Perception of Stress: A Comparative Multimodal Analysis of Time-Continuous Stress Ratings from Self and Observers

Ehsanul Haque Nirjhar
Texas A&M University
College Station, Texas, USA
nirjhar71@tamu.edu

Winfred Arthur, Jr.
Texas A&M University
College Station, Texas, USA
w-arthur@tamu.edu

Theodora Chaspari
University of Colorado Boulder
Boulder, Colorado, USA
theodora.chaspari@colorado.edu

Abstract

Time-continuous ratings of stress are necessary for designing robust stress detection algorithms that operate in real-time. Common methods for obtaining these ratings in the field of affective computing are through self-reports or by employing multiple external observers. However, limited research has explored the association between these two methods, as well as their respective relation with multimodal bio-behavioral features. Using a mock job interview as a stress inducing task, this paper investigates time-continuous ratings of stress from self-reports and external observers. By analyzing the data from 223 question/answer exchanges from 31 participants, results suggest that observer ratings display low correlation with self ratings ($r = 0.145, p < 0.05$) and this degree of association varies depending on the inter-rater reliability of external observers. Findings also indicate that multimodal bio-behavioral features show higher association with observer ratings compared to self ratings, and therefore, machine learning models based on this multimodal data can estimate observer ratings ($CCC = 0.4688 \pm 0.247$) better than self ratings ($CCC = 0.2172 \pm 0.205$).

CCS Concepts

• **Human-centered computing → Ubiquitous computing**; • **Computing methodologies → Supervised learning**.

Keywords

Stress; Rating; Self-report; Observer; Job interview

ACM Reference Format:
Ehsanul Haque Nirjhar, Winfred Arthur, Jr., and Theodora Chaspari. 2024. Perception of Stress: A Comparative Multimodal Analysis of Time-Continuous Stress Ratings from Self and Observers. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685746

1 Introduction

Stress is ubiquitous in the modern world and is experienced by a considerable portion of the general population [33, 61]. It is defined as an individual's physiological and psychological response to challenges (i.e., cognitive demands [1, 50], social interaction [29, 62]) coming from the environment [16, 35]. Continued exposure to stress has been linked to the deterioration of physical and mental well-being (e.g., mental health complications, cardiovascular diseases) [22, 28]. Therefore, monitoring and detecting stress are necessary steps to reduce such adverse health outcomes. Continuous and unobtrusive detection of stress for providing timely interventions has been an active focus of research in affective computing and human-computer interaction domains. Stress is manifested by bio-behavioral signals from various modalities (e.g., speech, physiology, language, video), and analyzing these multimodal indices can contribute to continuous stress monitoring [15, 37, 57]. However, implementing successful continuous stress detection models requires access to reliable moment-to-moment ratings of stress that can be obtained from self-reports or external observers. Self-reported rating of stress (or 'self rating') involves an individual reporting their own 'felt' stress [1, 59]. On the other hand, 'observer ratings' involve external observers rating the 'perceived' stress of the target individual (i.e., how stressed the target individual seems to be) [51, 57]. Self ratings and observer ratings tend to capture different elements of the perception of stress, therefore the outcome of the continuous stress detection models developed through machine learning (ML) depends heavily on the choice of the type of ratings used to train these models [36, 38].

Differences between the self ratings and observer ratings of affect have been theorized in the Brunswik's lens model [9] and its subsequent modified versions [53]. According to these models, the expression of an emotion by individuals and the corresponding perception of the emotion by external observers follow encoding and decoding steps. Individuals express their mental state by altering their communication cues (e.g., facial expression, speech), referred to as distal cues, to encode their felt emotion. External observers perceive these transmitted cues, known as proximal cues, and decode them to understand the perceived emotion. Although the proximal cues are based on the distal cues, their perception to observers might not be same as intended initially, due to the individual differences among individuals and ambiguity of emotion [10, 11, 54]. Prior work has extensively studied the mismatch between self ratings and observer ratings in terms of categorical emotion labels (e.g., happy, sad, angry) or affect dimensions (e.g., arousal, valence) [8, 14, 42, 49]. These studies highlighted that there remains a low to moderate association between affect ratings obtained from self and observers. Findings from these studies also indicate that observer ratings of perceived emotion are predicted better by ML models, compared to self ratings of felt emotion [58, 60]. However, the majority of these studies used single-valued and discrete affect ratings. Few studies have explored the mismatch in self and observer ratings using time-continuous (i.e., moment-to-moment) ratings obtained

ICMI '24, November 04–08, 2024, San Jose, Costa Rica
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685746

from both parties [42]. Prior work mostly focused on general affect dimensions or labels (i.e., categorical emotion, valence, arousal). Differences in ratings obtained from self and observer in the context of specific affect content, such as stress, have not been received much attention in the literature. Moreover, the effect of inter-rater reliability on the association between self and observer ratings has not been explored. Finally, in order to develop a robust continuous stress detection model, it is necessary to understand how bio-behavioral measures from different modalities are associated with time-continuous ratings of stress, and how the association varies between self ratings and observer ratings. However, this has also not received much attention in prior work.

In this paper, we aim to address these research gaps in prior work by examining different aspects of time-continuous ratings of stress obtained from self-reports and multiple external observers, and investigating their association with multimodal bio-behavioral features. We pose the following research questions to facilitate further analysis:

RQ1: What is the degree of association between time-continuous self and observer ratings of the perception of stress? In which aspects are observer ratings different from self ratings?

RQ2: Which bio-behavioral features are most associated with self and observer ratings? Is there a difference in the degree of association across different modalities and raters?

RQ3: Does the prediction performance of ML models in continuous estimation of stress vary between self ratings and observer ratings when these ratings are employed as labels?

For this purpose, we conducted a study to collect self-reported time-continuous ratings of stress from 31 participants who were asked to complete a stressful task. The job interview is used as the stress-inducing task in this study as it is known to elicit stress among individuals due to being a zero-acquaintance high-stake interaction between an interviewer and an interviewee [3, 44]. Self ratings were provided at the end of the interview by each participant when they were asked to retrospectively watch the recorded video of the interview and rate their felt stress in a continuous manner. Next, we employed four raters as external observers to obtain their ratings of perceived stress of the participants while viewing video recordings of the participants who completed the job interview. Results from analyzing 223 question/answer (Q&A) exchanges indicate that time-continuous observer ratings of perceived stress display low correlation (i.e., $r = 0.145, p < 0.05$) with self-reported ratings of felt stress, and this degree of association is significantly affected by the inter-rater reliability of the external observers. Moreover, multimodal bio-behavioral features exhibit higher association with observer ratings, and therefore, these ratings can be estimated better (i.e., $CCC = 0.4688 \pm 0.247$) by ML models compared to self ratings (i.e., $CCC = 0.2172 \pm 0.205$).

2 Related Work

Prior work in affective computing has examined the relationship between the self ratings and observer ratings in various domains, such as public speaking anxiety [8, 49], emotion [10, 14, 42, 59], and stress detection [1, 46]. Busso *et al.* indicated that self ratings of affect tend to contain more extreme values compared to observer ratings [10]. Behnke *et al.* hypothesized that the anxiety rating obtained from self-reports and the level of anxiety perceived by external observers would exhibit a low to moderate correlation [8]. To test this hypothesis, they conducted a study with 95 participants who performed a public speaking task in front of an audience in a classroom setting and then self-reported their anxiety. Audience members also provided their perception of the degree of anxiety the participant felt which were moderately correlated (i.e., Pearson's $r = 0.37, p < 0.01$) with self-reported anxiety, supporting their initial hypothesis. In a similar work, Pörhölä found a low, positive correlation (i.e., $r = 0.10$, p-value not reported) between self-reported trait anxiety and the external observers' perceived anxiety during a public speaking performance completed by 47 participants in front of their peers [49]. Cheng *et al.* reported an even lower correlation (i.e., Pearson's $r = 0.08 - 0.21$, p-value not reported) between self-reported and observed affect [14]. However, the self and observer ratings used in these studies were not time-continuous, rather they were single-valued in nature offering an aggregate perception of a focal construct over the entire session.

Few studies have investigated the relationship between moment-to-moment affect ratings from self and observer. Truong *et al.* obtained time-continuous ratings of affect dimensions (i.e., valence, arousal) from 28 participants who self-reported their affect after a study involving multi-player gaming [59, 60]. Three external raters also provided their perceived affect rating in the same manner while watching the participants' video recording. Low to medium correlation (i.e., $r = 0.35 - 0.41$ for valence, $r = 0.24 - 0.33$ for arousal, p-value not reported) has been observed between these ratings from the two sources. Their study also indicated that ML models trained using observer ratings exhibited better prediction performance. In AMIGOS dataset [42], Miranda-Correa *et al.* obtained self-reported single-valued valence and arousal scores from participants after they watched emotional videos. In addition, the authors collected multiple time-continuous valence and arousal ratings from external observers who watched the video recordings of the participants. They compared the mean external observer rating with self-reports and found a significant positive correlation (i.e., $r = 0.44, p < 0.05$ for valence, $r = 0.15, p < 0.05$ for arousal). Aigrain *et al.* obtained self-reported stress ratings from 25 participants who completed mental arithmetic tasks and compared the ratings to the perceived stress ratings obtained from external observers [1]. A moderate correlation (i.e., $r = 0.41, p < 0.05$) was found between these ratings. The large magnitude of the correlation is potentially attributed to the well-defined and constrained stressors.

Collectively, these studies suggest that there is a low to moderate correlation between the self-reported and external observer ratings across various affect dimensions. The magnitude of the association depends on several factors, such as the affect content being evaluated, the emotion elicitation process, and the nature of ratings (i.e., continuous/singled-valued). However, there are limitations in prior work in terms of investigating the conditions under which the self-observer mismatch occurs and examining the effect of different multimodal bio-behavioral features on the self-observer mismatch. In addition, the effect of inter-rater reliability on the association between self and observer ratings has not been studied. This paper addresses these limitations and contributes to the current body of research in the following ways: (1) analyzing the association between time-continuous ratings of stress from both self-reports and external observers obtained from 223 Q&A exchanges from a research study involving mock job interviews; (2) examining the relationship between inter-rater reliability and the mismatch found in self and observer ratings; and (3) investigating how bio-behavioral features from various modalities exhibit different association with ratings from the two sources, and how this affects the performance of ML models that estimate stress in a time-continuous manner.

3 Data Description

3.1 Stress Elicitation through the Job Interview

In order to simulate a stress-inducing situation, we conducted a user study involving mock job interviews that were used as a stress-inducing task similar to prior work [6, 43]. We recruited 31 participants (27 male, 4 female) through campus-wide emails and advertisements for this study who participated as interviewees in the mock job interview. The average age of the participants was 38.48 years (*SD* = 10.48). Participants were military veterans who were transitioning (or had transitioned) to the civilian life after completing their military service. In order to make the mock interview more realistic, the job interviews were conducted by 11 interviewers who were industry representatives with prior experience in conducting interviews and recruiting personnel. The interview was conducted in a hybrid format, where the participants (i.e., the interviewees) came to our lab, while the industry representatives (i.e., the interviewers) participated remotely via Zoom video conferencing [63]. Before the day of the interview session, a customized mock job posting was crafted for each participant based on their résumé which they shared with the research team. The interviewers were provided with the résumé and the mock job posting for the corresponding participants, and they were asked to conduct the interviews as they would normally do as part of their work. Meanwhile, the participants were instructed to approach the task as if they had applied for the custom job posting and were interviewing for it. To motivate them further, they were informed to consider the interview as an opportunity to also practice their interviewing skill. These measures were implemented to ensure a naturalistic interaction during the mock job interview that would mimic the real-life interaction.

On the day of the interview, participants arrived at the lab and were briefed about the study. They were instructed to wear two wearable devices that captured their physiological signals during the entire duration of the study. These devices were the wrist-worn Empatica E4 wristband [21] and the chest-worn Actiheart 5 device [13]. The E4 wristband obtained electrodermal activity (EDA) signal sampled at 4 Hz, while the Actiheart 5, a single-lead electrocardiogram (ECG) recording device, collected ECG data at 512 Hz. Next, participants completed a set of measures pertaining to their demography, prior daily experience, and individual differences [5, 12, 18, 26, 27, 56]. After completing the measures, a relaxation session was administered in which the participants were shown a video of natural images with soothing music for 10 minutes to obtain their physiological reactivity at rest. Next, participants were introduced to the interviewer, who was connected through Zoom. Members of the research team were not present in the room during the interview. Audio and video of the interview session were recorded and downloaded from Zoom. Transcripts were also generated by Zoom and were later manually checked. In addition to the Zoom recording, a separate webcam was used to record only the participants. The interview sessions lasted approximately 19.2 minutes on average (*SD* = 5.7). After the interview, participants completed another set of measures where they recorded their thoughts about their performance in the interview. The study took about 2 hours for each participant to complete and provided multimodal data from different modalities, such as audio (i.e., speech), video, physiology (i.e., E4, Actiheart), and language (i.e., transcript).

3.2 Obtaining the Self Ratings of Felt Stress

After the completion of the interview session, participants were asked to provide a time-continuous self ratings of stress felt during the interview. For this purpose, we used the CARMA software [25] which is widely used in affective computing research. Participants watched their interview videos and provided moment-to-moment ratings of their stress using the computer mouse while watching the video in CARMA. The ratings were done on a continuous scale from 1 ('No Stress') to 5 ('High Stress'). The ratings were sampled at a rate of 1 Hz. The videos used for obtaining the self ratings included only the participants recorded by the separate webcam and did not include the interviewers that were available in the Zoom recording. This was done to expedite the process as the Zoom recordings were not available immediately after the end of the interview.

3.3 Obtaining the Observer Ratings of Perceived Stress

In order to obtain moment-to-moment ratings of perceived stress from external observers, we recruited four raters (one male, three female) who were undergraduate students majoring in psychology and had prior experience in emotion annotation tasks. They were trained to work with the CARMA software [25] before they started the tasks. The raters were asked to rate how stressed the participants were during the interview using the same scale as participants while watching the video. They watched the videos recorded by Zoom that captured both the participants and the interviewers.

As the videos of the entire interview session of individual participants were long in duration, providing ratings for the whole videos might cause fatigue, resulting in ratings with reduced quality [40, 52]. To address this issue, the interview videos were segmented into smaller videos of Q&A exchanges, where an exchange contained a question asked by the interviewers, the corresponding responses from the participants, and the back-channeling conversation during the questions and responses. The segmentation process resulted in 223 exchanges from 31 participants' interviews. The exchanges were approximately 2 minutes long on average (*SD* = 1.06). Raters were instructed to first watch all the videos of a participant so that they were aware of the context of the conversation before attempting to rate the perceived stress of the participants. Moreover, they were asked to rate the exchanges of the same participants in sequence to be uniform with the self-reported ratings. Continuous ratings from the raters were sampled at 1 Hz. A total of 80 exchanges from 12 participants were rated by four raters, while the remaining 143 exchanges from 19 participants were rated by three raters. The subsequent analysis presented in this paper is performed using the exchanges obtained from the segmentation instead of the whole videos. Therefore, the self ratings were also segmented into exchanges that are used in further analysis for uniform comparison with the ratings obtained from the raters.

4 Methodology

4.1 Inter-rater Reliability and Association between Self and Observer Ratings

The Pearson's correlation coefficient, r is chosen as a metric to quantify inter-rater reliability in the process of obtaining observer ratings of perceived stress (Section 3.3). For each exchange, the Pearson's r is computed for ratings obtained by all possible pairs of raters. There are six possible pairs for exchanges rated by four raters and three pairs for exchanges rated by three raters. An aggregated Pearson's r is obtained for each exchange using the Fisher's z-transformation [55]. An overall agreement metric for the entire dataset is also obtained by computing aggregated Pearson's r over all exchanges in a similar way.

Next, the arithmetic mean of the ratings from all raters for an exchange is used as the fused rating for that exchange and the mean rating is considered as a representation of the observer ratings. Computing the mean for fusing time-continuous ratings from multiple ratings is a common practice in prior work [40, 51]. In investigating the degree of association between self and observer ratings (**RQ1**), the association between self ratings and observer ratings is obtained for each exchange by computing the Pearson's r between self ratings and fused observer ratings. Similar to the inter-rater reliability, an overall association metric between the self ratings and observer ratings is computed by employing Fisher's z-transformation [55] to aggregate Pearson's r over all exchanges.

We identify the exchanges that exhibited higher inter-rater reliability compared to other exchanges. For this purpose, we inspect the Pearson's r for all possible rater pairs for each exchange and select the exchanges that had at least one pair of raters exhibiting Pearson's r over a given threshold value r_{th}. We choose $r_{th} = 0.4$ empirically based on prior work [41] and the preliminary observation of our data. The exchanges that have at least one pair of raters exhibiting $r > r_{th}$ are considered to be in the 'High reliability' group, while the remaining exchanges are assigned to the 'Low reliability' group. We examine the effect of inter-rater reliability on the association between the self-reports and the observer ratings by performing a t-test between Pearson r metrics obtained for the exchanges in these groups.

Finally, we inspect how the self ratings of stress differ from observer ratings by examining their distribution over all participants and exchanges. To quantify potential differences, we build a linear mixed effect (LME) model with the self ratings and the observer ratings as the independent variable and the dependent variable, respectively. Along with the fixed effect of self ratings on the observer ratings, both random intercept and random slope are considered for the LME model to account for the individual differences of the participants. The model is defined as:

$$r^{i,j}_{obs} = (\beta_0 + \beta_{0_i}) + (\beta_1 + \beta_{1_i}) \times r^{i,j}_{self} \qquad (1)$$

where $r^{i,j}_{obs}$ and $r^{i,j}_{self}$ refer to the self rating and observer rating obtained at the timestamp j for participant i. β_0 and β_1 denote the fixed intercept and slope, respectively while β_{0_i} and β_{1_i} indicate the random intercept and slope, respectively for participant i. We inspect the model parameters $(\beta_0, \beta_1, \beta_{0_i}, \beta_{1_i})$ to quantify the differences between self and observer ratings of stress.

4.2 Feature Extraction from Multimodal Bio-behavioral Signals

In order to identify the bio-behavioral features associated with self and observer ratings of stress and examine the difference among different ratings to answer **RQ2**, we perform a set of pre-processing and feature extraction steps on the obtained multimodal signals (Section 3.1). We extract features from different modalities, namely, acoustic, visual, physiological, and linguistic modalities.

4.2.1 *Acoustic Features.*

Audio signals obtained from the Zoom recording of the interviews contain speech signals from both the participants and the interviewers. The transcripts associated with the recording contain the timestamps corresponding to both speakers. Voice activity detection (VAD) is performed at the timestamps during which only the participants were speaking. Next, these segments are further used for acoustic feature extraction using the OpenSMILE toolkit [24]. For our experiment, we choose the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [23] due to its conciseness and extensive usage in prior work [6, 32, 57]. The eGeMAPS feature set consists of 88 features that include amplitude-related parameters (e.g., loudness, shimmer), frequency-related features (e.g., formants, jitter), and spectral parameters (e.g., Alpha ratio, harmonic difference). Features are computed over a 600 ms window with 100 ms overlap, and then averaged over non-overlapping 1 second windows similar to the ratings of stress.

4.2.2 *Visual Features.*

We use the OpenFace toolkit [7] to capture visual features from the participants' video recordings during the interview session. Different types of visual features are obtained, such as 2 gaze-related features, 3 head pose features, and 17 facial expression features. The gaze-related features include the eye gaze direction in radians in world coordinates along both X- and Y-axes. These features are related to the participants' eye contact with the interviewer, which is found to be an indicator of stress [4, 30]. Head pose features consist of head rotation angles around X-, Y- and Z-axes that are widely used in prior work related to job interview analysis [29, 44]. Finally, the facial expression features provide the intensity of 17 facial action units (AUs) in the frames of the video. These AUs are defined by the Facial Action Coding System (FACS) [20], and different AUs are found to be associated with stress and anxiety [1, 29]. These features are extracted at the same rate as the video frame rate and then aggregated over non-overlapping 1 second window using the mean and maximum values. This has resulted in 44-dimensional visual features for further analysis.

4.2.3 *Physiological Features.*

Visual inspection is performed on both the EDA signals and ECG signals to identify visible artifacts, followed by outlier removal and noise suppression [45]. In order to extract the features from clean EDA signals, the NeuroKit toolbox [39] is utilized. Five features are extracted from EDA signals, specifically, skin conductance level (SCL) parameters (i.e., mean, standard deviation) and skin conductance response (SCR) parameters (e.g., amplitude, onset and peak frequency). These features are extracted over a 10 second window with 1 second step size to match the sampling rate of the ratings of stress. Meanwhile, 23 heart rate variability (HRV) features are computed from the R-R interval series obtained from ECG signals over a 4 second window with 1 second step size using hrv-analysis toolbox [31]. These features contain 16 time-domain features (e.g., functionals of the N-N interval (NNI), heart rate (HR)) and 7 frequency-domain features (e.g., low frequency (LF), and high frequency (HF) power components, LF-HF ratio). Therefore, the physiological modality is characterized by a 28-dimensional feature set for each exchange.

4.2.4 Linguistic Features. To understand the psycholinguistic content of participants' speech, we utilize the Linguistic Inquiry and Word Count (LIWC) toolbox [48]. LIWC computes the count or percentage of words describing various constructs known as LIWC categories. The LIWC categories consist of general descriptors (e.g., word count, words per sentence), summary constructs (e.g., analytical thinking, clout), linguistic dimensions (e.g., nouns, verbs), psychological concepts (e.g., cognition, emotion), and informal language identifiers (e.g., filler words, confluency). Punctuation related features are excluded from the analysis. Overall, 81 linguistic features are computed for each exchange using the participants' responses to the questions asked by the interviewer. It is to be noted that the linguistic features extracted in this work are not time-series sampled at 1 Hz for each exchange. This is because these features are not continuous time signals and require the context of the entire exchange to be reliably computed. Instead, the 81-dimensional feature set is computed over the entire exchange.

4.3 Estimation of Time-continuous Ratings of Stress using Multimodal Features

We use a long short-term memory (LSTM) neural network to estimate the time-continuous self and observer ratings of stress using the multimodal bio-behavioral features (Section 4.2) to answer **RQ3**. Features obtained from acoustic, visual, linguistic, and physiological modalities are employed for this purpose. This model is similar to the baseline model in the MUSE-STRESS sub-challenge of the MuSe 2021 challenge [57]. The LSTM model consists of 2 hidden layers where each layer contains 64 hidden states, followed by a fully connected layer with 64 states before the output layer. For incorporating the static linguistic features into the ML model with other time-continuous modalities, we employ a two-layer feedforward neural network (64 hidden states, 64 output states) parallel to the LSTM to obtain their embeddings, and fuse them with the embeddings from LSTM before feeding them to the fully connected layer. A window size of 60 seconds and hop size of 10 seconds are chosen for the LSTM model. The model is trained with a learning rate of 0.005 with a scheduler with 5 epoch patience and L2-regularization penalty of 0.001. These model hyperparameters are chosen based after a set of experiments involving hyperparameter tuning. The model is trained for 200 epochs, with an early stopping criterion of no performance improvement over 15 consecutive epochs. A custom loss $\mathcal{L}$ based on the concordance correlation coefficient (CCC) loss (i.e., $\mathcal{L}_{CCC} = 1 - CCC$) and mean square error (MSE) loss (i.e., $\mathcal{L}_{MSE}$) is used to optimize the model where $\mathcal{L} = \mathcal{L}_{CCC} + \mathcal{L}_{MSE}$. This is done to prioritize both the increase of CCC and decrease of MSE, as these metrics do not always exhibit one-to-one mapping [47]. The model is trained via a leave-one-subject-out cross validation scheme using 223 exchanges from 31 participants. Training and testing of ratings from two sources are performed separately. The experiments are repeated 10 times for each configuration to account for any randomness. Both CCC and MSE metrics are used as the evaluation metric to measure the performance of the model in estimating the moment-to-moment self or observer ratings of stress using the multimodal features.

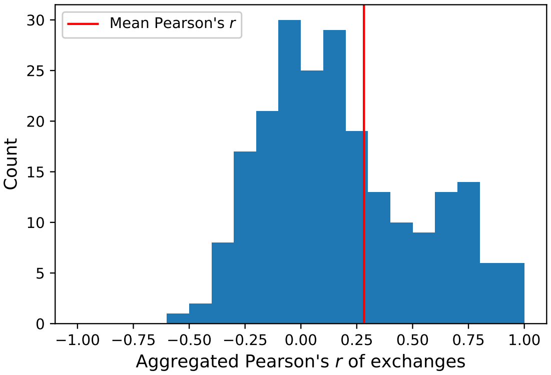

Figure 1: Distribution of inter-rater reliability measured by aggregated Pearson's r over all exchanges.

5 Results

5.1 Examining Observer Ratings: Inter-rater Reliability and Association with Self Ratings

The aggregated Pearson's r for all possible rater pairs is used as the inter-rater reliability metric in this work. Fig. 1 exhibits the distribution of the reliability metric over the entire dataset (i.e., 223 exchanges from 31 participants). The distribution is bimodal in nature with a larger peak around $r = 0.1$ and a smaller peak around $r = 0.7$. The mean Pearson's r of over all exchanges is $r = 0.282, p < 0.05$ which indicates a low to moderate inter-rater reliability when raters were asked to rate the perceived stress of the participants [2, 17]. Such low correlation has been found in prior work [59], as the perception of affect content such as stress is subjective. However, this distribution presents the aggregated Pearson's r for each exchange, and the aggregated r can be affected by some rater pairs showing lower agreement compared to other pairs. To examine this, the exchanges are separated into a 'High Reliability' and a 'Low Reliability' group based on whether an exchange has a pair of raters with Pearson's r higher than the threshold value of $r_{thres} = 0.4$ (Section 4.1). This results in 144 exchanges falling into the 'High Reliability' group, while the remaining 79 exchanges being assigned to the 'Low Reliability' group. The inter-rater reliability for the exchanges in 'High Reliability' group is found to be $r = 0.453, p < 0.05$ compared to the 'Low Reliability' group (i.e., $r = -0.0714, p < 0.05$), which is higher than the overall reliability.

Next, we investigate the association between the self ratings and the observer ratings by computing Pearson's r, where the observer rating is obtained by computing the arithmetic mean over ratings from all raters for an exchange. A low correlation has been found (i.e., $r = 0.145, p < 0.05$) between the self and observer ratings of stress, which is consistent with prior work [14, 42, 59]. However, a significant difference in the self-observer rating association is found between the exchanges of the 'High Reliability' and 'Low Reliability' groups, as shown in Fig. 2. The t-test confirms (i.e., $t(221) = 3.409, p < 0.05$) that the exchanges in the 'High Reliability' group exhibit higher correlation (i.e., $r = 0.216, p < 0.05$) between

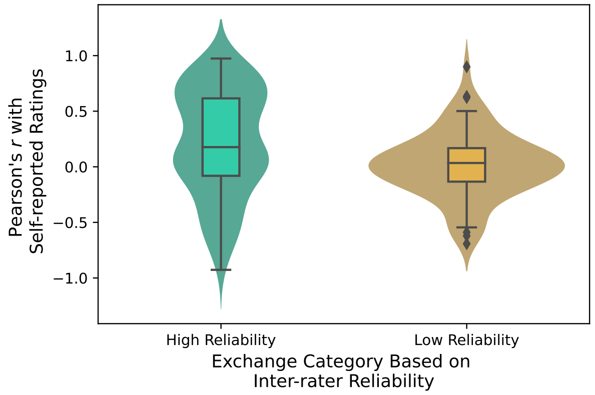

Figure 2: Difference in the distribution of the association between self-reported ratings and observed ratings based on the inter-rater reliability of the exchanges.

self and observer ratings, compared to the correlations (i.e., $r = 0.015, p < 0.05$) in the 'Low Reliability' group. Therefore, the result indicates that the inter-rater reliability plays an important role on the association between self ratings of felt stress and observer ratings of perceived stress.

Next, we examine the difference in the nature of ratings of felt stress from the participants' self-reports and the rating of perceived stress by the observers for the corresponding exchanges. A scatter plot showing the self ratings and the corresponding observer ratings is presented in Fig. 3. It is evident that the observer ratings rarely exceeded 3 on the 1–5 point scale, and in very few cases, the observer ratings were around 4. Conversely, the self ratings tend to contain more extreme values, even when the observer rating is lower. In Fig. 3, the majority of the data points (i.e., 60.01%) fall below the identity line $y = x$. Significant fixed effects are found by the LME model, where the fixed intercept is $\beta_0 = 1.208, p < 0.05$ and the fixed slope is $\beta_1 = 0.17, p < 0.05$. The variance of the random intercept and random slope are 0.216 and 0.05, respectively. The fitted line in Fig. 3 is described by considering the fixed effects in (1), while the shaded area denotes the 95% confidence interval resulting from the random effects. Results suggest that self-reported ratings contain more extreme values than the observer ratings, and the participants tend to consider themselves more stressed even when they are perceived as being less stressed by external observers.

5.2 Inspecting the Association between Bio-Behavioral Features and Stress Ratings

To answer **RQ2**, we investigate the association between the bio-behavioral features and time-continuous self and observer ratings of stress. In case of time-continuous modalities (i.e., acoustic, visual, and physiological), the Pearson's r is computed for each exchange and the aggregated Pearson's r is computed over all exchanges. However, a different approach is taken for the static linguistic features as they were computed over the entire duration of an exchange, and they do not have the same temporal resolution as the ratings. Therefore, the Pearson's r between the linguistic features and the maximum rating of exchanges is obtained, since the maximum rating captures the most salient rating in terms of felt or perceived stress. Table 1 presents the correlation of the features

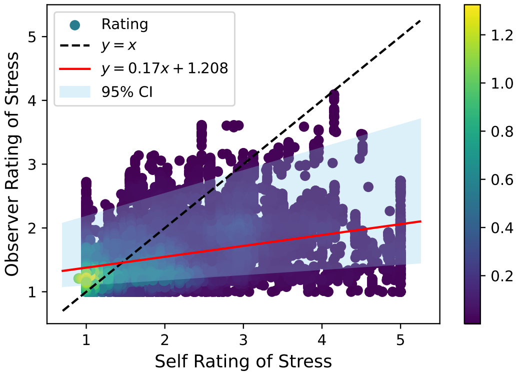

Figure 3: Self-reported ratings and the corresponding observer ratings for all the timestamps of all exchanges. The colorbar on the right shows the density of the scatter plot.

from different modalities with self ratings and observer ratings of stress. Detailed correlations for all extracted features are available in the supplementary material. From the result, it is clear that all types of features exhibit stronger associations with observer ratings compared to self ratings. This might have been attributed to the process of obtaining the ratings. The raters had to watch the audio-visual recording of the interview multiple times to understand the context to provide the ratings. On the other hand, participants performed the retrospective rating after completing the interview and they might have relied on recalling their experience based on the linguistic content, instead of relying on the audio-visual recording. Among the different modalities, observer ratings depict the highest correlations with acoustic features, while self ratings depict the highest correlations with linguistic features. Despite the magnitude of correlations, the most predictive features are similar for both self and observer ratings. The most prominent acoustic features are the formant frequencies (i.e., F3, F2), fundamental frequency, and shimmer. Among the visual features, head rotation and gaze angle, potentially associated with eye contact and head movement [29, 43], are positively correlated with ratings of stress. The use of nonfluencies and informal language are found to be positively correlated, while the word count is negatively correlated with both self-reported ratings and observer ratings of stress. Finally, mean SCL and HR are the more prominent physiological features exhibiting positive correlations with observer ratings.

5.3 Estimating Moment-to-moment Stress using Bio-Behavioral Features

Finally, we answer **RQ3** by training an LSTM model with multimodal bio-behavioral features to estimate the moment-to-moment ratings of stress obtained from self-reports and external observers (Section 4.3). Features from different combinations of the available modalities (i.e., Audio, Video, Physiology, Language) are provided as input to the model. Table 2 presents the outcome of the prediction experiments in terms of CCC and MSE metrics. Results suggest that the observer ratings are easier to estimate by the LSTM model using unimodal or multimodal features, compared to the self-reported ratings. The best performance (i.e., $CCC = 0.4688 \pm 0.247$, $MSE = 0.0988 \pm 0.141$) is obtained when audio-visual features

Table 1: Aggregated Pearson's correlation coefficient (r) between bio-behavioral features from different modalities and ratings of stress obtain from self-reports and raters.

Modality	Feature	Self	Observer
Audio	F3 frequency	0.1234*	0.5577*
	F2 frequency	0.1234*	0.5564*
	F0 frequency	0.1119*	0.5116*
	Shimmer	0.1017*	0.4439*
	Jitter	0.0789*	0.3282*
Video	AU25 (Lips part)	0.0919*	0.3639*
	Head rotation (Y-axis)	0.0484*	0.1720*
	AU01 (Inner Brow Raiser)	0.0382*	0.1413*
	Gaze angle (X-axis)	0.0446*	0.1385*
	AU09 (Nose Wrinkler)	0.0387*	0.1174*
	Head rotation (X-axis)	0.0338*	0.1144*
	AU07 (Lid Tightener)	0.0160*	0.1028*
Physiology	Mean SCL	0.0564*	0.2242*
	Mean HR	0.0132*	0.1777*
	Max HR	0.0152*	0.1743*
	SCR Frequency	0.0159*	0.1026*
Language	Nonfluencies	0.1915*	0.2703*
	Informal language	0.1669*	0.2589*
	Words per sentence	0.1951*	0.2209*
	Clout	-0.0498	0.0780
	Filler words	0.1337*	0.0716
	Word count	-0.0971	-0.1413*

*: p <0.05

fused with linguistic features are employed in estimating the observer ratings of perceived stress. Meanwhile, the self-reported ratings of felt stress are estimated the best in terms of CCC (i.e., $CCC = 0.2172 \pm 0.205, MSE = 0.5962 \pm 0.729$) by the model that combines linguistic features with audio-visual features, and in terms of MSE (i.e., $CCC = 0.2079 \pm 0.186$, $MSE = 0.5801 \pm 0.682$) by the model combining physiological features with audio-visual ones. These observations are further confirmed by performing one-way ANOVA test on CCC values obtained for different cases. Significant differences are found for CCC from different modalities for both self (i.e., $F(9, 90) = 95.9, p < 0.01$) and observer ratings (i.e., $F(9, 90) = 595.7, p < 0.01$). Post-hoc Tukey HSD tests indicate that best combination (A+V+L) for observer ratings significantly outperforms other combinations. An independent t-test is conducted between the best cases for self and observer ratings. CCC for observer ratings is significantly higher (i.e., $t(18) = -48.5, p < 0.01$) than self ratings. The inclusion of the physiological feature does not improve the prediction performance of the observer rating. This is expected as the raters only relied on the audio-visual recording while rating their perceived stress of the participants. On the other hand, physiological indices are known to capture the participants' reactivity when stressed, and therefore, they can be used to improve the estimation of the self-reported ratings of stress. Among the unimodal features, acoustic features are found to be the best performing features in estimating both self-reported and observer ratings of stress. The inclusion of linguistic features to the acoustic features seems to improve the performance for the observer ratings. The visual features further improve the predictive performance for observer ratings when added to the acoustic and linguistic measures. These suggest that external observers rely on the multimodal content of the video when rating, while self observers tend to rely mostly on acoustic and linguistic features.

6 Discussion

In this work, we investigate the interplay between self-reported ratings of felt stress and external observers' ratings of perceived stress in terms of their association with multimodal bio-behavioral signals. Although prior work examined the mismatch between self and observer ratings [10, 42, 59], this has not been examined for time-continuous ratings in the context of stress. Using data from 223 Q&A exchanges from 31 participants' mock job interviews, we pose three research questions to address this knowledge gap. In **RQ1**, we investigate the inter-rater reliability of the ratings obtained from the four raters in our study and determine how the reliability of the observer rating affects its association with self ratings. Time-continuous observer ratings of perceived stress obtained from the raters exhibit low to moderate inter-rater reliability (i.e., $r = 0.282$). This is not unprecedented as rating affect content is a complex process, influenced by ambiguity and subjectivity in perceiving emotion [19, 40]. The time-continuous nature of the ratings introduces additional complexity to the rating process. The exchanges in the 'High Reliability' group exhibit higher reliability (i.e., $r = 0.453$) compared to the remaining exchanges, which attests to the ambiguity in the process of rating perceived stress. There is also a significant difference between exchanges in 'High Reliability' and 'Low Reliability' groups in terms of their association with self-reported ratings (i.e., $r = 0.216$ and $r = 0.015$ for 'High Reliability' and 'Low Reliability' group, respectively). Although the overall correlation between self and observer ratings is low in our work, similar to prior work [14, 42], these findings indicate that certain exchanges exhibit less ambiguity compared to others. When interpreting these findings through the lens of Brunswik's lens model [9], it could imply that exchanges categorized in the 'High Reliability' group may present proximal cues (i.e., transmitted cues perceived by the raters) that are easily decoded and interpreted by the raters. These proximal cues likely convey similar information to the distal cues (i.e., how cues are encoded by the individuals experiencing the stressor), resulting in a high degree of association between self and observer ratings. On the contrary, the proximal and distal cues within the 'Low Reliability' group might present high ambiguity, leading to a very weak association between self and observer ratings.

Furthermore, we highlight the difference in the nature of ratings obtained from self and observer. Data from our study suggest that participants generally consider themselves more stressed than they are perceived by others for the majority of the interview duration (i.e., 60%). Self ratings exhibit more extreme values compared to the corresponding observer ratings, a trend observed also in prior research [10, 34, 59]. This mismatch might have been caused by the ambiguity of distal cues for the extreme values that might have increased the complexity of the decoding process for the raters. Next, **RQ2** examines how bio-behavioral features from different modalities (i.e., acoustic, visual, linguistic, physiological) exhibit

Table 2: Prediction performance of the LSTM model in estimating self and observer ratings of stress using features from different modalities (i.e., acoustic (A), visual (V), physiological (P), linguistic (L)). CCC ($\mu \pm \sigma$) and MSE ($\mu \pm \sigma$) are reported as the evaluation metrics.

Modality	CCC ($\mu \pm \sigma$)		MSE ($\mu \pm \sigma$)	
	Self	**Observer**	**Self**	**Observer**
A	0.1987 ± 0.188	0.3916 ± 0.286	0.6285 ± 0.751	0.1186 ± 0.167
V	0.1141 ± 0.117	0.3166 ± 0.228	0.6761 ± 0.896	0.1218 ± 0.151
P	0.0848 ± 0.134	0.2016 ± 0.152	0.7126 ± 0.876	0.1510 ± 0.192
A + L	0.2002 ± 0.190	0.4396 ± 0.257	0.6278 ± 0.750	0.1102 ± 0.157
V + L	0.1658 ± 0.184	0.3840 ± 0.212	0.6118 ± 0.716	0.1047 ± 0.123
P + L	0.1338 ± 0.153	0.2872 ± 0.145	0.6437 ± 0.787	0.1184 ± 0.162
A + V	0.2027 ± 0.172	0.4092 ± 0.272	0.6243 ± 0.827	0.1045 ± 0.115
A + V + L	**0.2172 ± 0.205**	**0.4688 ± 0.247**	0.5962 ± 0.729	**0.0988 ± 0.141**
A + V + P	0.2079 ± 0.186	0.3939 ± 0.257	**0.5801 ± 0.682**	0.1125 ± 0.146
A + V + P + L	0.1886 ± 0.192	0.4448 ± 0.241	0.5979 ± 0.758	0.1045 ± 0.150

association with moment-to-moment ratings of stress from self and observer. The acoustic features depict the highest association with observer ratings, followed by the visual and linguistic features. This suggests that as the raters watched the videos of the exchanges to rate the perceived stress, they might have focused heavily on the acoustic, visual, and linguistic cues. Conversely, the self ratings exhibit the highest correlations with linguistic features, followed by acoustic features, and very low correlations with visual features. This indicates that the participants might have focused more on the content of the interviews while rating their felt stress retrospectively, instead of relying on the visual cues.

Finally, we inspect the variation of the prediction performance of ML model in estimating time-continuous ratings of stress from self and observer as part of **RQ3**. We find that observer ratings are better estimated by ML models than self ratings, and this is the case across all modalities. This suggests that perceived stress might be easier to be modeled by multimodal indices, compared to felt stress [58, 59]. ML models trained and tested using observer ratings perform the best when acoustic, visual, and linguistic features are used. This result comes as no surprise as these features exhibit higher association with observer ratings. Also, these are the modalities that the raters perceived as proximal cues. Physiological features did not contribute to improved performance for observer ratings, as raters did not have access to these cues. However, the addition of physiological features with audio-visual features resulted in a slightly better performance for estimating self ratings. Physiological indices act as distal cues, as they are more related with felt stress, due to their association with individuals' reactivity to their exposure to stressful situation [8, 62]. The variation of the results for self and observer ratings across different modalities suggests that the choice of feature modality and rating source has important implications for the design of continuous stress detection systems. Self ratings are found to be more difficult to estimate compared to observer ratings, which is consistent with prior work [1, 14]. However, self ratings were obtained from different participants while observer ratings were obtained from the same raters who have had the chance to review all participants [59]. This might have introduced more subjectivity from individual differences in the self ratings compared to the observer ratings, which adds to the complexity of the estimation of self rating by ML model, as we used a participant-independent (i.e., leave-one-subject-out) experimental design. A participant-dependent framework might be a better option in training ML models to estimate self ratings of stress, which can serve as a future research direction.

Results from this study highlight the importance in understanding the nuances of time-continuous rating of stress from self-reports and multiple observers. Different elements of affective processes are captured while obtaining the ratings of felt stress and perceived stress from self and observer, respectively. This results in the differences in ratings obtained from two sources which in turn, can affect the ML models being used in estimating moment-to-moment stress ratings. Therefore, understanding the difference in the perception of stress in terms of multimodal bio-behavioral signals is needed to ensure effective continuous stress detection. Results from the current study come with some limitations. The job interview as a stress-inducing task is not as well-defined and constrained as other stressors (e.g., mental arithmetic task, cold pressor) commonly used in prior work [1, 6]. This might have introduced additional subjectivity in the time-continuous ratings. Next, the number of raters used as external observers is not very high, which might have decreased the generalizability of the result. Finally, the linguistic features had lower temporal resolution than features from other modalities, which makes it difficult to perform a uniform comparison in terms of their association with ratings of stress.

7 Conclusion

This work investigates how the perception of stress differs across time-continuous ratings from self-report and multiple observers. Association between self and observer ratings has been explored and the effect of inter-rater reliability on this association has been investigated. Findings indicate that observer ratings of stress are better correlated with multimodal indices and these ratings are better estimated by ML models. As part of future work, we plan to examine the individual differences of the participants and the raters, and inspect their effect on the mismatch between time-continuous ratings of stress obtained from self and observer.

Acknowledgments

The authors would like to acknowledge the support of the National Science Foundation (#1956021) for this work.

References

[1] Jonathan Aigrain, Michel Spodenkiewicz, Severine Dubuisson, Marcin Detyniecki, David Cohen, and Mohamed Chetouani. 2016. Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing* 9, 4 (2016), 491–506.

[2] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* 18, 3 (2018), 91–93.

[3] Nalini Ambady, Mark Hallahan, and Robert Rosenthal. 1995. On judging and being judged accurately in zero-acquaintance situations. *Journal of personality and social psychology* 69, 3 (1995), 518.

[4] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In *International Conference on Advances in Computer Entertainment Technology*. Springer, 476–491.

[5] W Arthur Jr. 2017. An unproctored internet-based test of general mental ability. *A validation report. College Station, TX: Author* (2017).

[6] Alice Baird, Andreas Triantafyllopoulos, Sandra Zänkert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, Brigitte M Kudielka, et al. 2021. An evaluation of speech-based recognition of emotional and physiological markers of stress. *Frontiers in Computer Science* 3 (2021), 750284.

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[8] Ralph R Behnke, Chris R Sawyer, and Paul E King. 1987. The communication of public speaking anxiety. *Communication Education* 36, 2 (1987), 138–141.

[9] Egon Brunswik. 1956. *Perception and the representative design of psychological experiments*. Univ of California Press.

[10] Carlos Busso and Shrikanth S Narayanan. 2008. The expression and perception of emotions: comparing assessments of self versus others.. In *Interspeech*. 257–260.

[11] Federico Cabitza, Andrea Campagner, and Martina Mattioli. 2022. The unbearable (technical) unreliability of automated facial emotion recognition. *Big data & society* 9, 2 (2022), 20539517221129549.

[12] H Calsbeek, M Rijken, GPB Henegouwen, and J Dekker. 2003. Factor structure of the Coping Inventory for Stressful Situations (CISS-21) in adolescents and young adults with chronic digestive disorders. *The Social Position of Adolescents and Young Adults With Chronic Digestive Disorders. Utrecht: NIVEL* (2003).

[13] CamNtech. 2024. *Actiheart 5*. https://www.camntech.com/actiheart-5/

[14] Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, Peter W Foltz, Alex S Cohen, Terje B Holmlund, and Brita Elvevåg. 2018. Modeling Self-Reported and Observed Affect from Speech.. In *Interspeech*. 3653–3657.

[15] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2011. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine* 16, 2 (2011), 279–286.

[16] Sheldon Cohen, Ronald C Kessler, and Lynn Underwood Gordon. 1997. *Measuring stress: A guide for health and social scientists*. Oxford University Press on Demand.

[17] Christine P Dancey and John Reidy. 2007. *Statistics without maths for psychology*. Pearson education.

[18] Leonard R Derogatis and Rachael Unger. 2010. Symptom checklist-90-revised. *The Corsini encyclopedia of psychology* (2010), 1–2.

[19] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18, 4 (2005), 407–422.

[20] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[21] Empatica. 2024. *E4 wristband*. https://www.empatica.com/en-gb/research/e4/

[22] George S Everly, Jr, Jeffrey M Lating, George S Everly, and Jeffrey M Lating. 2019. The anatomy and physiology of the human stress response. *A clinical guide to the treatment of the human stress response* (2019), 19–56.

[23] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[25] Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software* 2, 1 (2014), e5. https://doi.org/10.5334/jors.ar

[26] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[27] Matt J Gray, Brett T Litz, Julie L Hsu, and Thomas W Lombardo. 2004. Psychometric properties of the life events checklist. *Assessment* 11, 4 (2004), 330–341.

[28] Paul Grossman. 1983. Respiration, stress, and cardiovascular function. *Psychophysiology* 20, 3 (1983), 284–300.

[29] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 573–581.

[30] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 697–706.

[31] hrvanalysis. 2024. hrvanalysis 1.0.0 documentation. https://aura-healthcare.github.io/hrv-analysis/

[32] Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2020. MuSE: a multimodal dataset of stressed emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1499–1510.

[33] Madhu Kalia. 2002. Assessing the economic impact of stress [mdash] The modern day hidden epidemic. *Metabolism-clinical and experimental* 51, 6 (2002), 49–53.

[34] Hyunji Kim, Stefano I Di Domenico, and Brian S Connelly. 2019. Self–other agreement in personality reports: A meta-analytic comparison of self-and informant-report means. *Psychological science* 30, 1 (2019), 129–138.

[35] Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.

[36] Iulia Lefter, Gertjan J Burghouts, and Léon JM Rothkrantz. 2015. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing* 7, 2 (2015), 162–175.

[37] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.

[38] Yuliya Lutchyn, Paul Johns, Mary Czerwinski, Shamsi Iqbal, Gloria Mark, and Akane Sano. 2015. Stress is in the eye of the beholder. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 119–124.

[39] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[40] Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The MSP-conversation corpus. *Interspeech 2020* (2020).

[41] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. 2013. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing* 31, 2 (2013), 137–152.

[42] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing* 12, 2 (2018), 479–493.

[43] Iftekhar Naim, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing* 9, 2 (2016), 191–204.

[44] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia* 16, 4 (2014), 1018–1031.

[45] Ehsanul Haque Nirjhar, Md Nazmus Sakib, Ellen Hagen, Neha Rani, Sharon Lynn Chu, Winfred Arthur, Amir H Behzadan, and Theodora Chaspari. 2022. Investigating the interplay between self-reported and bio-behavioral measures of stress: A pilot study of civilian job interviews with military veterans. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[46] Matthias Norden, Oliver T Wolf, Lennart Lehmann, Katja Langer, Christoph Lippert, and Hanna Drimalla. 2022. Automatic detection of subjective, annotated and physiological stress responses from video data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[47] Vedhas Pandit and Björn Schuller. 2019. The many-to-many mapping between the concordance correlation coefficient and the mean square error. *arXiv preprint arXiv:1902.05180* (2019).

[48] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[49] Maili Pörhölä. 1997. Trait anxiety, experience, and the public speaking state responses of Finnish university students. *Communication research reports* 14, 3 (1997), 367–384.

[50] Andrew Raij, Patrick Blitz, Amin Ahsan Ali, Scott Fisk, Brian French, Somnath Mitra, Motohiro Nakajima, Minh Hoai Nguyen, Kurt Plarre, Mahbubur Rahman, et al. 2010. mstress: Supporting continuous collection of objective and subjective measures of psychosocial stress on mobile devices. *ACM Wireless Health 2010*

San Diego, California USA (2010).

[51] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[52] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 896–903.

[53] Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40, 1-2 (2003), 227–256.

[54] Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan. 2019. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360* (2019).

[55] N Clayton Silver and William P Dunlap. 1987. Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology* 72, 1 (1987), 146.

[56] Charles D Spielberger. 1983. State-trait anxiety inventory for adults. (1983).

[57] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. 5–14.

[58] Khiet P Truong, Mark A Neerincx, and David A Van Leeuwen. 2008. Assessing agreement of observer-and self-annotations in spontaneous multimodal emotion data. (2008).

[59] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech communication* 54, 9 (2012), 1049–1063.

[60] Khiet P Truong, David A van Leeuwen, Mark A Neerincx, and FM Jong. 2009. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. (2009).

[61] David Wainwright and Michael Calnan. 2002. *Work stress: The making of a modern epidemic*. McGraw-Hill Education (UK).

[62] Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir H Behzadan, and Theodora Chaspari. 2022. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1168–1182.

[63] Zoom. 2024. Video Conferencing, Cloud Phone, Webinars, Chat, Virtual Events. https://zoom.us/