Towards an Affective Interface for Assessment of Psychological Distress

Gale M. Lucas, Jonathan Gratch, Stefan Scherer, Jill Boberg & Giota Stratou
Institute for Creative Technologies, University of Southern California, Los Angeles, USA
Email: {lucas, gratch, scherer, boberg, stratou}@ict.usc.edu

ABSTRACT

***Abstract*—Even with the rise in use of TeleMedicine for health care and mental health, research suggests that clinicians may have difficulty reading nonverbal cues in computer-mediated situations. However, the recent progress in tracking affective markers (i.e., displays of emotional expressions on face and in voice) has opened the door to new clinical applications that might help health care providers better read nonverbal behaviors when employing TeleMedicine. For example, an interface that automatically quantified affective markers could assist clinicians in their assessment of and treatment for psychological distress (i.e., symptoms of depression and PTSD). To move towards this prospect, we will show that clinicians' judgments of these nonverbal affective markers (e.g., smile, frown, eye contact, tense voice) could be informed by such technology. The results of our evaluation suggest that clinicians' ratings of nonverbal affective markers are less predictive of psychological distress than automatically quantified affective markers. Because such quantifications are more strongly associated with psychological distress than clinician ratings of these same nonverbal behaviors, an affective interface providing quantifications of nonverbal affective markers could potentially improve assessment of psychological distress.**

Keywords—Affective interface, nonverbal behaviors, affective markers, psychological distress, assessment, clinical judgments

1. INTRODUCTION

Little is known about how accurate clinicians are at reading the nonverbal behavior of their clients; however, research suggests that they could have difficulty reading nonverbal cues in computer-mediated situations. Indeed, expert interviewers are less able to accurately read the nonverbal behavior of their interviewees in computer-mediated environments than face-to-face interviews [1]. However, the recent progress in automatically tracking affective markers (i.e., displays of emotional expressions on face and in voice) has opened the door to new applications that could help improve the ability of clinicians to read nonverbal behaviors more accurately during computer-mediated interaction with patients. For example, such affective interfaces could assist clinicians and mental health care providers in their assessment of and treatment for psychological disorders, such as depression, Post Traumatic Stress Disorder (PTSD), and anxiety, that manifest with affective symptoms.

In this paper, we are motivated by this possibility - that an affective interface providing quantifications of nonverbal affective markers (e.g., smile, frown, tense voice) could improve assessment of psychological distress (e.g., symptoms of depression and PTSD). To move towards this prospect, we will show that clinicians' judgments of these nonverbal affective cues could be informed by such technology. Clinical assessment of psychological distress could be improved to the extent that automatic quantifications of nonverbal affective markers are more predictive of psychological distress than clinicians' (and non-clinicians') judgments of these nonverbal affective markers.

2. PRIOR WORK

A. Nonverbal affective markers associated with psychological distress

A body of research has examined the relationship between nonverbal behavior and clinical conditions such as depression, PTSD, and anxiety. There are relevant studies dating back to the 1970's, when academic examination of this topic began. Most of this research resides in clinical and social psychology and the vast majority relied solely on manual annotation of facial and vocal expressions of emotion. Only very recently has automatic tracking of such nonverbal affective markers been employed, and such automatic quantifications have only been attempted in the field of computer science.

Despite at least forty years of research, few quantified relationships have been established between clinical disorders and expressed nonverbal affective behavior. This lack of agreement is due to a number of factors like difficulty in manually annotating data as well as inconsistencies in how psychological distress and affective markers have been defined across these studies.

In spite of the incongruences, there are some general trends in the relationship between psychological distress in depression and PTSD and nonverbal affective markers. In this work, we focus on four of the most well-researched affective markers: lack of smile, frown, lack of eye contact with interviewer, and tense voice.

The frequency of smiles and frowns has been shown to be predictive of depression and PTSD [2-4]. Specifically, depressed patients frequently display fewer and less intense smiles, and

more frowns. Even though this pattern is fairly consistent across studies, some of the findings suggest the picture is more complicated. For example, depressed patients may frequently smile, but these are perceived as less genuine and often shorter in duration than what is found in non-clinical populations.

As eye contact with an interlocutor often reflects more "settled" nondistressed affective states, we also see that psychological distress is associated with atypical patterns of gaze [4-6]. Depressed patients have a tendency to maintain significantly less mutual gaze, gazing instead to the left or to the right rather than directly at the interlocutor. The pattern for PTSD is similar to that for depression, with patients suffering from symptoms of either disorder avoiding direct eye contact with the interviewer.

The picture for tense voice is more mixed, as both tense and breathier voice quality has been associated with psychological distress. Some recent research has shown more tense voice quality among participants who report being psychologically distressed [7-11]. Authors were able to distinguish speakers with moderate to severe depression from speakers without depression using the Normalised Amplitude Quotient (NAQ), which was extracted fully automatically using the IAIF algorithm [12]. This research echoes older findings that more tense voice quality is present in distressed individuals [13-15]. However, other work [16-17], shows more breathy voice quality among individuals with increasing levels of depression. Therefore, it is unclear which direction the relationship between voice quality and distress might be, and the current work can contribute another investigation towards answering this question.

B. Automatic quantifications of nonverbal affective markers via MultiSense

The publicly-available multimodal sensing framework MultiSense [18] (Figure 1) assesses clinically-relevant aspects of patients' mental state via multimodal analysis of their vocal and visual communication using data-driven methods in computer vision and audio processing. Computer microphones and webcams provide audio and video, and a Microsoft Kinect sensor sends gaze-tracking data to MultiSense, which uses a suite of analyses software to process the data.

MultiSense enables the acquisition and integration of multimodal behavior, including facial and vocal expression, as well as patterns of eye gaze. Specifically, MultiSense tracks facial expression, gaze, and vocalization in real-time by integrating a number of pieces of technology: *CLM-Z FaceTracker*, iMotion's Facet, *GAVAM HeadTracker*, *OMRON's OKAO Vision,* and Microsoft Kinect SDK to track facial expression, eye and head position, and acoustic analysis of vocalizations, respectively [19-21]. These commercial software systems have been shown to be reliable in a number of scientific publications (e.g., [21]). Although in the current work we consider these features separately, it is important to note that, not only can all of these features be automatically identified (as we will do here in this work), but also, through machine learning, they can be summed via multimodal fusion [22-24].

Motivated by the body of research that has examined the relationship between nonverbal behavior and psychological distress (described in Section A above), MultiSense was used to automatically quantify these markers so that the association between automatically quantified nonverbal affective markers and psychological distress could be studied [7-10,25]. This research has verified that each of the nonverbal affective markers that were found in previous research to be predictive of psychological distress based on manual annotations were also predictive when annotated automatically. Specifically, self-reported depression and/or PTSD was associated with less intense smiles and shorter durations of smile, more frequent and intense frowns, decreased eye contact (increased angle of eye gaze), and more tense voice.

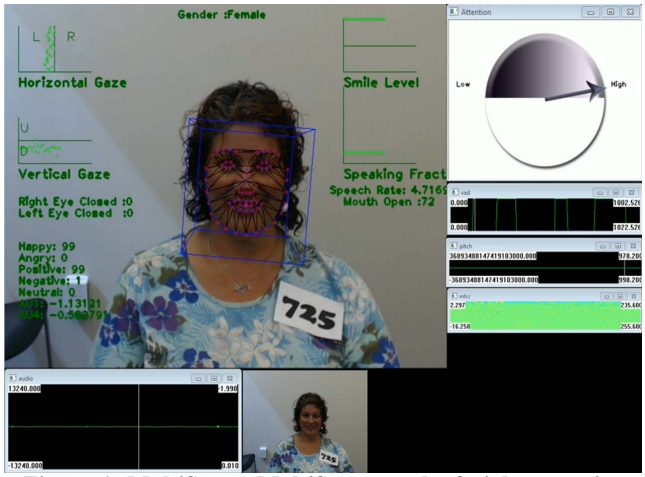

Figure 1. MultiSense. MultiSense tracks facial expression, gaze, and vocalization, automatically quantifying nonverbal affective markers.

3. CURRENT WORK

While this prior work demonstrates that quantifications of nonverbal affective markers (e.g., smile, frown, eye contact, tense voice) through MultiSense significantly predict psychological distress (e.g., symptoms of depression and PTSD), the current paper works toward the possibility that an affective interface that provides such quantifications can improve assessment of distress. We will show that clinicians' judgments of these nonverbal affective cues could be informed by such technology. In an evaluation, we will test whether clinicians' ratings of nonverbal affective markers are less (or more) predictive of psychological distress than automatically quantified affective markers.

A. Distress Assessment Interview Corpus videos

In this study, we utilized six videos (2 males, 4 females) from the Distress Assessment Interview Corpus [DAIC; 26].[1] In the DAIC, interviews were designed to simulate standard protocols for identifying people at risk for major depression or PTSD and to elicit nonverbal and verbal behavior indicative of such psychological distress. In order to increase the comparability of behaviors between individuals, we use a virtual human as an interviewer. A virtual human, i.e. a digital graphical representation of a human, in the present work allows for a higher level of control for the administration of stimuli (e.g. asking questions of varying levels of intimacy or acoustic parameters of the interviewer). Research suggests that when human interviewers are used, accommodation effects or mirroring is persistent in these human interlocutor studies [27-29] and could lead to biases in the observed results [30]. The DAIC interviews were collected as part of a larger effort named SimSensei to create a virtual agent that

interviews people and identifies verbal and nonverbal indicators of mental illness [31].

The DAIC was recorded at the USC Institute for Creative Technologies (ICT). Participants are drawn from two distinct populations: veterans of the U.S. armed forces and U.S. general population. The population of subjects consisted of individuals recruited from Craigslist and the direct recruitment of veterans at a US Vets facility in Long Beach. One posting on Craigslist asked for participants who had been previously diagnosed with depression or PTSD, while another asked for any subjects between the ages of 18 and 65. In this dataset, all participants were asked about their history of psychological disorders, and 54% reported that they have been diagnosed with depression in their past and 32% reported PTSD. In the DAIC, the self-reported symptoms of depression and PTSD are significantly correlated, as in previous work [8].

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. Standard clinical screening measures were used to assess symptoms of depression and PTSD, specifically: the Patient Health Questionnaire's depression module (PHQ-9) and the PTSD Checklist-Civilian version (PCL-C), respectively. 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 DSMIV [32]. The PHQ-9 is typically used as a screening tool for assisting clinicians in diagnosing depression as well as selecting and monitoring treatment. Further, it has been shown to be a reliable and valid measure for severity of depression symptoms [33]. Scores range from 0-27, with higher scores indicating higher depression severity. Due to IRB requirements, we used a 9-question PHQ-9 instrument, excluding question 9 about suicidal thoughts. Severity of depression symptoms is calculated by totaling the answers to all of the questions we asked.

The PTSD Checklist-Civilian version (PCL-C) [34] is a self-report measure that evaluates all 17 PTSD criteria using a 5-point Likert scale. It is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSMIV). Scores range from 17-85, and symptom severity is reflected in the size of the score, with larger scores indicating greater severity of PTSD symptoms. The scores are added to assess the severity of symptoms. The PCL is widely used across various research endeavors on PTSD.

Upon completion of the distress questionnaires, the participants were asked to sit down in a chair facing the virtual human interviewer directly, which was displayed on a large 50 inch monitor at about 1.5 meter distance. Within this work we utilize the SimSensei virtual human platform designed to create an engaging interaction through both verbal and nonverbal communicative channels [31]. The DAIC participants were video recorded with an HD webcam (Logitech 720p). An experimenter helped the participant set up the head mounted microphone (Sennheiser HSP 4-EW-3) and then the virtual human appeared and proactively started the semi-structured conversation. The audio was recorded at 16 kHz and a 16 bit resolution. The interaction between the participants and the fully automatic virtual human was designed as follows: the virtual human explains the purpose of the interaction and that it will ask a series of questions. It further tries to build rapport with the participant in the beginning of the interaction with a series of ice-breaker questions about Los Angeles, the location of the recordings. Then a series of more personal questions with varying polarity follow. The positive phase included questions like: "What would you say are some of your best qualities?" or "What are some things that usually put you in a good mood?". The negative phase included questions such as: "Do you have disturbing thoughts?" or "What are some things that make you really mad?". Neutral questions included: "How old were you when you enlisted? " or "What did you study at school?". This entire process took from 30-60 minutes, depending on the participant.

Three minute clips selected from the interviews were used in the present study. Only participants' answers, and none of the inciting questions, were included in the clips. All of the three minute videos created for the present study included answers to questions from both the negative and neutral questions. The videos were shown to our participants (described below in section B) online through the Qualtrics interface, via which they answered questions regarding the prevalence of the four affective markers. Furthermore, to garner automatic quantifications for comparison, the videos were run through MultiSense (described below in Section C).

B. Human ratings of affective markers

Three hundred and fifty two participants were recruited to participate in our study online. Forty eight of these participants had clinical training, and the remaining 304 were not clinicians. Clinician participants were recruited through online advertisements and flyers posted in psychology departments with clinical psychology training programs. All of our clinicians reported receiving training in clinical interviewing and/or assessment. They had been treating patients for an average of 3.91 years (SD = 3.76). Non-clinician participants were recruited via Amazon's Mechanical Turk. All participants who met requirements (i.e., native English speaker living in the United States) were accepted. Before completing the experiment, for clinicians, experience was confirmed through questions about their clinical training.

After giving consent and receiving a brief explanation of the study, participants then watched and rated six videos from the Distress Assessment Interview Corpus [DAIC; 26]. Specifically, they were asked to estimate the prevalence of four nonverbal affective markers: lacking a smile, frowning, lacking eye contact with interviewer, and tense voice. For each of these affective markers, participants used a scale ranging from 1 (none) to 7 (a lot) to describe the extent to which participants in the video displayed that particular affective marker during their DAIC interview. Upon completion, participants were thanked and paid.

C. MultiSense quantifications of affective markers

The MultiSense sensing platform [18] was used in this study to automatically detect four affective makers (lacking a smile, frowning, lacking eye contact with interviewer, and tense voice) from the six videos from the DAIC. 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 [20] 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), iMotion's Facet for facial expression recognition, GAVAM HeadTracker by [20] 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 real-time version 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 real-time. 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.

To quantify lack of smile, we considered the inverse of average smile level of the subject during the exchange captured in the video. MultiSense returns the smile level, which can range in the span from 0 to 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.

To quantify frowning, we considered the average level of AU4 (Brow Lowerer) that the subject displayed during the exchange captured in the video. MultiSense returns the level of AU4, which can range in the span from 0 to 100, where 0 is the absence of frown and 100 a strong frown. Since MultiSense returns not only the existence but also the intensity of the frown in every frame, averaging that signal over the whole conversation includes the factors of how frequent, how strong, and how long the subject is frowning.

To quantify lack of eye contact with interviewer, we used a measure of the horizontal eye gaze of the subject during the exchange captured in the video. MultiSense returns the horizontal gaze direction that can range in the span from -60 to 60 degrees. To assess gaze away from the interviewer, we used the absolute value (allowing for deviation from the interviewer in either direction) of the average vertical gaze.

To quantify tense voice, Normalized Amplitude Quotient (NAQ) was used. NAQ is derived from the glottal source signal estimated by iterative adaptive inverse filtering (IAIF, [11]). The output is the differentiated glottal flow. The Normalized Amplitude Quotient (NAQ, [35]) is calculated using this equation: $NAQ = f_{ac} / d_{peak} * T_0$, where d_{peak} is the negative amplitude of the main excitation in the differentiated glottal flow pulse, 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 [36], which is frequently noisy. These six videos were selected because they both returned sufficient (90%) confidence from MutliSense's metric of accuracy (which ranges from 1 to 100) and represented a range of scores from low to high in distress on the PHQ and PCL.

4. RESULTS

In order to consider the possibility that an affective interface that provides quantifications of nonverbal affective markers (e.g., smile, frown, eye contact, tense voice) can improve assessment of psychological distress (e.g., symptoms of depression and PTSD), we will test whether clinicians' (and non-clinicians') ratings of nonverbal affective markers are less (or more) predictive of distress than automatically quantified affective markers.

A. Predictive power of nonverbal affective markers for psychological distress

Table 1. Correlations between nonverbal affective markers and PHQ Depression symptom scale or PCL PTSD symptom scale

Nonverbal affective marker	Clinician ratings		Non-clinician ratings		MultiSense ratings	
	PHQ	PCL	PHQ	PCL	PHQ	PCL
Lack of smile	.43	.41	.38	.36	.61	.55
Frown	.31	.31	.32	.33	-.94	-.91
Lack eye contact	.18	.22	.20	.22	.86	.95
Tense voice	.04	.01	.16	.16	-.46	-.34

As can be seen in Table 1 and Figures 2 and 3, the correlations between MultiSense's quantifications of nonverbal affective markers and distress (symptoms of depression as measured by self-report on PHQ, symptoms of PTSD as measured by self-report on PCL) are substantially different than the correlations between human ratings of nonverbal affective markers and distress. Indeed, t-tests reveal that all correlations for MultiSense are significantly different from the corresponding correlation for clinicians' ratings of these nonverbal affective markers (all $ts(1,46) > 3.36$, $ps < .001$). Likewise, all correlations for MultiSense are also significantly different from the corresponding correlation for ratings made by non-clinicians (all $ts(1,303) > 10.31$, $ps < .001$). Across all these comparisons, correlations between MultiSense's quantifications of nonverbal affective markers and distress are significantly stronger than the correlations between human ratings and distress. However, for two of these affective markers -frown and tense voice- the correlation between MultiSense's quantifications and distress is also in the opposite direction from the correlation between human ratings of these affective markers and distress, finding less frowns and breathier voices associated with distress.

In contrast, when we considered whether the clinicians and non-clinicians differed in the predictiveness of their ratings for distress, clinicians' correlations tended not to differ from non-clinicians' correlations. However, the correlation between clinicians' ratings of tense voice and PTSD symptoms as

measured by self-reports on the PCL was significantly lower than the correlation for non-clinicians ($t(1, 341) = -2.10$, $p = .04$).

Likewise, the correlation between clinicians' ratings of tense voice and depressive symptoms as measured by self-reports on the PHQ was marginally lower than the correlation for non- clinicians ($t(1, 341) = -1.85$, $p = .065$). However, all other contrasts between clinicians' correlations and non-clinicians' correlations failed to reach significance ($ts(1,344) < 1.02$, $ps > .30$).

Not only did clinicians' correlations tend not to differ from non-clinicians' correlations, not all of the clinicians' correlations were

even significantly predictive of distress. Specifically, the correlations between clinician ratings of tense voice and distress were not significant ($ts(1,46) < 0.90$, $ps > .37$). On the other hand, all other correlations between human ratings of nonverbal affective markers and distress were significant (all other clinician $ts(1,46) > 3.38$, $ps \leq .001$, all non-clinician $ts(1,296) > 6.19$, $ps < .001$.

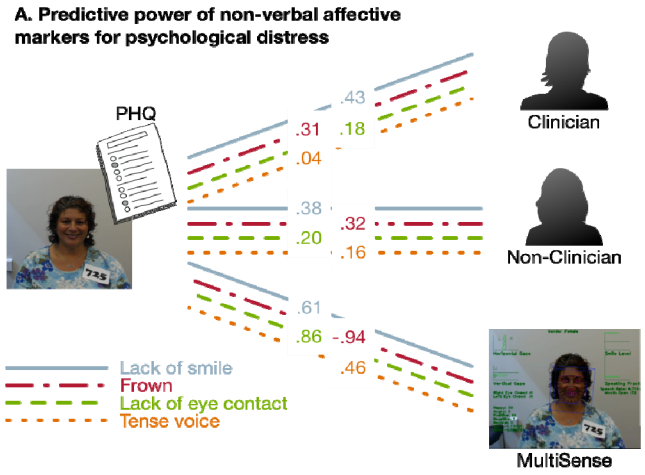

Figure 2. Correlations between nonverbal affective markers and PHQ Depression symptom scale

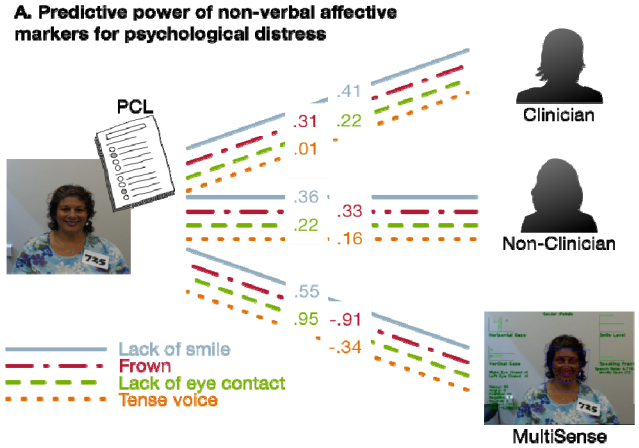

Figure 3. Correlations between nonverbal affective markers and PCL PTSD symptom scale

B. Relationship of human rated nonverbal affective markers to automatic quantifications by MutliSense

Table 2. Correlations between human ratings and MultiSense quantifications of the nonverbal affective markers

Nonverbal affective marker	Rater	
	Clinician	Non-clinician
Lack of smile	.71	.67
Frown	-.19	-.18
Lack eye contact	.39	.37
Tense voice	-.06	-.06

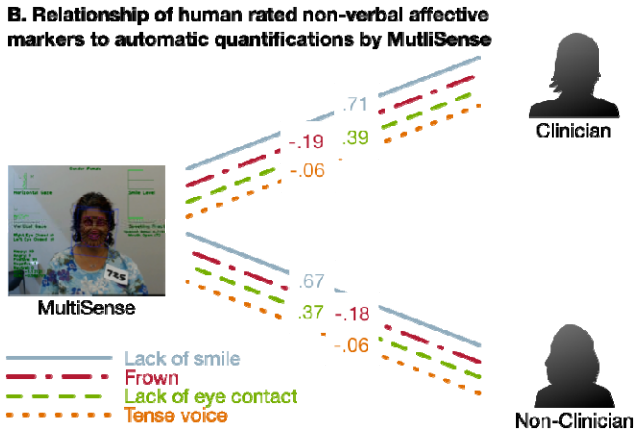

Figure 4. Correlations between human ratings and MultiSense quantifications of the nonverbal affective markers

We next considered the extent to which clinician (and non-clinician) ratings of nonverbal affective markers match the automatic quantifications made by MutliSense for these affective markers. To the extent that these ratings are unrelated to the quantifications, or even negatively related, there is more room for clinicians' assessments to be informed by an affective interface that includes these automatic quantifications made by MutliSense.

As can be seen in Table 2 and Figure 4, the direction and size of the correlations between human rated nonverbal affective markers and MultiSense's quantifications of these affective markers varied. Ratings of smile were strongly correlated ($ts(1,46) > 16.13$, $ps < .001$), eye contact more moderately correlated ($ts(1,46) > 8.81$, $ps < .001$), tense voice barely correlated ($t(1, 46) = -0.90$, $p = .37$ for clinicians, $t(1, 296) = -2.47$, $p = .01$ for non-clinicians), and ratings of frown were moderately negatively correlated ($ts(1,46) > 3.65$, $ps \leq .001$). Although these correlations vary across the type of affective marker, there are no significant differences within type of affective marker between the correlations with clinician ratings and the correlations with non-clinician ratings ($ts(1,349) < 0.74$, $ps > .46$).

5. DISCUSSION

Even with a rise in use of TeleMedicine, research suggests that clinicians may have difficulty reading nonverbal cues in computer-mediated situations. The results of our evaluation suggest that clinicians' ratings of nonverbal affective markers are less predictive of distress than automatically quantified affective markers. Not only are the correlations between MultiSense's quantifications of nonverbal affective markers and distress *different* than the correlations between human ratings of the affective markers and distress, but they are also *stronger*. The fact that they are *different* suggests that, in an affective interface, adding such quantifications could provide users with additional affective information that they are not otherwise accessing. Indeed, while ratings of some nonverbal affective markers were correlated with MultiSense's quantifications, others were not correlated, or even negatively correlated. The fact that the quantifications are *stronger* implies that this information could potentially be useful for improving assessments of distress.

Although the correlations are stronger, for two of these affective markers -frown and tense voice- the correlation between MultiSense's quantifications and distress is *in the opposite direction* from the correlation between human ratings of the affective markers and distress. While this does allow more room for such quantifications to provide users with additional affective information that they are not otherwise accessing, it also raises questions about what exactly is being measured by these automatic quantifications. Indeed, our quantifications did not track with human assessments for some of these affective markers: while human ratings of smile and eye contact at least moderately correlated with MultiSense's quantifications, ratings for tense voice barely correlated if at all, and ratings of frown were actually somewhat negatively correlated. Moreover, the associations with psychological distress were also contrary to at least some prior work on nonverbal affective behaviors among psychologically distressed populations. For tense voice, the literature has been mixed (as reviewed above), some work finding that psychological distress is associated with more tense voice quality, and other research demonstrating that distress is related to a breathier voice. In contrast to the research showing that distress is linked to more tense voice, the present work finds that distress is related to a breathier voice when quantified by MultiSense, but a more tense voice when rated by humans. Indeed, future research should endeavor to disentangle when and why distress is sometimes found to correspond with a more tense voice, and other times a breathier voice, and perhaps there are clues in differences between what NAQ measures and what is perceived by humans.

More surprising was the observed negative relationship between MultiSense's quantification of frown and psychological distress. This finding is contrary to more consistent (albeit small) body of literature showing that distress is associated with more frequent and/or stronger frowns. Although research has recently found that this relationship only holds for men [25], because we selected DAIC videos for this study such that an equal proportion of men and women were distressed (50% of each gender were distressed), gender differences cannot account for the observed negative relationship between quantification of frown and distress.

It is possible instead that, in this subset of DAIC videos, those who were quantified by MultiSense as having increased frowning behavior (i.e., those with lower distress) were in fact simply concentrating harder. As AU4 is activated when frowning and when concentrating, it is possible that, in the absence of actual frowning behavior in this subset, our "frowning" marker is solely tapping an expression of concentration. It is also reasonable that less distressed participants would display stronger expressions of concentration, given some heightened engagement with the task. Future research should endeavor to disentangle when and why activation of AU4 as measured by MultiSense might be capturing expressions of displeasure vs. expressions of concentration. Along these lines, another limitation of the current study is that these findings -more generally- may be an artifact of the specific videos analyzed. Follow-up studies could test this with different videos.

Overall, these results suggest that clinicians' ratings of nonverbal affective markers are less predictive of psychological distress than automatically quantified affective markers. Clinicians' ratings were also no more predictive than non-clinicians' (and actually *less* predictive for tense voice). Therefore, across both metrics, it seems that there is room to improve clinicians' assessments

Providing clinical care through *non-augmented* computer-mediated interactions can only obscure the observation of such nonverbal affective markers compared to face-to-face sessions [1]. However, augmenting computer-mediated interactions with quantifications of nonverbal affective markers could improve this otherwise impoverished means of communicating. Quantifications of nonverbal affective markers could be included both online and after the diagnostic interview or clinical session, therefore creating the potential to improve both online clinical judgment and post-session assessment. These quantifications would be designed to support the clinician's assessment overall, just like self-report, direct observation and other forms of assessment.

While automatic quantifications of markers like smile, frown, eye contact, tense voice may not always precisely track common perceptions of these nonverbal behaviors, they are more useful for an affective interface than manual annotations. Even if human ratings were available online (through crowdsourcing), there would always be more of a delay for human ratings than automatic ones. Even if human ratings could be obtained immediately, our results suggest that they would be less predictive of distress than automatic quantifications. For all the reasons outlined above, this paper makes a call for, and progress towards, an affective interface to improve clinical assessment.

REFERENCES

[1] Dunbar, N. E., et al. (in press). Synchronization of nonverbal behaviors in detecting mediated and non-mediated deception. J*ournal of Nonverbal Behavior*.

[2] Fairbanks, L. A., et al. (1982). Nonverbal interaction of patients and therapists during psychiatric interviews. Journal of Abnormal Psychology, 91, 109–119.

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

[4] Perez, J. E., & Riggio, R. E. (2003). Nonverbal social skills and psychopathology. In P. Philippot, R. S. Feldman, & E. J. Coats (Eds.) Nonverbal behavior in clinical settings, pp. 17–44. New York, New York: Oxford University Press.

[5] Schelde, J. T. M. (1998). Major depression: Behavioral markers of depression and recovery. The Journal of Nervous and Mental Disease, 186, 133–140.

[6] Waxer, P. (1974). Nonverbal cues for depression. Journal of Abnormal Psychology, 83, 319–322.

[7] Scherer, S., et al. (2013). Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD, in: Proceedings of Interspeech. ISCA, Lyon, France, pp. 847–851.

[8] Scherer, S., et al. (2013). Automatic Behavior Descriptors for Psychological Disorder Analysis, in Automatic Face and Gesture Recognition, 1 – 8.

[9] Scherer, S., et al. (2013). Audiovisual Behavior Descriptors for Depression Assessment, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI), 135–140.

[10] Scherer, s., et al. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. Image and Vision Computing.

[11] Quatieri, T.F., & Malyska, N. (2012). Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity, in: Proceedings of Interspeech, 1059–1062.

[12] Alku, P. (1992). Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. Speech Commun. 11, 109–118.

[13] Darby, J.K. (1984). Speech and voice parameters of depression: A pilot study. J. Commun. Disord. 17, 75–85.

[14] Flint, A.J. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. J. Psychiatr. Res. 27, 309–319.

[15] France, D.J. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Bio-Eng, 47, 829–837.

[16] Hönig, F. (2014). Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender, in: Proceedings of Interspeech. Singapore, pp. 1248–1252.

[17] Low, L.S.A. (2011). Detection of Clinical Depression in Adolescents; Speech During Family Interactions. Biomed. Eng. IEEE Trans. 58, 574–586.

[18] Hartholt, A., et al. (under review). All Together Now: Introducing the Virtual Human Toolkit. Intelligent Virtual Agents Conference.

[19] Baltrusaitis, T., et al. (2012). 3D constrained local model for rigid and non-rigid facial tracking. In Proceedings of IEEE Computer Vision and Pattern Recognition.

[20] Morency, L.P., et al. (2008). Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In Proceedings of IEEE International Conference on Automatic Face Gesture Recognition.

[21] Bartlett et al. (2006). Automatic recognition of facial actions in spontaneous expressions. Journal of multimedia,1(6), 22-35

[22] Quattoni, A. et al. (2007). Hidden-state Conditional Random Fields. In Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence.

[23] Morency, L.P., et al. (2007). Latent-Dynamic Discriminative Models for Continuous Gesture Recognition, In Proceedings of IEEE Conference on Vision and Pattern Recognition.

[24] Scherer, S., et al. (2013). Investigating Fuzzy-Input Fuzzy-Output Support Vector Machines for Robust Voice Quality Classification. Computer Speech and Language, 27, 263-287.

[25] Stratou, G. et al. (2013). Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In Proceedings of International Conference on Affective Computing and Intelligent Interaction.

[26] Gratch, J., et al. (2014). The Distress Analysis Interview Corpus of human and computer interviews. Proceedings of Language Resources Evaluation Conference.

[27] De Looze, C., et al. (2014) Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. Speech Communication 58: 11-34.

[28] Levitan, R. & Hirschberg, J. (2011) Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proceedings of Interspeech, 3081-3084.

[29] Shepard, C.A., Giles, H, & Le Poired, B.A. (2001) Communication Accommodation Theory. Wiley.

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

[31] DeVault, D., et al. (2014) Simsensei: A virtual human interviewer for healthcare decision support. In: Proceedings of Autonomous Agents and Multiagent Systems (AAMAS). pp. 1061-1068.

[32] Kroenke, K., & Spitzer, R.L. (2002) The phq-9: A new depression and diagnostic severity measure. Psychiatric Annals 32: 509-521.

[33] Kroenke, K., Spitzer, R.L. & Williams, J.B.W. (2001) The phq-9. Journal of General Internal Medicine 16: 606-613.

[34] Blanchard E.B., et al. (1996) Psychometric properties of the ptsd checklist (pcl). Behaviour Research and Therapy 34: 669-673.

[35] Alku, P. et al. (2002). Normalized amplitude quotient for parameterization of the glottal flow, J. Acoust. Soc. Am, 112, 701–710.

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

NOTE

1. Videos were selected randomly from those that met these criteria: 1) participants consented to sharing, 2) tracking worked for all frames, 3) half qualified were distressed, and 4) an equal proportion of men and women were distressed (50% of each).
2. This work was 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.