriate and adequate assessment for studying anxiety [49]. Trait anxiety is assessed by adding up all scores and using the experimental STAI-T population mean of 34.84 plus one standard deviation (σ 9.21) for a rounded total cut-off of 44.

4.2.3. Patient Health Questionnaire-Depression 9 (PHQ-9)

The Patient Health Questionnaire-Depression 9 (PHQ-9) is a ten-item self-report measure based directly on the diagnostic criteria for major depressive disorder in the DSM-IV [50]. The PHQ-9 is typically used as a screening tool for assisting clinicians in assessing depression severity as well as selecting and monitoring treatment. Further, it has been shown to be a reliable and valid measure of depression severity [51]. Scores range from 0 to 27, with higher scores indicating higher depression severity. Due to IRB requirements, we used a 9-question PHQ-9 instrument, leaving off question 9 about suicidal thoughts. When scoring the PHQ-9, response categories 2–3(more than half the days or above) are treated as symptomatic and responses 0–1 (several days or below) as non-symptomatic. At least five of the first eight questions must be checked as symptomatic, including at least one of the first two questions. Additionally, question 10 must be checked as at least somewhat difficult. Severity is calculated by totaling the answers to all of the questions. A PHQ-9 score of at least 10 was used to determine a positive assessment, in addition to the previous requirements.

4.2.4. General distress

We observed significant correlations between the disorders (i.e. PTSD, anxiety, and depression) as assessed with the self-assessment questionnaires, with a significance level of $p < .01$. The screening outcome (i.e. positive or negative scoring) for depression correlated with PTSD with $\phi = .64$, using Pearson's correlation, screening outcome for depression and anxiety correlated with $\phi = .40$, and PTSD with anxiety correlated with $\phi = .43$.

When directly considering the scalar severity measures or scores of the three inventories, we found even stronger correlations with $\rho > .8$, as seen in Fig. 1. Based on this analysis, and several findings in theliterature that confirmed these co-morbidities [52,42], we decided to combine the three measures using principal component analysis to that of general distress. This measure of general distress forms the basis of our behavioral indicator analysis.

We performed a principal component analysis using oblique rotation (direct Oblimin) on the forty-six variables that made up the pooled metrics (i.e. STAI, PHQ-9, and PCL-C). Nearly all of the variables had high commonalities (>.7), while none had low commonalities (<.4), indicating that the results of principal component analysis are likely to be a little different from those of factor analysis. Additionally, the factors are expected to be interrelated, indicating the use of an oblique rotation.

The Kaiser–Mayer–Olkin measure of .940 was high, and Barlett's test of sphericity is $\chi^2\,(1035) = 6723.76$, $p < .001$, indicating that inter-item correlations were sufficiently high. An initial analysis was run, resulting in seven components with Eigenvalues over 1, explaining 72.12% of the total variance. The scree plot showed an inflection indicating that two components should be retained. Horn's parallel analysis also indicated to retain two factors. The two factors explain 59.5% of variance.

Table A.1 shows the pattern matrix after rotation. The clustering of items on the first and second factors suggest that component 1 represents anxiety and depression (the STAI and PHQ-9 scales) and component 2 PTSD (the PCL-C scale).Reliability is high, with Cronbach's $\alpha > .97$ for each of the factors.

The measure of general distress was computed as a linear combination of the participant's response to each question in the three questionnaires weighted by the observed factor loadings shown in Table A.1. The population was separated into tertiles: The upper third was considered as “distressed”, the middle third was “unclear” and was discarded, and the lower third was labeled “not distressed”. We opt for tertiles over a median split as such dichotomization treats people who arejust above the median and just below the median as categorically different, whereas using tertiles more appropriately includes only true “high” scorers in the distressed category and “low” scorers in the non-distressed category.

4.3. Procedure

For the recording of the dataset we adhered to the following procedure: After a short explanation of the study and giving consent, participants were left alone to complete a series of questionnaires at a computer. These included the following: The PTSD Checklist-Civilian version (PCL-C), the Patient Health Questionnaire, depression module (PHQ-9), the Spielberger State-Trait Anxiety Inventory (STAI-T), the Balanced Inventory of Desirable Responding (BIDR), the Big Five Inventory (BFI), the Reading the Mind in the Eyes (RME) scale, and the Positive and Negative Affect Schedule (PANAS). The following section describes the main three questionnaires used in this paper. This process took 30–60 min, depending on the participant.

Upon completion of the questionnaires, the participants were asked to sit down in a chair facing the interviewer directly. Both of them were video recorded with an HD webcam (Logitech 720p) and a depth sensor (i.e. Kinect). The participant and interviewer were about 7 ft apart. This distance was required for the Kinect to record depth information for the whole body of the subject/interviewer. This was not a problem for most of the participants, as only 5% said that it had a large effect on their interaction and only about 9% were uncomfortable or very uncomfortable with the distance.

Lavaliere microphones were attached to the lapel of the subject (Audio-Technica Pro 88 W; wireless microphone), and the recording

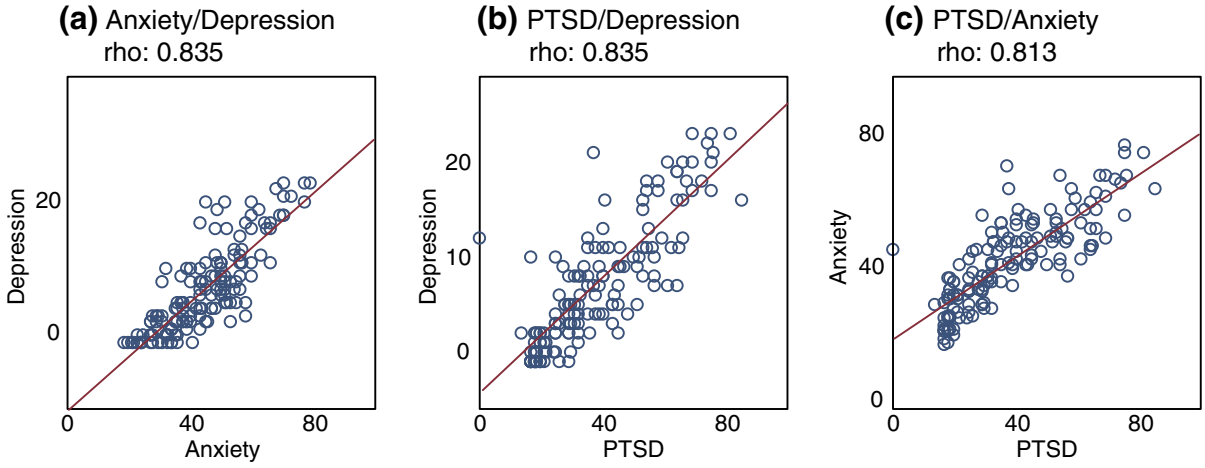

Fig. 1. Scatterplots showing the correlations between the conditions (a) anxiety and depression, (b) depression and post-traumatic stress disorder, and (c) anxiety and post-traumatic stress disorder. It is clearly seen that strong correlations are found ($\rho > .8$) for all combinations. The regression line fit to the data is shown in red.

was started. The interviewer then began a series of semi-structured questions. The questions were based partly on answers given by the participant during the questionnaire phase about their self-assessment questionnaire results and symptoms of PTSD or depression. The initial questions were neutral, but became more specific about possible symptoms and traumatic events as the interview progressed and as the participant's willingness to talk dictated. Interviews lasted between 30 and 60 min.

Finally, the participant was asked to complete the final set of questionnaires, which included a second PANAS, situational motivation questions, and questions about the participant's reactions to the interviewer. This phase took between 10 and 20 min. Participants were then debriefed, paid $25 to $35, and escorted out.

5. Automatic behavior analysis

In this section we describe the automatic analysis conducted in this paper in more detail. The goals of the automatic behavior analysis utilizing current state of the art behavior descriptors is two-fold: First, we would like to confirm findings from previous work that have identified several nonverbal behaviors that are characteristic of psychological disorders; and second, we seek to enrich previous findings, which have until recently predominantly relied on manual behavior annotations, with the quantitative analysis of the behavior dynamics. In the following we introduce our automatic analysis system MultiSense and the automatic behavior descriptors analyzed in the present study.

5.1. Automatic analysis system

For the automatic analysis we employ a multimodal sensor fusion framework called MultiSense. This is a flexible framework that was based on the Social Signal Interpretation framework (SSI) by [53] and it is created as a platform to integrate and fuse sensor technologies and develop probabilistic models for human behavior recognition. The modular setup of MultiSense allows us to integrate multiple sensing technologies including the following: *CLM*-Z FaceTracker by tebaltrusaitis-3d-2012 for facial tracking (66 facial feature points), GAVAM HeadTracker by [2] for 3D head position and orientation, OMRON's OKAO Vision for the eye gaze signal, smile level, and face pose and skeleton tracking by Microsoft Kinect SDK. It also includes RGB video capture via webcam device, synchronized audio capture and depth image capture via Microsoft Kinect sensor. The extracted acoustic measurements are currently not integrated in the realtimeversion of the sensing framework, but we plan to incorporate them in the near future.

MultiSense utilizes a multithreading architecture enabling all these different technologies to run in parallel and in realtime. Moreover MultiSense's synchronization schemes allow for inter-module cooperation, synchronized data recording, and information fusion. We can employ MultiSense for the fusion of the different tracker results to create a multimodal feature set that can be used to infer higher level information on perceived human behavioral states such as attentiveness, emotional state, agitation, and agreement by building probabilistic models for these states. Within this work, we are processing the synchronously recorded audiovisual tracker results in parallel, i.e. no explicit information fusion is utilized.

5.2. Automatic behavior descriptors

Based on our research goals (cf. Section 3) and our tracking technology we designed a few key behavioral descriptors that are informative for the psychological disorders, namely general distress, anxiety, depression, and PTSD. According to literature presented in Section 2 and summarized in Table 1, gaze and head turns are important features to observe (gaze aversion, gaze down and head turning are some of those behaviors associated with these features), as well as overall smile level (amount of smiling, and expression of genuine joy are associated with this feature). Additionally, we investigate acoustic parameters related to the tenseness and monotonicity of the voice to complement the visual analysis, as motivated by related work in Section 2. We seek to confirm these findings and add quantitative evidence to them by utilizing the automatic behavior description processes described above. We analyze and discuss the results in terms of our research goals in Section 7.

5.2.1. Visual behavior descriptors

To extract the features for this study we used the output from MultiSense to estimate the head orientation, the eye-gaze direction, smile level, and smile duration. The following are the visual behavior descriptors we analyzed in detail.

5.2.1.1. Vertical head gaze. This is a measure of how much the person is facing up or down during the conversation. MultiSense returns the 3D head orientation per video frame in radians [2]. The average head rotation is measured based on the x-axis (i.e. pitch).

5.2.1.2. Vertical eye gaze. This is a measure of the gaze vertical direction of the subject during the conversation. MultiSense returns the vertical gaze direction that can range in the span: [−60,60] degrees. We are measuring the average vertical gaze.

5.2.1.3. Smile intensity. This is the average smile level of the subject during the conversation. MultiSense returns the smile level, which can range in the span: [0,100], where 0 is the absence of smile and 100 a strong smile. Since MultiSense returns not only the existence but also the intensity of the smile in every frame, averaging that signal over the whole conversation includes the factors of how frequent, how strong, and how long the subject is smiling.

5.2.1.4. Smile duration. This is the average duration of the smiles of the subject during the conversation. It is again extracted using MultiSense. In this case, the smile level signal was thresholded to leave only in instances where the smile level is greater than 60. We proceeded with a small window smoothing process to get a binary smile pulse signal that allows us to count the number of strong smiles and approximate the duration of each. Based on the literature [5], these arefactors that can help differentiate between genuine and non genuine smiles.

The MultiSense signals that we extracted provide a confidence level for their output. We used the average confidence over the whole session as a screening measure to discard noisy videos.

5.2.2. Acoustic behavior descriptors

For the processing of the speech signals, we use the freely available COVAREP toolbox, a collaborative speech analysis repository available for Matlab and Octave [54].[2]COVAREP provides an extensive selection of open-source robust and tested speech processing algorithms enabling comparative and cooperative research within the speech community. The automatically extracted acoustic features werechosen based on previous encouraging results in classifying voice patterns of suicidal adolescents and distressed adults in [55,19] as well as the features' relevance for characterizing voice qualitieson a breathy to tense dimension [56,57]. Additionally, we extract features related to the monotonicity of the speech to complement the voice quality features. All features are sampled at 100 Hz.

5.2.2.1. Normalized amplitude quotient (NAQ). The first two features are derived from the glottal source signal estimated by iterative adaptive

[2] http://covarep.github.io/covarep/.

inverse filtering (IAIF, [58]). The output is the differentiated glottal flow. The normalized amplitude quotient (NAQ, [59]) is calculated using:

$$\mathrm{NAQ} = \frac{f_{ac}}{d_{peak} \cdot T_0}, \quad (1)$$

where d_{peak} is the negative amplitude of the main excitation in the differentiated glottal flow pulse, while f_{ac} is the peak amplitude of the glottal flow pulse and T_0 the length of the glottal pulse period.

NAQ is a direct measure of the glottal flow and glottal flow derivative and as an amplitude based parameter, was shown to be more robust to noise disturbances than parameters based on time instant measurements and has, as a result, been used in the analysis of conversational speech [60], which is frequently noisy. The parameter, however, may not be as effective as a voice quality indicator when a speaker is using a wide f_0 range [61].

5.2.2.2. Quasi-open quotient (QOQ). The quasi-open quotient (QOQ, [62]) is also derived from amplitude measurements of the glottal flow pulse and is a frequently used correlate of the open quotient OQ, i.e. the period the vocal folds are open. OQ is a salient measurement of the glottal pulse, thought to be useful for discriminating breathy to tense voice [63–65]. OQ can be defined as the duration of the glottal open phase normalized to the local glottal period. The quasi-open period is measured by detecting the peak in the glottal flow and finding the time points previous to and following this point that descends below 50% of the peak amplitude. The duration between these two time-points is divided by the local glottal period to get the QOQ parameter.

5.2.2.3. Spectral stationarity. To characterize the monotonicity used over utterances and the monotonicity of the speech, we make use of the so-called spectral stationarity measure SpecStat. This measurement was previously used in [66] as a way of modulating the transition cost used in a dynamic programming method used for f_0 tracking. Spectral stationarity, SpecStat is measured with:

$$SpecStat = \frac{0.2}{\mathrm{itakura}(f_i, f_{i-k}) - 0.8} \quad \in [0, 1], \quad (2)$$

where itakura(.) is the Itakura distortion measure [67] of the current speech frame f_i and f_{i-k} is the previous frame with $k = 1$. We use a relatively long frame length of 60 ms (with as shift of 10 ms; sampling rate100 Hz), and frames are windowed with a Hamming window function before measuring SpecStat. The long frame length was used in an attempt to characterize relatively long periods of maintained vocal tract articulation. SpecStat is close to 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference.

5.2.2.4. Intensity variation. Lastly, we investigate the intensity variation IntensityVar of the speech as the standard deviation of the signal energy over an utterance. The direct measure of energy is not suitable for evaluation in the contextof this work as the recording conditions over the two recording sites cannot be guaranteed to be held constant.

6. Manual behavior annotation

As mentioned in Section 3, one of the goals of this work is to identify the typical regions of self-adaptors and fidgeting behaviors, which were found to be correlated with psychological disorders as stated in [28]. As there are no automatic behavior descriptors currently available that robustly detect these behaviors, we complement the capabilities of our automatic descriptors with manual annotations. In the future, we plan to develop and train automatic descriptors for those behaviors based on the annotations. Particular interest was directed to the behaviors of people with PTSD, as this population is relatively understudied. The cues that were selected were divided into the following two annotation tiers.

6.1. Hand self-adaptors

For this annotation tier self-adaptors were annotated along with hand fidgets. These include hand tapping, stroking, grooming, playing with fingers or the hair, and similar fidgeting behaviors. These self-adaptors were separated into three distinct regions, namely head, torso, and hands. We split the manual annotation into these regions in order to be able to later disambiguate the regions on the body where these self-adaptors predominantly occur. We then compare the average durations of self-adapters to either (self-adaptors head) the head, face and hair region, (self-adaptors hands) the hands touch, or (self-adaptors torso) the arms and torso, in Section 7.5.

6.2. Leg fidgeting

In addition to self-adaptors, we annotated leg fidgets that include behaviors such as leg shaking and foot tapping. In our evaluation in Section 7.5, we then compare the average length ofthe subjects tapping or shaking their legs.

In total, four student annotators were recruited to carry out the full annotation. Each pair got one annotation tier assigned to them and went through a training phase. Both sets of annotators showed great agreement between annotations. Self-adaptors resulted after training in a Krippendorff's alpha of $\alpha = .77$; for the leg fidgets $\alpha = .84$ was observed [68]. These manual annotations were performed using ELAN [69].

After the training phase, each annotator started to annotate videos separately. To monitor the reliability of the coding in the post-training full annotation phase, every 10–15 videos each pair got assigned to the same video to annotate without knowledge that the other teammate was also annotating the same video, and inter-rater agreement was re-checked. Since findings suggest that annotators perform better when they know that their reliability is being assessed [70,71], annotators were informed that their reliability was measured but did not know which of the videos they worked on were used for cross-checking.

7. Evaluation and discussion

In this section we report the results of our investigations. The results are separated into two parts: The automatic behavior analysis using MultiSense and the manual nonverbal behavior annotation. In Sections 7.1 and 7.2, we report the results of the automatic nonverbal behavior descriptors. We analyze vertical head gaze, the overall vertical directionality of the gaze direction, as well as vertical eye gaze, the overallvertical directionality of the gaze direction. Further, we compare smile intensity, the average intensity of smiles as well as smile duration, the average duration of a smile. In Sections 7.3 and 7.4 we report results based on the assessed overall voice quality on a breathy to tense dimension, and the overall monotonicity of the speech. Finally, we report some supplementary findings based on the manual annotations in Section 7.5.

7.1. Automatic gaze descriptors

The vertical gaze measurements provided by MultiSense show significant results for the condition distress vs. no-distress. As head gaze and eye gaze are at least moderately correlated (Pearson's $\rho = .45$), we conducted a MANOVA with gender entered as an additional factor. With gender entered into the MANOVA, results indicate that there is no significant effect of gender ($F(2, 54) = 2.33$, $p = .11$) and the interaction between gender and distress is not significant ($F(2, 54) = 1.04$, $p = .36$).This analysis also did not reveal a significant effect of distress on gaze ($F(2, 54) = 2.65$, $p = .08$).

However, when gender is omitted from the MANOVA, distress shows a marginally significant effect ($F(2, 56) = 2.80$, $p = .07$) such that distressed subjects gaze downward marginally more than non-distressed subjects. Follow-up F tests for the individual gaze variables revealed that, distressed participants exhibited more of a downward eye gaze than non-distressed (distressed: 11.84 vs. non-distressed: 16.46; $F(1, 57) = 5.55$, $p = .02$), but did not differ on head gaze (distressed: −0.14 vs. non-distressed: −0.16; $F(1, 57) = 0.47$, $p = .50$). The distributions for the groups distress and no-distress are visualized in Fig. 2.

7.2. Automatic smile descriptors

The measurements of smiles provided by MultiSense show differences by distress condition. As smile intensity and duration are at least moderately correlated ($\rho = .52$), we conducted a MANOVA with gender entered as an additional factor. When genderis entered into the MANOVA, results indicate that females smile more than males ($F(2, 54) = 3.42$, $p = .04$); however, follow-up F tests for the individual smile variables revealed that, although females exhibit more intense smiles than males (female: 25.49 vs. male: 14.01; $F(1, 55) = 6.15$, $p = .02$), smile duration does not differ by gender ($F(1, 55) = 0.17$, $p = .68$). Additionally, in the MANOVA, there is no interaction between gender and distress group ($F(2, 54) = 0.17$, $p = .84$). This analysis also did not reveal a significant effect of distress on gaze ($F(2, 54) = 2.45$, $p = .096$). However, when gender is omitted from the MANOVA, distress shows a significant effect ($F(2, 54) = 3.86$, $p = .03$), such that distressed subjects smile less than non-distressed subjects. Follow-up F tests for the individual smile variables revealed that, distressed participants exhibited less intense smiles than non-distressed (distressed: 11.18 vs. non-distressed: 22.37; $F(1, 57) = 7.50$, $p = .008$), and exhibited marginally shorter smile duration (distressed: 0.87 vs. non-distressed: 1.14; $F(1, 57) = 3.35$, $p = .07$). The distributions for the groups distress and no-distress are visualized in Fig. 2.

Hence, based on our findings using the automatic behavior descriptors to estimate smile intensity and smile duration, we can confirm that our quantitative analysis of the smiling behavior is indeed correlated with psychological disorders of subjects. In particular, the automatic detection of decreased average intensity of smiles has strong benefits over traditional manual annotation approaches, as the coding of expression intensities can prove to be a very tedious and time consuming procedure.

These findings correspond to those in [27], where significantly attenuated positive emotional reactions were confirmed in a large meta-analysis across self-reported, physiological, behavioral and emotional reactivities in major depressive disorder studies. Even though we observed reduced smile intensities and reduced smile durations for subjects with psychological disorders, the nonverbal behavior of smiling might require some further analysis. For example, it is stated in [5] that an increase in masking behaviors of smiles was observed for depressed subjects. These masking behaviors might be of further interest in future analysis. Hence, we plan to annotate such masking behaviors (e.g. AU14 or AU12 of the facial action coding scheme (FACS) [72]) in a further annotation effort in order to confirm the hypothesis of [5] and to create training examples for the training of future automatic behavior descriptors.

7.3. Automatic voice quality descriptors

We utilized two common parameters to assess voice quality in the speaker's voice, i.e. NAQ and QOQ (cf. Section 5.2). Both measures are inversely correlated with the tenseness of the voice, i.e. the smaller the value the more tense the voice, and are highly correlated with each other ($\rho = .94$). When gender is entered into the MANOVA, results indicate that females exhibit more breathy voice qualities than males ($F(2, 39) = 3.65$, $p = .04$), and distressed participants exhibit more breathy voice than non-distressed ($F(2, 39) = 3.86$, $p = .03$). However, the difference between distressed and non-distressed speakers is not qualified by gender ($F(2, 39) = 0.15$, $p = .86$). When follow-up F tests for the individual variables are performed, these results are consistent across measures of voice quality. While females differ from males when voice quality is evaluated by either NAQ (female: 0.10 vs. male: 0.08; $F(1,40) = 7.28$, $p = .01$) or QOQ (female: 0.35 vs. male: 0.30; $F(1,40) = 5.39$, $p = .03$), the effect of gender never interacts with

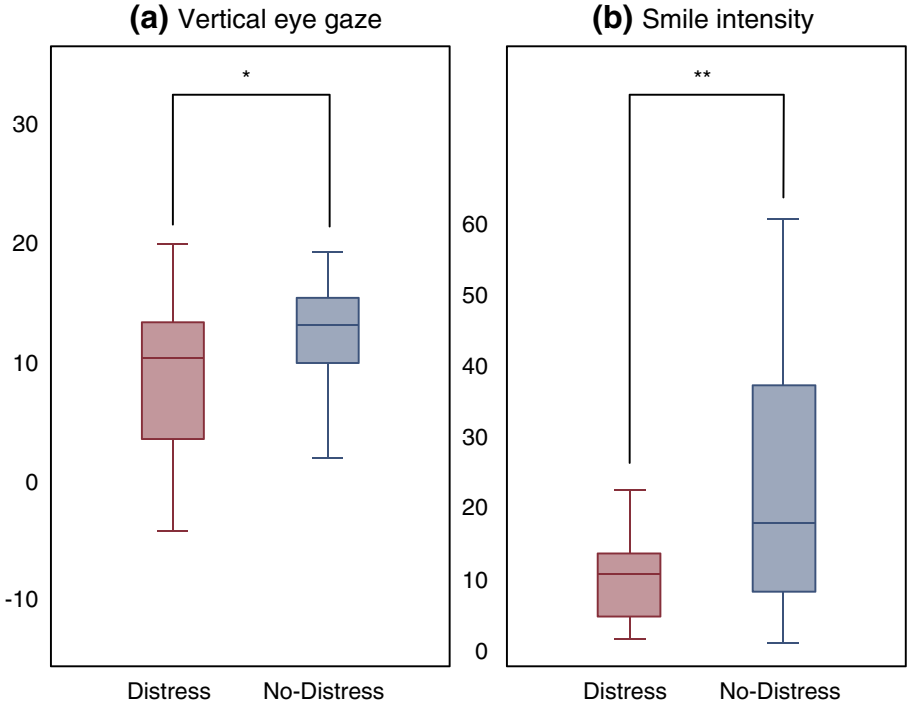

Fig. 2. Example of two automatic behavior descriptors. Boxplots show the significantly stronger overall downward angle of the (a) eye gaze ($p < .05$) and a significantly lowered average (b) smile intensity ($p < .01$) of subjects in the conditions distress vs. no-distress, as measured by MultiSense.

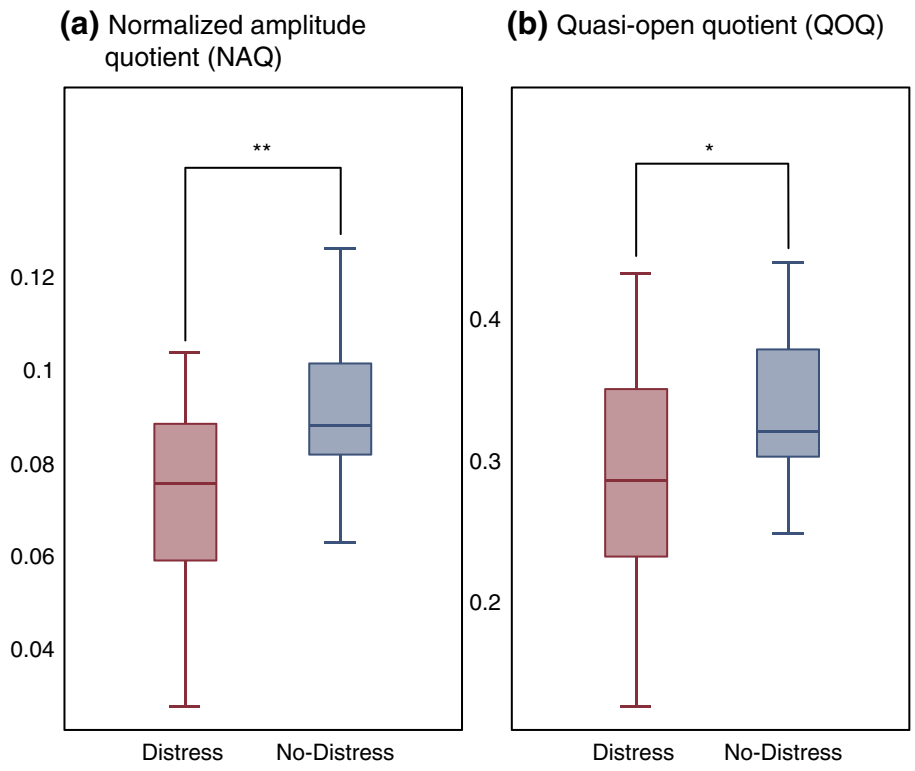

Fig. 3. Example of two automatic acoustic behavior descriptors. Boxplots show the significantly stronger overall tense voice qualities using (a) normalized amplitude quotient (NAQ; $p < .01$) and (b) quasi-open quotient (QOQ; $p < .05$) of subjectsin the conditions distress vs. no-distress.

distress (all $F < 0.07$ and $p > .79$). A MANOVA omitting gender confirms that distressed subjects exhibit more tense voice qualities than non-distressed subjects ($F(2, 41) = 4.66$, $p = .02$). Follow-up F tests reveal that distressed differs from non-distressed significantly when evaluated by NAQ (distressed: 0.07 vs. non-distressed: 0.09; $F(1, 42) = 7.81$, $p = .01$) or QOQ (distressed: 0.29 vs. non-distressed: 0.33; $F(1, 42) = 4.56$, $p = .04$). The distributions for the groups distress and no-distress are visualized in Fig. 3.

These findings correspond to those in [35,34], and in our a previous work of [19], where we found significantly more tense voice quality measures in the voice of depressed subjects and those with PTSD[3] within a virtual human interview corpus following a very similar protocol as the one employed in this study. The effects in [19], however, were stronger, which might be due to the more robust recording conditions using a microphone that has a better signal-to-noise ratio than the one utilized within this work.

7.4. Automatic monotonicity descriptors

With respect to the two investigated monotonicity measures, i.e. spectral stationarity and speech intensity variation (which were correlated at $\rho = .49$), we unfortunately cannot find any significant results within the dataset. MANOVA analyses with and without gender did not reveal any effect of distress, gender, or their interaction on monotonicity (all $F < 1.51$ and $p > .23$). Prior literature suggests that participants with psychological disorders should show more monotonous speech onaverage than those without [33,34].

We believe that the recording conditions might have been too variable and noisy for these measures to work properly. Further, the interviews contain a large variety of situations which might further dilute these measures to the point where no significant differences can be observed. Hence, one of the future avenues we would like to investigate with respect to this is to contextualize the analysis into categories of rapport building, question phase, and cool-down at the end of the conversations. In addition, we would like to experiment with the analysis window size for the spectral stationarity computation in the future. As monotonicity is most likely perceptually observed over longer periods of analysis than the investigated windows. Further, we would like to see if we can observe a significantly increased monotonicity within the virtual human interview recordings analyzed in [19], as they are recorded in much more stable conditions at only one site and with a more precise head-mounted microphone.

7.5. Manual annotation evaluation

As introduced in Section 6, we manually annotated the recordings on two tiers self-adaptors and leg fidgeting. Here, we report several results and indicators based on these. As head, hand and torso self-adaptors are moderately correlated with each other (ρ ranging from .43 to .54), we conducted a MANOVA that includes all three of these self-adaptors as dependent variables. Leg fidgeting was excluded from this analysis because it only produced in small correlations (ρ ranging from .11 to .19) with each of these self-adaptors. The results for leg fidgeting were therefore analyzed separately. We unfortunately cannot find any significant results within the dataset. Analyses with and without gender did not reveal any effect of distress, gender, or their interaction on self-adaptors or fidgeting (all $F < 2.80$ and $p > .10$).

Prior literature suggests that participants with psychological disorders should show more self-adaptors and fidgeting on average than those without [28]. Our results did not confirm the correlation between the longer durations of hands/legs fidgeting and psychological distress. In [28], hand tappings[4] were identified to be correlated with anxiety/

[3] Anxiety and distress were not investigated per se.

[4] This behavior falls in our analysis under the general term of hand fidgeting.

depression disorders. It is possible that the distance required by these sensors caused some behavioral changes in the participants that interfered with their self-adaptor and fidgeting behavior, however, only a small percentage of participants reported to feel uncomfortable with the setup, so this may not explain our failure to replicate findings [28]. As part of our future work, we plan to develop an automatic descriptor for such behaviors so that they can be automatically detected and further investigated in future analysis using the depth information collected using the Kinect sensors.

8. Conclusion

In this study we analyzed the Distress Assessment Interaction Corpus (DAIC) of face-to-face interactions with a confederate interviewer and a paid participant. Within the DAIC we investigated the nonverbal behaviors of subjects with psychological disorders (i.e. general distress, depression, anxiety, and PTSD) as measured with self-assessment questionnaires using both audiovisual automatic behavior descriptors and manual annotations. As we observed significant correlations between the disorders (i.e. PTSD, anxiety, and depression), we report results on a general distress measure as described in Section 4.2.

We focused our efforts on the behaviors, vertical gaze directionality, smile intensity and average duration, voice quality, speech monotonicity, and self-adaptors and leg fidgeting. The gaze and smile behaviors as well as the acoustic characteristics were analyzed using automatic behavior descriptors, whereas the hand self-adaptors and fidgets were analyzed using manual annotations, as there are no current robust automatic descriptors for such behaviors available.

As reported in Section 7, we found several statistically significant differences in the nonverbal behavior of subjects based on levels of general distress. Based on the four research goals stated in Section 3 we could identify the four main findings: (1) There are significant differences in the automatically estimated gaze behavior of subjects with distress. In particular, an increased overall downward angle of the gaze could be automatically identified using two separate automatic measurements, for both the face as well as the eye gaze; (2) using automatic measurements, we could identify on average significantly less intense smiles for subjects with distress as well as shorter average durations ofsmiles; (3) within the analyzed acoustic parameters we identified significantly tenser voice qualities for subjects with distress utilizing low-level glottal source parameters and no significant differences were found for the investigated measure of monotonicity; (4) based on the manual analysis, subjects with distress do not exhibit longer hand self-touches nor leg fidgets.

Whereas, we mainly analyzed the subject's behavior in the present study, for future work we plan to investigate audiovisual dyadic behaviors and patterns between the interviewer and the participant, in order to reveal additional indicators for both the presence and severity evaluation of psychological conditions. In [73], for example it was found that the clinician's behavior was strongly correlated with the patient's condition. Additionally, nonverbal attunement and entrainment was a strong predictor for the subsequent improvement of the patient's condition [74].

Acknowledgments

The effort described here is supported by DARPA under contract W911NF-04-D-0005 and the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

Appendix A. Factor analysis results

The original questions for STAI are protected and can be found in [47].

Table A.1
Questions from all utilized questionnaires with factor loadings.

Source	Question	Comp. 1	Comp. 2
STAI	Q3 — satisfied	1.012	
STAI	Q10 — happy	0.960	
STAI	Q13 — secure	0.923	
STAI	Q5 — failure	0.914	
STAI	Q16 — content	0.855	
STAI	Q1 — pleasant	0.810	
STAI	Q15 — inadequate	0.796	
STAI	Q7 — calm	0.760	
PHQ-9	Feeling bad about yourself or that you are a failure or have let yourself or your family down	0.748	
STAI	Q20 — steady	0.696	
STAI	Q8 — difficulties	0.696	
PHQ-9	Feeling down, depressed, or hopeless	0.682	
STAI	Q12 — self-confidence	0.661	
STAI	Q14 — decisions	0.652	
PHQ-9	Feeling tired or having little energy	0.647	
STAI	Q21 — turmoil	0.628	
STAI	Q4 — happy as others	0.625	
STAI	Q9 — worry	0.597	
STAI	Q2 — nervous	0.584	
PCL-C	Feeling distant or cut off from other people	0.575	0.346
PHQ-9	Little interest or pleasure in doing things	0.573	
PCL-C	Feeling emotionally numb or being unable to have loving feelings for those close to you	0.548	0.315
STAI	Q11 — disturbing thoughts	0.522	0.324
PHQ-9	If you checked off any problems, how difficult have these problems made it for you to do your work	0.513	0.338
STAI	Q6 — rested	0.481	
STAI	Q19 — disappointments	0.471	
PCL-C	Feeling irritable or having angry outbursts	0.467	0.435
PCL-C	Loss of interest in things that you used to enjoy	0.449	0.411
PHQ-9	Trouble falling or staying asleep, or sleeping too much	0.414	0.376
PHQ-9	Poor appetite or overeating	0.383	0.370
STAI	Q17 — unimportant thoughts	0.359	0.358
PCL-C	Trouble remembering important parts of a stressful experience		0.865
PCL-C	Feeling jumpy or easily startled		0.836
PCL-C	Having physical reactions when something reminded you of a stressful experience from the past		0.806
PCL-C	Being super alert or watchful on guard		0.761
PCL-C	Repeated, disturbing dreams of a stressful experience from the past		0.734
PHQ-9	Moving or speaking so slowly that other people could have noticed		0.727
PCL-C	Suddenly acting or feeling as if a stressful experience were happening again (as if you were reliving it)		0.724
PCL-C	Feeling very upset when something reminded you of a stressful experience from the past		0.699
PCL-C	Avoid activities or situations because they remind you of a stressful experience from the past		0.659
PCL-C	Having difficulty concentrating		0.626
PCL-C	Avoid thinking about or talking about a stressful experience from the past or avoid having feelings related to it		0.624
PCL-C	Repeated, disturbing memories, thoughts, or images of a stressful experience from the past	0.430	0.541
PCL-C	Trouble falling or staying asleep	0.412	0.452
PHQ-9	Trouble concentrating on things, such as reading the newspaper or watching television		0.450
PCL-C	Feeling as if your future will somehow be cut short	0.399	0.445

References

[1] T. Baltrusaitis, P. Robinson, L.-P. Morency, 3D constrained local model for rigid and non-rigid facial tracking, IEEE Computer Vision and Pattern Recognition (CVPR 2012), 2012.
[2] L.-P. Morency, J. Whitehill, J. Movellan, Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation, 8th IEEE International Conference on Automatic Face Gesture Recognition (FG08), 2008, pp. 1–8, http://dx.doi.org/10.1109/AFGR.2008.4813429.
[3] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2106–2111.

[4] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, L.-P. Morency, Automatic behavior descriptors for psychological disorder analysis, Proceedings of IEEE Conference on Automatic Face and Gesture Recognition, IEEE, 2013.

[5] L.I. Reed, M. Sayette, J.F. Cohn, Impact of depression on response to comedy: a dynamic facial coding analysis, J. Abnorm. Psychol. 116 (2007) 804–809.

[6] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Ying, M.H. Nguyen, M.T. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1–7.

[7] J. Joshi, R. Goecke, G. Parker, M. Breakspear, Can body expressions contribute to automatic depression analysis? Proceedings of IEEE Conference on Automatic Face and Gesture Recognition, 2013.

[8] N. Cummins, J.R. Epps, M.J. Breakspear, R. Goecke, An investigation of depressed speech detection: features and normalization, Proceedings of Interspeech 2011, ISCA, 2011.

[9] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J.R. Epps, G. Parker, M. Breakspear, Multimodal assistive technologies for depression diagnosis and monitoring, Journal on Multimodal User Interfaces, 72013. 217–228.

[10] J. Joshi, A. Dhall, R. Goecke, J.F. Cohn, Relative body parts movement for automatic depression analysis, Proceedings of Affective Computing and Intelligent Interaction (ACII), IEEE, 2013, pp. 492–497.

[11] N. Cummins, J.R. Epps, V. Sethu, M. Breakspear, R. Goecke, Modeling spectral variability for the classification of depressed speech, Proceedings of Interspeech 2013, ISCA, 2013, pp. 857–861.

[12] N. Cummins, J.R. Epps, E. Ambikairajah, Spectro-temporal analysis of speech affected by depression and psychomotor retardation, Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, 2013, pp. 7542–7546.

[13] J.R. Williamson, T.F. Quatieri, B.S. Helfer, R. Horwitz, B. Yu, D.D. Mehta, Vocal biomarkers of depression based on motor incoordination, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, ACM, 2013, pp. 41–48, http://dx.doi.org/10.1145/2512530.2512531.

[14] A.C. Trevino, T.F. Quatieri, N. Malyska, Phonologically-based biomarkers for major depressive disorder, EURASIP Journal on Advances in Signal Processing (42) (2011).

[15] D. Sturim, P. Torres-Carrasquillo, T. Quatieri, N. Malyska, A. McCree, Automatic detection of depression in speech using gaussian mixture modeling with factor analysis, Proceedings of Interspeech 2011, 2011, pp. 2981–2984.

[16] G. Stratou, S. Scherer, J. Gratch, L.-P. Morency, Automatic nonverbal behavior indicators of depression and ptsd: exploring gender differences, Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2013.

[17] Y. Zhou, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, J. Cassell, Multimodal prediction of psychological disorder: learning nonverbal commonality in adjacency pairs, Proceedings of Workshop Series on the Semantics and Pragmatics of Dialogue, 2013.

[18] J. Gratch, L.-P. Morency, S. Scherer, G. Stratou, J. Boberg, S. Koenig, T. Adamson, A. Rizzo, User-state sensing for virtual health agents and telehealth applications, Studies in Health Technology and Informatics, 1842012. 151–157.

[19] S. Scherer, G. Stratou, J. Gratch, L.-P. Morency, Investigating voice quality as a speaker-independent indicator of depression and ptsd, Proceedings of Interspeech 2013, ISCA, 2013, pp. 847–851.

[20] S. Scherer, G. Stratou, L.-P. Morency, Audiovisual behavior descriptors for depression assessment, Proceedings of International Conference on Multimdoal Interaction, ACM, 2013.

[21] D. DeVault, K. Georgilia, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, L.-P. Morency, Verbal indicators of psychological distress in interactive dialogue with a virtual human, Proceedings of SigDial 2013, Association for Computational Linguistics, 2013, pp. 193–202.

[22] D. DeVault, R. Artstein, G. Benn, et al., Simsensei: A virtual human interviewer for healthcare decision support, in: Proceedings of Autonomous Agents and Multiagent Systems (AAMAS), to appear, 2014. 1061–1068.

[23] J. Gratch, R. Artstein, G. Lucas, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, L.-P. Morency, The distress analysis interview corpus of human and computer interviews, Proceedings of Language Resources and Evaluation Conference (LREC), 2014.

[24] P. Waxer, Nonverbal cues for depression, J. Abnorm. Psychol. 83 (3) (1974) 319–322.

[25] J.T.M. Schelde, Major depression: behavioral markers of depression and recovery, J. Nerv. Ment. Dis. 186 (3) (1998) 133–140.

[26] J.E. Perez, R.E. Riggio, Nonverbal social skills and psychopathology, Nonverbal Behavior in Clinical Settings, Oxford University Press, 2003. 17–44.

[27] L.M. Bylsam, B.H. Morris, J. Rottenberg, A meta-analysis of emotional reactivity in major depressive disorder, Clin. Psychol. Rev. 28 (2008) 676–691.

[28] L.A. Fairbanks, M.T. McGuire, C.J. Harris, Nonverbal interaction of patients and therapists during psychiatric interviews, J. Abnorm. Psychol. 91 (2) (1982) 109–119.

[29] J.A. Hall, J.A. Harrigan, R. Rosenthal, Nonverbal behavior in clinician-patient interaction, Appl. Prev. Psychol. 4 (1) (1995) 21–37.

[30] A. Kirsch, S. Brunnhuber, Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects, Psychopathology 40 (5) (2007) 296–302.

[31] R. Menke, Examining nonverbal shame markers among post-pregnancy women with maltreatment histories, Ph.D. thesis Wayne State University, 2011.

[32] P. Ekman, W.V. Friesen, The repertoire of nonverbal behavior: categories, origins, usage, and coding, Semiotica 1 (1969) 49–98.

[33] A. Nilsonne, Speech characteristics as indicators of depressive illness, Acta Psychiatr. Scand. 77 (3) (1988) 253–263.

[34] J.K. Darby, N. Simmons, P.A. Berger, Speech and voice parameters of depression: a pilot study, J. Commun. Disord. 17 (2) (1984) 75–85.

[35] A.J. Flint, S.E. Black, I. Campbell-Taylor, G.F.G. Gailey, C. Levinton, Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression, J. Psychiatr. Res. 27 (3) (1993) 309–319.

[36] M. Elliott, M.A. Clements, J.W. Peifer, L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Biomed. Eng. 55 (1) (2008) 96–107.

[37] Y. Yang, C. Fairbairn, J.F. Cohn, Detecting depression severity from vocal prosody, IEEE Trans. Affect. Comput. 4 (2) (2013) 142–150.

[38] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, D.P. Rosenwald, Social risk and depression: evidence from manual and automatic facial expression analysis, Proceedings of IEEE Conference on Automatic Face and Gesture Recognition, IEEE, 2013.

[39] J.A. Russell, L.F. Barrett, Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant, J. Pers. Soc. Psychol. 76 (5) (1999) 805–819.

[40] J.D. Elhai, L. de Francisco Carvalho, F.K. Miguel, P.A. Palmieri, R. Primi, B. Christopher Frueh, Testing whether posttraumatic stress disorder and major depressive disorder are similar or unique constructs, J. Anxiety Disord. 25 (3) (2011) 404–410.

[41] P.J. Bieling, M.M. Antony, R.P. Swinson, The state-trait anxiety inventory, trait version: structure and content re-examined, Behav. Res. Ther. 36 (7–8) (1998) 777–788.

[42] G.N. Marshall, T.L. Schell, J.N.V. Miles, All ptsd symptoms are highly associated with general distress: ramifications for the dysphoria symptom cluster, J. Abnorm. Psychol. 119 (1) (2010) 126–135.

[43] P.A. Arbisi, M.E. Kaler, S.M. Kehle-Forbes, C.R. Erbes, M.A. Polusny, P. Thuras, The predictive validity of the ptsd checklist in a nonclinical sample of combat-exposed national guard troops, Psychol. Assess. 4 (24) (2012) 1034–1040.

[44] E.B. Blanchard, J. Jones-Alexander, T. Buckley, C. Forneris, Psychometric properties of the ptsd checklist (pcl), Behav. Res. Ther. 34 (8) (1996) 669–673.

[45] E.E. Bolton, M.J. Gray, B.T. Litz, A cross-lagged analysis of the relationship between symptoms of ptsd and retrospective reports of exposure, J. Anxiety Disord. 20 (7) (2006) 877–895.

[46] C.W. Hoge, C.A. Castro, S.C. Messer, D. McGurk, D.I. Cotting, R.L. Koffman, Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care, N. Engl. J. Med. 351 (1) (2004) 13–22.

[47] C.D. Spielberger, R.L. Gorsuch, R.E. Lushene, Manual for the State-Trait Anxiety Inventory, Consulting Psychologists Press, 1970.

[48] A. Beck, N. Epstein, G. Brown, R. Steer, An inventory for measuring clinical anxiety: psychometric properties, J. Consult. Clin. Psychol. 56 (1988) 893–897.

[49] A. Sesti, State trait anxiety inventory in medication clinical trials, Qual. Life Newsl. 25 (2000) 15–16.

[50] K. Kroenke, R.L. Spitzer, The phq-9: a new depression and diagnostic severity measure, Psychiatr. Ann. 32 (2002) 509–521.

[51] K. Kroenke, R.L. Spitzer, J.B.W. Williams, The phq-9, J. Gen. Intern. Med. 16 (9) (2001) 606–613.

[52] D. Campbell, B. Felker, C.-F. Liu, E. Yano, J. Kirchner, D. Chan, L. Rubenstein, E. Chaney, Prevalence of depression–PTSD comorbidity: implications for clinical practice guidelines and primary care-based interventions, J. Gen. Intern. Med. 22 (2007) 711–718, http://dx.doi.org/10.1007/s11606-006-0101-4.

[53] J. Wagner, F. Lingenfelser, N. Bee, E. André, Social signal interpretation (ssi), KI – Kuenstliche Intelligenz, 252011. 251–256, http://dx.doi.org/10.1007/s13218-011-0115-x.

[54] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep — a collaborative voice analysis repository for speech technologies, To Appear in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), 2014.

[55] S. Scherer, J.P. Pestian, L.-P. Morency, Investigating the speech characteristics of suicidal adolescents, Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, 2013, pp. 709–713.

[56] S. Scherer, J. Kane, C. Gobl, F. Schwenker, Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification, Comput. Speech Lang. 27 (1) (2013) 263–287, http://dx.doi.org/10.1016/j.csl.2012.06.001.

[57] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, C. Gobl, Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech, Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, 2013, pp. 7982–7986.

[58] P. Alku, T. Bäckström, E. Vilkman, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Comm. 11 (2–3) (1992) 109–118.

[59] P. Alku, T. Bäckström, E. Vilkman, Normalized amplitude quotient for parameterization of the glottal flow, J. Acoust. Soc. Am. 112 (2) (2002) 701–710.

[60] N. Campbell, P. Mokhtari, Voice quality: the 4th prosodic dimension, Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03), ICPhS, 2003, pp. 2417–2420.

[61] C. Gobl, A.N. Chasaide, Amplitude-based source parameters for measuring voice quality, Proceedings of ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), ISCA, 2003, pp. 151–156.

[62] T. Hacki, Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie, Folia Phoniatr. (1989) 43–48.

[63] H. Hanson, K. Stevens, H. Kuo, M. Chen, J. Slifka, Towards models of phonation, J. Phon. 29 (2001) 451–480.

[64] N. Henrich, C. d'Alessandro, B. Doval, Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data, Proceedings of EUROSPEECH, Scandanavia, 2001, pp. 47–50.

[65] R. Timcke, H. von Leden, P. Moore, Laryngeal vibrations: measurements of the glottic wave. Part 1: the normal vibratory cycle, Arch. Otolaryngol. Head Neck Surg. 68 (1) (1958) 1–19.

[66] D. Talkin, A robust algorithm for pitch tracking, in: W.B. Kleijn, K.K. Paliwal (Eds.), Speech Coding and Synthesis, Elsevier, 1995, pp. 495–517.
[67] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 23 (1975) 67–72.
[68] K. Krippendorff, Agreement and information in the reliability of coding, Communication Methods and Measures, 5 (2)2011. 93–112.
[69] H. Lausberg, H. Sloetjes, Coding gestural behavior with the NEUROGES-ELAN system, Behav. Res. Methods 41 (3) (2009) 841–849 (URL http://www.lat-mpi.eu/tools/elan/).
[70] B. Wildman, M. Erickson, R. Kent, The effect of two training procedures on observer agreement and variability of behavior ratings, Child Dev. (1975) 520–524.
[71] F. Harris, B. Lahey, Recording system bias in direct observational methodology: a review and critical analysis of factors causing inaccurate coding behavior, Clin. Psychol. Rev. 2 (4) (1982) 539–556.
[72] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978.
[73] A.L. Bouhuys, R.H. van den Hoofdakker, The interrelatedness of observed behavior of depressed patients and of a psychiatrist: an ethological study on mutual influence, J. Affect. Disord. 23 (1991) 63–74.
[74] E.N. Geerts, A.L. Bouhuys, R.H. van den Hoofdakker, Nonverbal attunement between depressed patients and an interviewer predicts subsequent improvement, J. Affect. Disord. 40 (1–2) (1999) 15–21.