This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

Modeling Gold Standard Moment-to-Moment Ratings of Perception of Stress from Audio Recordings

Ehsanul Haque Nirjhar, *Student Member, IEEE,* and Theodora Chaspari, *Member, IEEE*

Abstract—Enabling continuous and unobtrusive monitoring of stress is essential for delivering personalized stress interventions at opportune moments. To achieve automatic stress detection on a time-continuous basis, reliable moment-to-moment ratings of stress are required. However, the current research lacks a large-scale multimodal dataset that provides time-continuous ratings of perceived stress. Existing datasets mainly consist of single-valued self-reported ratings obtained after the stress-inducing task or rely on audio-visual recordings to capture moment-to-moment ratings from multiple annotators. The collection of time-continuous ratings of stress based solely on audio recordings has not been extensively explored. In this paper, we introduce an updated version of the publicly available VerBIO dataset that contains moment-to-moment ratings of perceived stress from multiple annotators. These annotators rated their perception of stress by listening to participants who had conducted a public speaking task. Time-continuous ratings of stress are obtained from four annotators using 22 hours of audio recordings from 339 public speaking sessions performed by 53 individuals. These time-continuous ratings of stress perception were obtained from the annotators solely based on speech, without incorporating the visual modality as an expressive marker. We examine the reliability of the annotation scheme employed in this study and investigate the factors contributing to the observed variation in perceived stress among annotators. Next, we introduce an annotation fusion technique based on expectation-maximization to obtain a reliable gold standard rating by aggregating the ratings from multiple annotators. Results indicate that the proposed annotation fusion technique yields aggregated ratings that can be estimated more reliably using acoustic features compared to the ratings yielded from conventional annotation fusion techniques. The newly generated annotations are publicly available within the proposed updated version of the existing VerBIO dataset, facilitating research in the field of continuous stress detection.

Index Terms—Stress, speech, dataset, annotation fusion, moment-to-moment rating.

1 Introduction

Stress can be defined as an individual's physiological and psychological response towards a challenge coming from a threatening environment or situation [1], [2]. These challenges can include demanding work environments [3], tasks involving cognitive load [4], [5], physically burdensome activities [6], [7], or anxiety in interpersonal communication [8]–[10]. Stress exerts a significant influence on human performance. While exposure to a certain level of stress can enhance performance, high stress is known to decrease performance and cause burnout [11]. Prolonged exposure to stress can result in heightened levels of anxiety and fear, negatively impacting both physical and mental well-being. Long-term exposure to stress has been related to adverse health effects, such as cardiovascular diseases, mental health complications, and sleeping disorders [12], [13]. Due to its enduring impact on physical and psychological health, stress is often regarded as a modern epidemic [14], [15]. Consequently, the monitoring and detection of stress have become active areas of research in affective computing and human-computer interaction.

Stress, as an affective state, is associated with negative valence and high arousal, therefore it can be mapped in the top-left quadrant of the circumplex model of affect [16]. Stress is manifested via the physiological response of the autonomic nervous system (ANS) of the human body. The ANS comprises of two components–the sympathetic (SNS) and parasympathetic (PNS) nervous systems. The SNS is responsible for the "flight-or-fight" response under stress that causes physiological reactivity such as sweating or racing heart [17], [18]. Stress is also manifested via muscle activity that causes pupil dilation, change in facial expression, and increased vocal fold tension [19]–[21]. To enable continuous stress monitoring, it is beneficial to record these multimodal indices and further analyze them using signal processing and machine learning algorithms [22]–[24]. Prior work has produced several multimodal datasets [3], [10], [25] that have been valuable for the task of automatic stress detection. However, these datasets typically provide single-valued labels or ratings of stress obtained at the conclusion of the stress-inducing task.

To enhance the implementation of continuous stress detection on a more granular temporal level, it is necessary to obtain reliable moment-to-moment ratings of stress. While obtaining self-reported ratings of stress would be ideal, it is often burdensome and can be confounded by personal

- *E. H. Nirjhar is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, 77843 USA. E-mail: nirjhar71@tamu.edu*
- *T. Chaspari is with the Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309 USA. E-mail: theodora.chaspari@colorado.edu*

Manuscript received July 11, 2023; revised May 13, 2024.

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

subjectivity and recall bias [26]. To address this, multiple annotators are typically employed to rate their perception of the extent to which the target individual is stressed while observing the individual undergoing a stressful situation [6], [7], [27]. Annotators usually rely on audiovisual data either by observing the situation as it happens or by retrospectively reviewing the recorded videos. However, rating the perception of stress based solely on audio recordings is hardly explored. Once ratings are obtained from multiple annotators, they need to be aggregated to construct gold standard ratings [28], [29] that can act as an approximation of the latent ground truth rating of stress. Previous research has actively investigated this process focusing primarily on valence and arousal dimensions of affect [28], [30]–[33]. Yet, annotation fusion in the context of stress ratings remains unexplored, presenting an opportunity to construct a multimodal dataset with continuous ratings of stress.

The most commonly used techniques for fusing ratings of affect dimensions (i.e., valence, arousal) are feature-independent since they solely rely on the ratings provided by the annotators without considering any behavioral cues (e.g., acoustic, visual) present in the data that is observed by the annotators [28], [34]. Among these feature-independent annotation fusion techniques, calculating the mean rating across all annotators is widely employed in affect recognition [28], [35]. While this method is straightforward to implement, it does not account for the inter-individual differences among annotators, which can be addressed by assigning weights to each annotator via a weighted annotation fusion approach [29], [36].

Prior work on feature-independent methods has overlooked the potential reaction delay in annotator ratings due to the time lag between observing an affective state and providing the corresponding rating using an annotation tool. To overcome this challenge, prior work on feature-dependent annotation fusion techniques has primarily focused on aligning visual features extracted from the individual expressing the emotion with the moment-to-moment rating from multiple annotators who observed the individual [37], [38]. These feature-dependent techniques leverage both individual annotator ratings and observable features and have been developed using algorithms such as Canonical Time Warping (CTW) [39], expectation-maximization (EM) [38], [40], delay estimation [32], triplet embedding [41], [42]. Previous research on feature-dependent annotation fusion has predominantly focused on the video modality, as the visual features have shown stronger correlations with affect ratings compared to audio data, which typically exhibit weaker correlations [32], [37]. Consequently, in scenarios where acoustic features are only relied upon, feature-independent techniques may be considered more suitable for annotation fusion compared to feature-dependent ones.

Moreover, previous studies on annotation fusion techniques have taken into account the reaction lag of annotators but have overlooked the adjustment shift, which refers to the temporal shift that annotators introduce when refining and revising their ratings using continuous-time annotation tools [28], [43], as they try to compensate for the pre-existing reaction time lag in their annotation. Additionally, most of the existing work employs the valence-arousal space for obtaining the rating of the perceived stress from the annotators while they observe video recordings of individuals experiencing stressful situations. The exploration of rating stress through a single time-continuous scale is limited, especially when considering only speech (i.e., audio recording) as the observable expressive behavior. Furthermore, stress is recognized as a sparsely occurring state [5], [44], which means it does not conform to the commonly assumed Gaussian distribution when developing computational frameworks [38], [40]. The distribution of stress labels exhibits skewness [5], yet skewness is often disregarded to maintain simplicity in stress models and their implementation.

To bridge the existing gap in current literature, we investigate how continuous (i.e., moment-to-moment) ratings of perception of stress can be obtained from multiple annotators when tasked with rating stress based on speech as the sole marker of expressive behavior, in the absence of rich visual information. Speech was investigated because of its ability to effectively detect stress [20], [45], while depicting reduced privacy leakage concerns compared to video data [46], [47]. We inspect the reliability of the annotation scheme and delve into the factors contributing to variations in perceived stress among annotators. Next, motivated by previous work [33], [38], [40], we propose a modified EM-based technique for fusing these moment-to-moment ratings of perceived stress to obtain the gold standard rating in a feature-independent manner. The modifications we introduced include accounting for adjustment shifts in addition to previously studied reaction delays, as well as incorporating the Skew-Normal distribution [48], [49] to accommodate the skewness in stress ratings. Finally, to evaluate the effectiveness of our proposed modifications compared to conventional feature-independent annotation fusion methods, we conducted prediction experiments assessing reliability and generalization. We assume that the reliable gold standard moment-to-moment ratings would maximally correlate with the input features (e.g., speech, physiology), hence machine learning models would be able to learn the mapping more effectively. This would maximize the prediction performance of the regressors, which is a commonly used assumption in related literature [41], [50], [51]. For this purpose, we have used VerBIO dataset [52], a publicly available dataset containing approximately 22 hours of audio data captured during a stress-inducing task involving public speaking. In addition, we inspect a multi-task learning framework to estimate individual annotator ratings which avoids annotation fusion and incorporates all available annotator ratings during training in an end-to-end manner. Its performance is compared with the predictive performance of the model trained with fused annotation to understand the necessity of obtaining gold standard ratings through annotation fusion and compare the performance of fused annotation with individual annotation. Moreover, we have examined how the fused gold standard time-continuous ratings of stress are associated with self-reported anxiety scores as well as the physiological indices available in the dataset. Finally, we investigate the performance of the proposed annotation fusion method on a synthetic dataset with a known ground truth developed in [41], [53], where multiple annotators performed two time-continuous rating tasks to rate the green color intensity values while watching a video of continuously changing green colored frames

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

TABLE 1
Summary of Existing Datasets Available for Stress Detection Research Where Labels for Stress Are Available Along With Multimodal Data. Modalities May Include Acoustic (A), Visual (V), Linguistic (L), And/or Physiological (P) Features.

Dataset	#Participants	Modalities	Annotation	#Annotators	Stressor task	Duration	Year
Driver Stress Data [6]	9	V, P	Continuous	2	Driving	36 hours	2005
SWELL-KW [3]	25	V, P	Single-valued	Self-reports	Knowledge work	75 hours	2014
WESAD [10]	15	P	Single-valued	Self-reports	Public speaking, Mental arithmetic	N/A	2018
AffectiveROAD [7]	10	V, P	Continuous	1	Driving	11 hours	2018
MuSE [25]	28	A, V, L, P	Single-valued	Self-reports	Monologue	10 hours	2020
Ulm-TSST [27]	69	A, V, L, P	Continuous	3	Public speaking	6 hours	2021
VerBIO	53	A, P	Single-valued [52] Continuous (this work)	Self-reports 4	Public speaking	22 hours	2022 2023

where actual intensity values of the video frames are known. The aim of this experiment is to analyze the efficacy of the proposed method in fusing annotations from multiple annotators to approximate the known ground truth and compare the performance with multiple feature-independent and feature-dependent annotation fusion techniques. The resulting gold standard moment-to-moment ratings of stress from the VerBIO data, obtained through the proposed method, along with the individual annotators ratings[1] will be added to the existing VerBIO dataset and become publicly available, facilitating research in the field of continuous stress detection and intervention (Appendix A).

The rest of this paper is organized as follows. Section 2 provides an overview of the available datasets for stress detection research and the previous work on annotation fusion. Section 3 outlines the process of obtaining moment-to-moment ratings of stress from multiple annotators using the VerBIO data. Section 4 describes the problem formulation, the proposed method of annotation fusion, and the evaluation methods that are employed to compare the proposed approach with existing methods. Next, Section 5 presents the results, and Section 6 provides the corresponding discussion of the findings and limitations of this work. Finally, the conclusion is provided in Section 7.

2 Related Work

2.1 Annotation Fusion for Aggregating Moment-to-Moment Affect Ratings

Utilizing multiple annotators to rate their perception of affect for a target individual can be a valuable approach to reduce bias stemming from the subjectivity of a single annotator. Several frameworks exist for obtaining moment-to-moment affect dimension ratings [26], [43], [54], [55]. Annotation fusion is employed to aggregate these moment-to-moment ratings from multiple raters, aiming to approximate the latent ground truth of affect, and the aggregated ratings are referred to as the gold standard ratings [28], [29], [56]. However, obtaining a reliable gold standard time-continuous rating that captures the subjective nature of human state from multiple annotators is not straightforward, given the inter-individual variability in perception. Consequently, this area has been actively explored in previous research. Findings from prior work point towards two different approaches—feature-independent and feature-dependent annotation fusion methods.

The feature-independent approach aims to fuse moment-to-moment affect ratings from multiple annotators using various statistical properties of the obtained ratings, without relying on the associated multimodal features. In many studies, the arithmetic mean of continuous-time ratings has been commonly employed for annotation fusion [28], [35]. However, alternative methods have been explored to assign weights to the annotators rather than treating them equally. One such approach is the Evaluator Weighted Estimator (EWE) [36], which calculates the weight of annotators by quantifying the similarity between individual annotator ratings and mean rating. Metallinou *et al.* proposed constructing a union set of selective annotators who exhibit a certain level of inter-annotator agreement with other annotators, and obtained the mean rating of this selective set [57]. The authors evaluated this method on the USC-CreativeIT data [34]. Their choice of inter-annotator agreement metric was Pearson's r with a threshold value of $r = 0.45$. Two additional feature-independent fusion techniques, namely the Dynamic Time Warping (DTW) [58] Based Barycenter Averaging (DBA) and the Rater Aligned Annotation Weighting (RAAW) [27], have been introduced in the MuSe-toolbox [29]. In DBA, a gold standard rating is estimated which acts as a barycenter that minimizes the DTW distance between multiple ratings. The RAAW incorporates Generic CTW to account for the reaction delay and EWE to capture annotator reliability through a weighted mean. Although the feature-independent methods are straightforward to compute, they are primarily evaluated empirically (e.g., via visual inspection of the original and fused ratings). A formal quantitative evaluation of these methods has not been conducted in prior work.

Feature-dependent methods leverage multimodal features, typically video, that demonstrate a high correlation with the target state being estimated through annotation fusion. Mariooryad *et al.* [32] estimated annotator-dependent lags (i.e., reaction delays) by maximizing the mutual information between visual features and individual ratings. Gupta *et al.* [38] and Ramakrishna *et al.* [40] treated individual annotators as a Linear Time Invariant (LTI) filter, where the ratings are modeled as a noisy and distorted version of the latent gold standard affect content. The latent gold standard affect content can be retrieved through an EM-based optimization that accounts for annotator-specific reaction delays. This approach of EM optimization extends the seminal work by Raykar *et al.* [33] to time-continuous ratings, providing a general framework for fusing single-valued labels. Kossaifi *et al.* [39] utilized CTW for annotation fusion while curating the SEWA DB dataset. Booth *et al.* [41], [53] employed a triplet embedding scheme to fuse time-continuous annotations using a synthetic dataset (i.e., green

1. The moment-to-moment ratings of stress used in this paper are available at https://tinyurl.com/mr2w592n

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

intensity task data) with known ground truth. Using the same dataset, Mundnich *et al.* [42] introduced a novel way of obtaining annotator ratings, where annotators were asked to compare between three samples and choose the closest pair from the triplet, instead of a time-continuous annotation process. Instead of obtaining fused gold standard ratings, Huang *et al.* [59] utilized a multi-task learning by setting up a crowd layer consisting of ratings from multiple annotators and learning the input features' mapping to all ratings. The majority of these methods present a comparative evaluation of the annotation fusion methods through stress estimation experiments. These methods predominantly rely on visual features due to their inherent association with affect ratings. Acoustic features, are rarely utilized for feature-dependent annotation fusion due to their weak individual correlation profile [32]. Furthermore, these methods do not account for the adjustment shift introduced by the annotators.

2.2 Current Datasets Available For Stress Detection

Over the past two decades, a considerable amount of research has been conducted in the area of multimodal detection of stress [24], [27], [45]. However, only a small portion of these studies have made their data publicly available. Table 1 provides an overview of the currently available datasets that include labels or ratings of perception of stress along with multimodal indices (e.g., acoustic, visual, linguistic, physiological). The ratings in the previous studies have been obtained either through self-reports or external annotators. In terms of temporal granularity, these labels can be single-valued acquired as a post-stressor task rating, or moment-to-moment captured in a continuous manner.

In one of the earliest studies on stress detection using physiological signals, Healey *et al.* [6] utilized a driving task to induce stress in participants. The video recordings of the driving task were independently coded by two raters, and a continuous stress metric was derived based on a filtered version of the frequencies of observable driving events sampled at a rate of 1 Hz. All modalities except the stress metric from this work are publicly available. A similar approach was taken by Haouij *et al.* [7] in the development of the AffectiveROAD dataset. In this study, an experimenter sat in the rear seat while the participant drove, and provided a continuous stress metric by assessing the complexity of the driving scene and the perceived workload. This metric was validated by the participants at the end of the experiment, and adjusted, if necessary. Most recently, two widely used datasets for stress detection include the SWELL-KW [3] and WESAD [10]. SWELL-KW involved stress-inducing knowledge work tasks (e.g., writing reports, making presentations, checking emails). Participants self-reported their level of perceived stress after performing each stressor task using a visual analog 10-point Likert scale. In WESAD, participants engaged in public speaking followed by mental arithmetic tasks. Their perceived stress levels were assessed through validated questionnaires (e.g., Positive and Negative Affect Schedule (PANAS) [60], State-Trait Anxiety Inventory (STAI) [61]). Jaiswal *et al.* [25] introduced a multimodal dataset where participants performed a monologue in response to a set of emotion-evoking questions and self-reported stress levels using the Perceived Stress Scale (PSS) [1], [62]. In a more recent work, Stappen *et al.* [27] presented the Ulm-TSST dataset, where three annotators rated their perception of participants' stress levels during a public speaking task. The rating was provided in a moment-to-moment manner and the annotators rated both the arousal and valence dimensions while watching the video recordings of the public speaking sessions. Yadav *et al.* developed the VerBIO dataset [52] where participants completed multiple public speaking tasks in a longitudinal study and self-reported their anxiety using the STAI questionnaire. This discussion underscores the scarcity of datasets that include both multimodal indices and continuous ratings of stress for facilitating continuous stress detection. To address this gap in the literature, a larger multimodal dataset with moment-to-moment ratings of stress is necessary.

2.3 Proposed study contributions

The major contributions of this work to the current state of knowledge are as follows:

- We examine the feasibility of obtaining annotators' perception of stress based on audio recordings, and investigate the level of agreement or disagreement among annotators in relation to the acoustic properties of the corresponding audio and the characteristics (e.g., variance) of the individual annotations.
- We propose a modified version of the annotation fusion technique presented by Gupta *et al.* [38] that enables feature-independent annotation fusion, incorporates an adjustment shift in addition to the previously proposed reaction delays (Section 1), and accommodates the inherent skewness in stress labels.
- We conduct a formal evaluation of the proposed feature-independent annotation fusion method through a prediction experiment and compare the performance of the proposed fusion method in automatically detecting stress from acoustic features with conventional feature-independent techniques [29], [36], [57].
- We introduce an enhanced version of the VerBIO dataset [52], which now includes moment-to-moment ratings of stress provided by individual annotators' ratings and the gold standard ratings obtained using the proposed annotation fusion technique (Appendix A).

3 Data Description

3.1 VerBIO Dataset

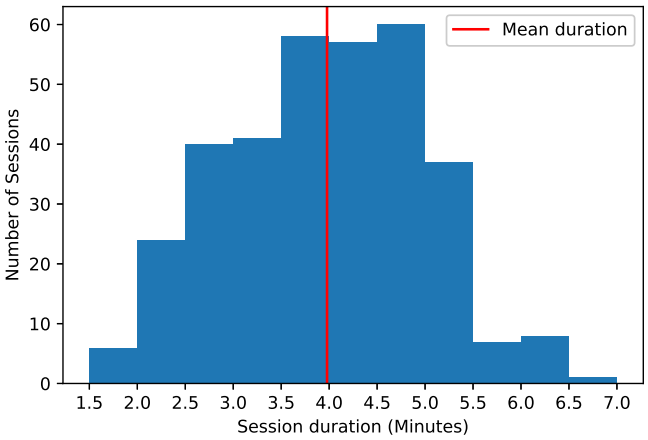
VerBIO dataset [52] is a publicly available multimodal signal corpus of individuals' affective responses during a public speaking task. This dataset consists of multimodal signals (e.g., speech, physiology) and self-reported measures (e.g., personality traits, anxiety rating) from 55 participants in a longitudinal research study, where the participants performed 344 public speaking tasks. Participants were asked to complete 10 public speaking sessions (i.e., 2 with real-life audience, 8 in virtual reality (VR)), distributed over 4 days, within a span of 2 weeks. In each session, participants were given a news article about a general topic and were provided 10 minutes to prepare a speech. Next, they spoke about the topic in front of a real or virtual

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

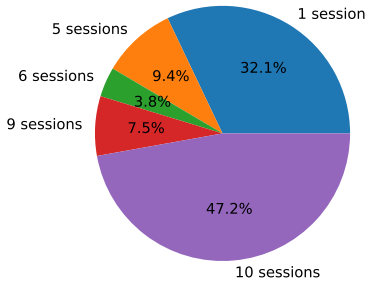
This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

audience, depending on the day of the study and the study protocol. After completing each session, participants self-reported their level of anxiety by responding to the 20-item State Anxiety Enthusiasm (SAE) questionnaire [63]. Physiological reactivity of the participants was captured using the Empatica E4 wristband [64] which recorded the electrodermal activity (EDA) at a sampling rate of 4 Hz. The audio data of each public speaking session has been recorded via a Creative lavalier microphone at a sampling rate of 16 kHz and 16-bit encoding. Audio recordings of two participants who completed only the first session were removed as their speeches were unintelligible due to the presence of excessive background noise. The total number of available audio files in the dataset is 340 which comes from 53 participants. One file has been excluded from this work due to technical issues resulting in ratings from 339 audio files. This results in a large audio corpus of an approximate duration of 22 hours. An overview of the dataset used in this work is presented in Table 2. Fig. 1(a) presents the distribution of the duration of each public speaking session (i.e., audio file). The smallest sessions are approximately 2 minutes long, while the longest sessions are over 6 minutes.

It is to be noted that only 25 out of 53 participants (i.e., 47.2%) completed all 10 sessions. Out of the total participants, 17 participants (i.e., 32.1%) dropped out after completing only the first session. The rest of the participants dropped out in the middle of the study. Therefore, the number of available sessions (i.e., audio files) is different for each participant, as demonstrated Fig. 1(b).



(a) Distribution of audio duration in VerBIO



(b) Participant-wise audio availability

Fig. 1. Properties of the Audio Data in the VerBIO Dataset.

3.2 Obtaining Moment-To-Moment Ratings of Stress

We have used the audio files of the public speaking sessions in the VerBIO dataset to obtain continuous time ratings of stress from multiple annotators. Instead of collecting continuous-valued ratings in a time-continuous manner [26], [54], we have aimed to obtain a continuous rating of stress from the audio signals on a 5-point Likert scale. The reasoning behind this design decision is that emotion is often considered ordinal in nature, and perceiving and labeling the change of stress between consecutive timestamps in an audio stream is easier than continuously providing ratings [65]–[67]. The definition of each point of the Likert scale and its associated perceived stress level is exhibited in Table 3. The choice of scale is consistent with prior work where a similar range of annotations has been employed to rate affect [68], [69] and stress [25], [45]. The Noldus Observer XT software [43] is selected as the annotation tool for our purpose since it is widely used in behavioral and social sciences [70], [71]. The points of the Likert scale has been mapped to the corresponding numeric button on the keyboard (i.e., 1-5). While listening to the audio, whenever annotators perceive a change in how stressed the participant was, they press the button associated with the current level of their perceived stress for the considered participant. The level stays the same until the annotator perceive another change in stress, and therefore, select another level. For all audio files, the initial rating is manually set to 1 (i.e., No Stress), which annotators have the option to change.

In order to obtain moment-to-moment ratings of stress, we have recruited four annotators (one male, three female) who are undergraduate students in psychology. Two of the annotators were college seniors, while the remaining two were college juniors at the time when they started working on this task. Two annotators had 2 years of experience in emotion annotation, and the others had 1 year of experience. In terms of ethnicity, three annotators were White/Caucasian and the other annotator was Hispanic/Latino. Each annotator is assigned an identity (ID), namely R1, R2, R4, and R5. Before starting the tasks, the annotators have been trained to work with the Observer XT software. Fig. 2 shows the typical layout of the Observer XT tool the annotators used. Annotators are instructed to first listen to the entire audio file and then rate the perceived change in stress in a moment-to-moment basis while listening to the audio the second time. To reduce individual response delays, annotators are asked to listen to each audio two to three times and modify their annotations accordingly. All audio files have been rated by all the annotators. Continuous ratings from the annotators are sampled at 1 Hz,

TABLE 2
Overview of VerBIO Dataset

Factor	Value
Number of Participants	53
Number of Female Participants	23
Average age (years)	22.32
Number of Sessions	339
Total audio duration (hh:mm:ss)	22:28:31
Average session length (m:ss)	3:58
Average number of sessions per participant	6.4

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

TABLE 3
Definition on Perceived Stress Level for Annotation

Point on Likert Scale	Perceived Stress Level
1	No Stress
2	Low Stress
3	Moderate Stress
4	High Stress
5	Extremely High Stress

which results in four time series of ratings per audio file. Note that the annotators are not asked to rate stress on a second-to-second basis; the 1 Hz sampling rate is used by the software so that it can register any potential changes that the annotators mark with a 1 second resolution. The ratings from each annotator are re-scaled from $[1, 5]$ range to $[0, 1]$ range for further processing. Inter-annotator agreement is measured through Spearman's correlation, ρ, between ratings from two annotators, which can be aggregated to obtain average ρ for all possible pairs of annotators of an audio file using the Fisher's z-transformation [72].

Fig. 3 presents examples of continuous-time ratings of stress from four annotators for two audio files. Fig. 3(a) shows that the annotators agree in most of the transitions (Average Spearman's $\rho = 0.541$) and there are slight delays in the transition among the annotators. The trends are similar among the ratings, despite the different amplitude of the ratings. Although this audio exhibits high agreement, there are more local variations by R4 between 60-80 seconds compared to others. This highlights the inter-individual differences between annotators. On the other hand, Fig. 3(b) shows a scenario where the annotators exhibit low inter-rater agreement (Average Spearman's $\rho = -0.029$). In this case, R4 and R5 showed opposite trends, while R1 did not exhibit variance. By examining these scenarios, it is clear that conventional approaches (i.e., arithmetic mean, weighted mean) might not result in a reliable fused annotation. This calls for a sophisticated method of annotation fusion to obtain the fused ratings of perceived stress.

3.3 Green Intensity Task Data: Evaluation with a Synthetic Dataset

In order to understand how good the fused ratings are in approximating ground truth, a dataset with known ground truth is needed. However, VerBIO dataset [52] does not contain the objective ground truth of stress rating. Meanwhile, Booth *et al.* [41], [53] introduced the green intensity task data, a synthetic dataset with known ground truth that can be used to evaluate the goodness of annotation fusion techniques in fusing multiple ratings to approximate the ground truth. The dataset contains time-continuous ratings from 10 annotators who were shown two videos of green color with varying intensity and were asked to rate the intensity in a time-continuous manner while watching the videos. This was done to mimic the subjective nature of affect annotation but the ground truth of the green intensity values of the video frames were known *a priori* unlike affect content (e.g., stress, valence, arousal). In the first video (Task A), the intensity values of green-colored frames changed at different times and speeds without abrupt transitions, while in the second video (Task B), the green intensity values featured a slow oscillation. The ratings are sampled at a rate of 1 Hz. We use the annotator ratings and the known ground truth to examine the efficacy of the proposed annotation fusion method.

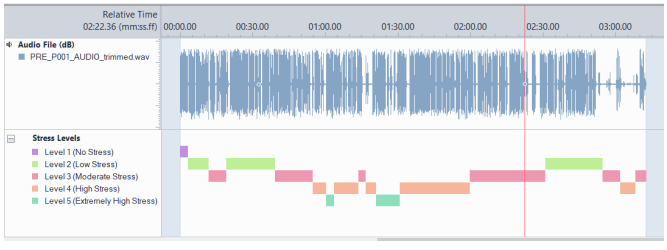
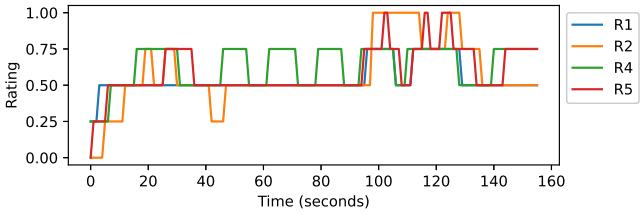
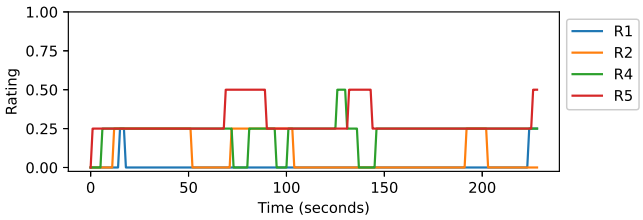


Fig. 2. Observer XT tool layout for moment-to-moment stress annotation.



(a) Participant: P044, Session: TEST08, $\rho = 0.541$



(b) Participant: P037, Session: POST, $\rho = -0.029$

Fig. 3. Moment-to-moment ratings of stress by four annotators for (a) a high inter-rater agreement session, and (b) a low inter-rater agreement session.

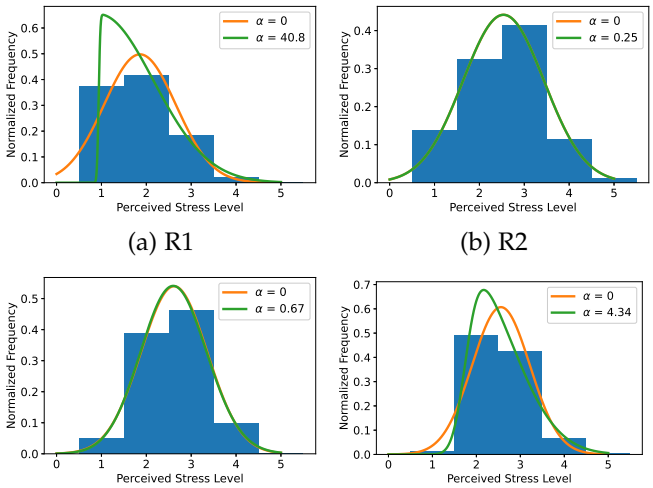
4 Methodology

4.1 Motivation

A common theme among the feature-dependent annotation fusion methods is the choice of feature modality. Visual features (e.g., facial landmark points, action unit intensity) are found to be most correlated with the affect dimensions being rated, therefore, these features have been used in previous methods to estimate the gold standard continuous-time rating of affect [32], [39]. In our case, the VerBIO dataset does not contain any visual information such as the videos of the public speaking sessions. The annotators are asked to rate the stress in a time-continuous manner while listening to the audio. Preliminary analysis show that the extracted acoustic feature time series exhibit poor correlation (i.e., mean correlation in the range $[-0.165, 0.142]$) with the individual annotator ratings (i.e., when computing correlations between individual feature time-series and individual annotator ratings), which is consistent with prior work [32], [37]. Therefore, incorporating the acoustic features in the annotation fusion might result in an unreliable gold standard of continuous-time rating of stress, thus the acoustic features are not used in the proposed annotator fusion method.

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502



(a) R1 (b) R2 (c) R4 (d) R5

Fig. 4. Annotator-wise distribution of moment-to-moment ratings of stress and the fitted Gaussian ($\alpha = 0$) and Skew-Normal ($\alpha \neq 0$) distributions, where α refers to the shape parameter of the Skew-Normal distribution.

Next, visual inspection of the annotator ratings (Fig. 3) shows that there are temporal shifts in the trends among annotators, which might have originated from the reaction delays or the adjustment shifts from multiple rounds of annotation. Prior work addressed the annotator-specific delays, but the adjustment-related shift has not been discussed. Hence, in order to exploit the strengths of both types of annotation fusion methods (i.e., feature-independent and feature-dependent) and address the abovementioned issues, we propose a feature-independent EM-based optimization scheme for annotation fusion, expanding upon the method proposed by Gupta *et al.* [38] and Ramakrishna *et al.* [40].

In addition, annotator ratings for commonly rated affect dimensions (e.g., valence, arousal) in public datasets [27], [39] tend to exhibit Gaussian distribution, as neutral emotions are more prominent than their positive and negative extremes. However, stress does not exhibit such two-sided extremes and the likelihood of stress has been found to be skewed [4], [5], which is not effectively modeled through a Gaussian distribution. The Skew-Normal distribution is a modification of the Gaussian distribution, introducing an extra parameter, known as shape parameter, α, that determines the degree of skewness. Negative values of α correspond to a left-skewed distribution, while positive values indicate a right-skewed distribution. When the shape parameter α is equal to 0, this results in the Gaussian distribution [48], [49]. Fig. 4 shows the distribution rating of each annotator over all sessions. R1 and R5 exhibit Skew-Normal distribution instead of Gaussian. Therefore, in order to account for this observed skewness, we propose a skew-normal approximation of annotation distribution, in contrast to the Gaussian distribution used in [38], [40].

4.2 Problem Formulation

Let S represent the number of sessions included in the dataset, and N be the number of annotators who have completed the moment-to-moment perceived stress rating for each session while listening to the associated audio file. The duration of s-th session is T_s, where $s = 1, \ldots, S$. The latent gold standard rating of stress for the s-th session is $\boldsymbol{a}^s \in \mathbb{R}^{T_s}$ and the rating obtained from the n-th annotator is $\boldsymbol{a}_n^s \in \mathbb{R}^{T_s}$. Motivated by prior work [38], [40], we consider each annotator acting as a Linear Time-Invariant (LTI) filter, since we can assume that each annotator observes the latent gold standard stress rating by listening to the corresponding audio file, and provides a noisy and distorted version of this gold standard rating. Therefore, we model the behavior of each annotator as a combination of an LTI filter of length W, an additive bias, and an annotator-specific noise vector of length T_s. The LTI filter is centered on the current timestamp and contains W_d windows on the left to account for the annotator-specific reaction delay, and W_a windows on the right for adjustment shift. Therefore, the total filter length is $W = W_d + W_a + 1$. We can represent the filter coefficient vector $\boldsymbol{d}_n \in \mathbb{R}^W$ as $\boldsymbol{d}_n = [d_n(-W_d), d_n(-(W_d - 1)), \ldots, d_n(-1), d_n(0), d_n(1), \ldots, d_n(W_a - 1), d_n(W_a)]$. This is a generalized version of the LTI filter used in [38], [40], where $W_a = 0$. The additive bias term and the noise vector are represented as $d_n^b \in \mathbb{R}$ and $\boldsymbol{\phi}_n \in \mathbb{R}^{T_s}$. Based on the problem formulation, $\boldsymbol{a}_n^s$ can be expressed as–

$$\boldsymbol{a}_n^s = (\boldsymbol{d}_n * \boldsymbol{a}^s) + (d_n^b \times \mathbf{1}_s) + \boldsymbol{\phi}_n \tag{1}$$

In (1), the operator '$*$' refers to the convolution operation, and $\mathbf{1}_s \in \mathbb{R}^{T_s}$ refers to a column vector of ones. The variable $\boldsymbol{\phi}_n$ is traditionally assumed as a zero mean Gaussian noise i.e., $\boldsymbol{\phi}_n \sim \mathcal{N}(0, \sigma_n^2 \times \boldsymbol{I}_s)$, where σ_n^2 is annotator-specific variance that is assumed constant over all sessions for the simplified calculation. Therefore, the likelihood an annotator's rating given the parameters of the Gaussian distribution for session s can be modeled as in (2)–

$$p(\boldsymbol{a}_n^s | \boldsymbol{a}^s, \boldsymbol{d}_n, d_n^b, \sigma_n) \sim \mathcal{N}((\boldsymbol{d}_n * \boldsymbol{a}^s) + (d_n^b \times \mathbf{1}_s), \sigma_n \times \boldsymbol{I}_s) \tag{2}$$

Here, in addition to the Gaussian noise, we propose a Skew-Normal approximation of noise, $\boldsymbol{\phi}_n \sim \mathcal{SN}(0, \sigma_n^2 \times \boldsymbol{I}_s, \alpha_n \times \mathbf{1}_s)$, where $\mathcal{SN}(\cdot, \cdot, \cdot)$ refers to the Skew-normal distribution, to account for the skewness present in the perceived stress rating of annotators. This approximation adds an additional shape parameter $\alpha_n \in \mathbb{R}$, which is annotator-specific, time-independent, and constant over all sessions. Hence, for the skew-normal approximation, the likelihood presented in (2) is modified as follows–

$$p(\boldsymbol{a}_n^s | \boldsymbol{a}^s, \boldsymbol{d}_n, d_n^b, \sigma_n, \alpha_n) \sim \mathcal{SN}((\boldsymbol{d}_n * \boldsymbol{a}^s) + (d_n^b \times \mathbf{1}_s), \sigma_n^2 \times \boldsymbol{I}_s, \alpha_n \times \mathbf{1}_s) \tag{3}$$

In order to simplify the optimization process, the skew-normal distribution in (2) can be expressed as a function of Gaussian distribution and a complementary error function [49], [73], as shown in Appendix C. This is demonstrated in (4) –

$$p(\boldsymbol{a}_n^s | \boldsymbol{a}^s, \boldsymbol{d}_n, d_n^b, \sigma_n, \alpha_n) \sim \mathcal{N}((\boldsymbol{d}_n * \boldsymbol{a}^s) + (d_n^b \times \mathbf{1}_s), \sigma_n \times \boldsymbol{I}_s) \times \operatorname{erfc}\left(\frac{\alpha_n}{\sqrt{2}\sigma_n}(\boldsymbol{a}_n^s - \boldsymbol{d}_n * \boldsymbol{a}^s - d_n^b \times \mathbf{1}_s)\right) \tag{4}$$

Given the annotator rating $\boldsymbol{a}_n^s$ and the initial values for the annotator-specific LTI filter parameters $(\boldsymbol{d}_n, d_n^b, \sigma_n, \alpha_n)$, the latent gold standard rating of perceived stress, $\boldsymbol{a}^s$, for all N annotators and S sessions can be obtained by maximizing the likelihood (i.e., minimizing the loss) over

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

the entire dataset. For the Gaussian approximation in (2), the associated loss function is shown in (5). The derivation of the loss function is provided in Appendix B and the optimization problem is solved via the EM algorithm [74] (Section 4.3.2).

$$L_{\mathcal{N}} = \sum_{s=1}^{S}\sum_{n=1}^{N}(T_s \log(\sqrt{2\pi}\sigma_n) + \frac{1}{2\sigma_n^2}\|\boldsymbol{a}_n^s - \boldsymbol{d}_n * \boldsymbol{a}^s - d_n^b \times \mathbf{1}_s\|_2^2) \tag{5}$$

Similarly, the loss function for the skew-normal approximation can be expressed as (6) and the details of derivation are available in Appendix C. The corresponding optimization problem is solved via the EM algorithm (Section 4.3.3).

$$L_{\mathcal{SN}} = L_{\mathcal{N}} - \sum_{s=1}^{S}\sum_{n=1}^{N}\log(\text{erfc}(\frac{\alpha_n}{\sqrt{2}\sigma_n}\|(\boldsymbol{d}_n * \boldsymbol{a}^s) + (d_n^b \times \mathbf{1}_s) - \boldsymbol{a}_n^s\|)) \tag{6}$$

4.3 Annotation Fusion Methods

4.3.1 Initialization

In this work, we experiment with two different types of initialization of the gold standard rating $\boldsymbol{a}^s$ to investigate how weighing annotators differently can affect the final outcome. These two types of initialization are referred to as 'mean initialization' and 'selective initialization'. In the first case, we assign the arithmetic mean of all annotators' ratings to a session as the initial value of $\boldsymbol{a}^s$. In the latter case, we calculate the Spearman's correlation ρ between all possible combinations of annotator pairs for a session and select the set of annotators who exhibit Spearman's ρ over a threshold value with at least one other annotator. The arithmetic mean over the set of selected annotators is assigned as the initial value of $\boldsymbol{a}^s$. If the corresponding set is empty, the mean value over all annotators is assigned. The initialization process of $\boldsymbol{a}^s$ has changed the ratings from discrete variables to continuous variables in the range of $[0, 1]$, which is a common practice in prior work [68], [75]. Therefore, we have chosen to use continuous distributions (i.e., Gaussian, Skew-Normal) for noise vector $\boldsymbol{\phi}_n$, instead of discrete distributions (i.e., Binomial, Geometric). We assigned $\rho > 0.4$ as the threshold value for the 'selective initialization', which is chosen based on prior work [76] and preliminary inspection of our data.

4.3.2 Gaussian Approximation

The loss function for the Gaussian approximation in (5) can be solved analytically for both the E-step and the M-step of the optimization, as shown in [40]. During the E-step, the current estimate of the gold standard rating of stress is obtained, while the annotator-specific parameters are calculated during the M-step. This iterative process continues until the rate of decrease in the loss value reaches a certain tolerance level (Appendix D). The analytical solution for the parameters is as follows-

$$\boldsymbol{a}^s = (\sum_{n=1}^{N} F_n^T F_n)^{-1}(\sum_{n=1}^{N}(F_n^T \boldsymbol{a}_n^s - F_n^T(d_n^b \times \mathbf{1}_s))) \tag{7}$$

$$\boldsymbol{d_n} = (\sum_{s=1}^{S}(A^s)^T A^s)^{-1}(\sum_{s=1}^{S}((A^s)^T \boldsymbol{a}_n^s - (A^s)^T(d_n^b \times \mathbf{1}_s))) \tag{8}$$

$$d_n^b = \frac{\sum_{s=1}^{S}(\mathbf{1}_s^T \boldsymbol{a}_n^s - \mathbf{1}_s^T F_n \boldsymbol{a}^s)}{\sum_{s=1}^{S} T_s} \tag{9}$$

$$\sigma_n = \sqrt{\frac{\sum_{s=1}^{S}\|\boldsymbol{a}_n^s - \boldsymbol{d}_n * \boldsymbol{a}^s - d_n^b \times \mathbf{1}_s\|_2^2}{\sum_{s=1}^{S} T_s}} \tag{10}$$

In the above, F_n and A^s are the matrix representation of $\boldsymbol{d}_n$ and $\boldsymbol{a}^s$, respectively, where $\boldsymbol{d}_n * \boldsymbol{a}^s = F_n \boldsymbol{a}^s = A^s \boldsymbol{d}_n$.

4.3.3 Skew-Normal Approximation

The loss function for the skew-normal approximation in (6) does not have a closed-form solution for the optimization problem. Therefore, we used gradient descent for minimizing the loss in both the E- and M-steps. During the E-step the loss function is minimized over all annotators for a given session through the SGD optimizer with a 0.001 learning rating and for 200 epochs. The current estimate of the gold standard rating of stress is obtained at this step. Meanwhile, during the M-step, the loss function is minimized over all sessions for a given annotator via the Adam optimizer with a 0.005 learning rate and for 100 epochs in order to compute the annotator-specific parameters. In both steps, batch size is set as the size of the entire training data and an early stopping is performed if no performance improvement over 15 consecutive epochs is observed or the decrease in loss in consecutive epochs falls below a tolerance value of 10^{-5}. Both steps are repeated for 50 iterations with an early stopping criteria based on the tolerance value of 0.0005 for the overall loss function. The aforementioned optimization parameters are chosen empirically.

4.4 Evaluation

In order to evaluate the proposed annotation fusion method and the effect of various design choices, we train a machine learning model to estimate the fused rating of stress using acoustic features. We have divided the VerBIO dataset into training and testing partitions. As there might be multiple sessions available for each participant, we have decided to stratify the dataset so that a participant's data is available in either the training or the testing set only. For each session, the Spearman's ρ over all possible annotator pair combinations is measured, and a selective set of annotators who exhibit $\rho > 0.4$ with at least one other annotator is formed. Participants who have a non-empty selective set for more than half of their total sessions are selected for the training partition. The rest of the participants are placed in the test set. This is conducted so that the training set contains reliable annotations so that the EM algorithm can effectively learn the LTI filter parameters. This results in 203 sessions from 32 participants in the training set, and 136 sessions from the remaining 21 participants in the test set. The list of the participant IDs used in the training and test sets is provided in Appendix A. The sessions of the training set are used in estimating the annotator-specific parameters (i.e., $\boldsymbol{d}_n$, d_n^b, σ_n, α_n) as well as the fused ratings of stress (i.e., $\boldsymbol{a}^s$) in the training set. The annotator-specific parameters obtained in the previous step are used for calculating the fused rating of stress in the test set.

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

4.4.1 Feature Extraction

We use the OpenSMILE [77] to extract acoustic features from the audio signals. We extract 88 features of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [78], as this feature set is concise and is widely used in the affective computing [25], [27]. It consists of frequency-related parameters (e.g., pitch, jitter, formant frequencies), energy estimates (e.g., shimmer, loudness, Harmonics-to-noise ratio), and spectral features (e.g., Alpha ratio, Hammarberg Index, MFCCs). These features are computed over a 600 ms window and then averaged over the 1 second segments of the public speaking sessions to match the sampling rate of the moment-to-moment ratings of stress with the acoustic measures. These acoustic features are used to examine potential associations between acoustic parameters and perceived stress rating by the different annotators. These features are used as the input to the machine learning algorithm that estimated the gold standard stress ratings that resulted from the annotation fusion.

4.4.2 Estimation of Gold Standard Stress Rating

A long short-term memory (LSTM) neural network is used for estimating the gold standard stress rating based on the aforementioned acoustic features (Section 4.4.1). The LSTM model is deemed appropriate for our purpose since it was used as a baseline model in the MUSE-STRESS subchallenge of the MuSe 2021 challenge [27], where the goal was to predict the valence and arousal in stressful situations using multimodal features. The LSTM model consists of 4 hidden layers each containing 64 hidden states and a fully connected layer before the output layer. The 88-dimensional acoustic feature time series is given as input and the goal of the model was to estimate the fused ratings of stress. The model is trained using the training data with a learning rate of 0.0002 for 200 epochs, with an early stopping if no performance improvement over 15 consecutive epochs was observed. Next, the model is used in estimating the gold standard time-continuous ratings of stress in the test set. The experiment is repeated 20 times for each configuration to account for any randomness. The Concordance Correlation Coefficient (CCC) loss is used to optimize the LSTM model and CCC is used as the evaluation metric to measure the predictive performance of the model on the testing data. We compare the proposed feature-independent annotation fusion technique with previous frequently used annotation fusion methods, namely Mean [28], DBA [29], EWE [36], RAAW [27], and Selective [76] through their predictive performance measured by the CCC. The underlying assumption is that the reliable gold standard moment-to-moment ratings would maximally correlate with the acoustic features and it will help the LSTM model to learn the mapping more effectively. Therefore, the prediction performance of the LSTM model would be higher for better and more reliable gold standard ratings. In addition, we evaluate how different design decisions (i.e., initialization, length of delay and adjustment windows, distribution approximation) affect the predictive performance of the model.

4.4.3 Association of Gold Standard Rating with Self-reports and Physiological Signals

In order to evaluate the reliability of the gold-standard time-continuous ratings of stress obtained by the proposed method, we investigate how these ratings are associated with other measures of stress, such as self-reported anxiety scores and physiological indices. First, we examine the degree of association between self-reported ratings of stress in the VerBIO dataset and moment-to-moment fused ratings of stress. However, self-reported moment-to-moment ratings of stress are not available in the VerBIO dataset. Instead, a single-valued SAE score is available for each session. The SAE score ranges between 20 and 100, where a higher SAE score indicates higher self-reported stress. The difference in temporal resolution renders it difficult to perform any direct comparison between the self-report and fused annotation of stress. Therefore, the Pearson's r is measured between the SAE score and the mean fused ratings of each session to examine their association, similar to prior work [79].

Next, to examine the association between fused time-continuous ratings and physiological signals, we utilize the EDA signals of each session available in the dataset. Skin Conductance Level (SCL) (i.e., tonic part of EDA) is extracted from the EDA signal using NeuroKit toolbox [80], as SCL is a well-known indicator of stress and it is known to increase when individuals experience higher stress [8], [21], [81]. Pearson's r correlation between the mean SCL signal and mean fused ratings of stress over 10 second window is measured for each session and the median value of the correlation coefficients is reported to exhibit the association.

4.4.4 Estimation of Individual Annotator Rating Instead of Annotation Fusion

Instead of performing annotation fusion, some of the prior work often relies on end-to-end learning by utilizing ratings available from all the annotators to train a machine learning model for affect detection [59]. The rationale behind this approach is that the annotation fusion step can be bypassed through the end-to-end learning, allowing the machine learning model to benefit from jointly learning the representation between the input features and the output ratings of all annotators. One of the common frameworks used for this purpose is multi-task learning, where the goal is to learn a feature embedding by maximizing the prediction performance of ratings from all annotators. To assess the effectiveness of constructing the gold standard rating through annotation fusion compared to such methods, we use a multi-task learning model to estimate individual annotator ratings (instead of the aggregate gold standard rating). We formulate a crowd layer similar to [59] by modifying the LSTM model to contain four separate fully connected layers followed by four output layers (i.e., each corresponding to an annotator), instead of one output layer (i.e., corresponding to a fused annotation). There is a fully connected layer after the hidden layers, which acts as a bottleneck and is shared by the crowd layer. The goal of the multi-task learning model is to maximize the prediction performance for estimating the ratings provided by all annotators using the acoustic features as the input, without needing to perform any annotation fusion step. The CCC

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

loss is calculated for each output layer and the total loss is computed as the sum of CCC loss values from individual annotators. The CCC between the actual rating by each annotator and the predicted rating by the multi-task model for the same annotator is reported as the evaluation metric that measures the predictive performance of the multi-task model. The prediction experiment is repeated 20 times by keeping the same model hyperparameters and the set of random seed values to ensure consistency in the evaluation.

4.4.5 *Effectiveness in Approximating Known Ground Truth*

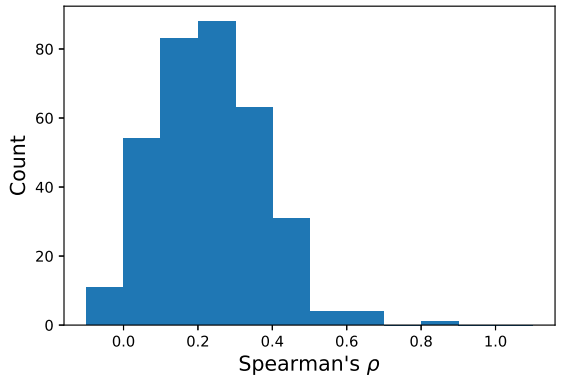
In order to understand how close the fused ratings are to the actual ground truth, we have used the green intensity task data (i.e., Task A, Task B), a synthetic dataset developed in [41], [53] (Section 3.3). We have used the moment-to-moment ratings from the annotators to construct the gold standard ratings for both tasks of the dataset. The fused ratings are obtained by both the proposed method, the baseline feature-independent annotation fusion methods (Section 4.4.2), and previously proposed feature-dependent methods (i.e, Triplet embedding [41], EM [38], Triplet comparison [42]). We compute the Mean Squared Error (MSE) and Pearson's r between the gold standard rating and the known ground truth rating. These metrics evaluate the performance of the annotation fusion methods in approximating the known ground truth.

5 Results

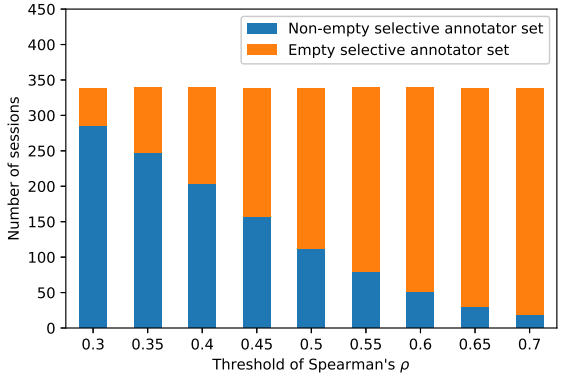
5.1 Inspecting the Quality of Moment-to-Moment Annotation of Stress

5.1.1 *Analyzing Inter-annotator Agreement*

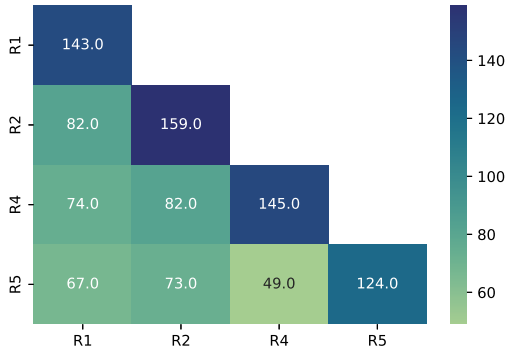
Here, we provide an analysis of the inter-annotator agreement that is achieved in our task. Spearman's ρ is computed for all six possible annotator pair combinations for a session, and they are aggregated to obtain the average ρ value for that session using Fisher's z-transformation. Fig. 5(a) presents the distribution of ρ values obtained from 339 sessions, where the average value of ρ is 0.237. It is to be noted that majority of the sessions exhibit positive average correlation among annotator ratings, over half of the sessions indicate $\rho > 0.2$, and 30% sessions have moderate to strong (i.e., $\rho >= 0.3$) inter-annotator agreement [82], [83]. However, this distribution presents the aggregated Spearman's ρ for each session that can be affected by some pairs of annotators exhibiting lower agreement compared to other pairs. In order to understand the inter-annotator agreement for pairs of annotators at the session level, a selective set of annotators who exhibit Spearman's ρ over a threshold value with at least one other annotator is constructed for each session. A non-empty selective set would indicate that at least two annotators exhibited some degree of agreement when rating stress. An empty selective annotator set would suggest no or low agreement. Fig. 5(b) exhibits the effect of the threshold $\rho = 0.3, 0.35, \ldots, 0.7$ on the distribution of the number of sessions based on the selective annotator set. The number of sessions with a non-empty selective annotator set decreases as ρ increases. With a lower threshold, more sessions contain a non-empty selective set, but this might cause reduced inter-annotator agreement. Conversely, the



(a) Distribution of Spearman's ρ for all sessions



(b) Distribution of #sessions for different threshold ρ



(c) Heatmap of #sessions with $\rho > 0.4$

Fig. 5. Inter-annotator agreement of the audio files in the VerBIO dataset.

selection of a higher threshold would ensure the quality of inter-annotator agreement in the sessions at the cost of a reduced number of sessions. Based on inspection of the results (Fig. 5(b)), $\rho = 0.4$ has been selected as the threshold for constructing the selective annotator set for each session, which ensured a more balanced distribution between the number of sessions with high inter-annotator agreement (i.e., 203 sessions where at least one annotator exhibits Spearman's $\rho > 0.4$ with at least one other annotator) and low inter-annotator agreement (i.e., 103 sessions where no annotator exhibits Spearman's $\rho > 0.4$ with another annotator). We have used the same threshold to construct the training and test partition for evaluation (Section 4.4). Note, that the average value of Spearman's ρ for sessions in the high agreement group is 0.505, which is significantly

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

higher than the average Spearman's ρ of 0.237 when all the sessions are considered. This is consistent with prior work [57], [76] where a similar ρ value has been used as a threshold for inter-annotator agreement.

Fig. 5(c) presents the contribution of each annotator in developing the selective set. The diagonal elements of the heatmap show the number of sessions for which the corresponding annotator (marked in the row or column) exhibited $\rho > 0.4$ with at least one other annotator. The non-diagonal elements of the heatmap visualize the number of sessions where the annotator of the corresponding row exhibited $\rho > 0.4$ with the annotator of the corresponding column. From Fig. 5(c), it is evident that among all the possible rater pairs, (R1, R2) and (R2, R4) agreed the most in rating the perceived stress, while (R4, R5) agreed the least. Overall, in terms of annotator agreement, R2 exhibited the highest with other annotators, followed by R1, R4, and lastly R5. This suggests that not all annotators rated stress in the same fashion, and that the simple averaging of annotators' ratings might not yield reliable fused ratings of stress.

5.1.2 Effect of Perceived Stress on Inter-Annotator Agreement

We examine how the overall variation of perceived stress rated by the annotators affects the inter-annotator agreement. For this purpose, we measure the mean and standard deviation of perceived stress ratings from the mean rating of all annotators for each session. We compare these statistics between the 203 sessions with high inter-annotator agreement (i.e., sessions with non-empty selective set; Section 5.1.1) and 136 with low inter-annotator agreement (i.e., sessions with empty selective set; Section 5.1.1) via an independent *t*-test. Results indicate that annotators tend to agree more when both the mean and the standard deviation of perceived stress are higher (Table 4(a)). Annotators tend to exhibit more agreement when they perceive a higher variation of stress during a session ($t(337) = 10.84,\ p < 0.0005$) and overall higher mean stress ($t(337) = 4.12,\ p < 0.0005$). Sessions with a lower variation of perceived stress present lower inter-annotator agreement which suggests that the subtle variation of stress might be difficult to perceive.

5.1.3 Effect of Acoustic Features on Inter-Annotator Agreement

Next, we investigate the effect of acoustic features on the inter-annotator agreement in perceived ratings of stress. For this purpose, we examine seven acoustic features from the eGeMAPS feature set [78] that are commonly associated with stress [21] and can be intuitively interpreted, namely, fundamental frequency (F0), $1^{st}/2^{nd}/3^{rd}$ formant frequencies (F1-F3), jitter, shimmer, and frequency of loudness peaks of speakers. These features are averaged over the entire session. We compare the aforementioned features between the 203 sessions with high inter-annotator agreement and 136 with low inter-annotator agreement via independent *t*-tests, similar to Section 5.1.2. In order to compensate for the multiple comparisons [84], we present the result without Bonferroni correction (i.e., $p < 0.05$) and with Bonferroni correction (i.e., $p < 0.0005$) in Table 4(b). The high and low inter-annotator agreement sessions depict significant differences in terms of formant frequencies (i.e., F1-F3), jitter, and frequency of loudness peaks of the speakers. Pitch or shimmer did not yield any significant difference. Sessions with high inter-annotator agreement tend to contain speech with lower F1-F3 and jitter compared to the low agreement counterparts. This indicates that annotators tend to agree in terms of the perceived rating of stress when jitter or F1-F3 are lower. An increase in jitter and an increase in F1-F3 indicates an increase in psychological stress. However, the aforementioned features are also dependent on the overall vocal characteristics of a person. If a speaker depicts high jitter and F1-F3 overall, the annotators might not be able to perceive when the speaker is under stress, since these acoustic parameters are always high. This phenomenon portrays the complexity of the task when annotators are asked to rate using the audio modality only.

TABLE 4
t-test Results Identifying Significant Differences between Sessions with High Inter-Annotator Agreement and Low Inter-Annotator Agreement Based on the Statistics of Perceived Ratings of Stress and Acoustic Features.

(a) Statistics of Perceived Ratings of Stress

Feature	High ρ Sessions $\mu \pm \sigma$	Low ρ Sessions $\mu \pm \sigma$	*t*-test Results
Mean	2.484 ± 0.43	2.284 ± 0.45	$t(337) = 4.12^{**}$
Standard Deviation	0.394 ± 0.09	0.299 ± 0.05	$t(337) = 10.84^{**}$

*: p <0.05, **: p <0.0005

(b) Statistics of Acoustic Features

Feature	High ρ Sessions $\mu \pm \sigma$	Low ρ Sessions $\mu \pm \sigma$	*t*-test Results
F0	151.20 ± 37.92	159.43 ± 37.03	$t(337) = -1.52$
F1	579.26 ± 75.32	596.61 ± 72.59	$t(337) = -2.10^{*}$
F2	1475.9 ± 93.9	1503.8 ± 77.1	$t(337) = -2.87^{*}$
F3	2367.2 ± 133.7	2416.5 ± 102.4	$t(337) = -3.63^{**}$
Jitter	0.029 ± 0.005	0.030 ± 0.006	$t(337) = -2.29^{*}$
Shimmer	1.25 ± 0.17	1.27 ± 0.17	$t(337) = -0.78$
Loudness Peaks/sec	3.62 ± 0.40	3.71 ± 0.77	$t(337) = -2.19^{*}$

*: p <0.05, **: p <0.0005

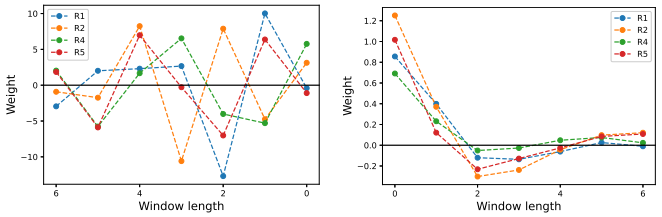
5.2 Examining the Outcomes of Annotation Fusion

5.2.1 Selecting the Delay and Adjustment Window Length

In order to identify the proper window length for both annotator-specific reaction delay and adjustment shift of the proposed annotation fusion algorithm (Section 4.3), we initially obtain the fused annotation of perceived stress for two cases, namely only delay window (i.e., $W_a = 0$) and only adjustment window (i.e., $W_d = 0$). A 6-second window length was used for this purpose since it is unlikely for both reaction delay and adjustment shift to depict a duration longer than 6 seconds [32]. Fig. 6 shows the annotator-specific filter weights for both cases. The filter weights for all annotators tend to saturate after 3 seconds when only the adjustment shift window is used. This suggests that the annotators are mostly able to compensate for reaction time lag in their annotation while they are refining their ratings, but they might make smaller adjustment shift (i.e., 1-3 seconds) that can be considered as overcompensation. On the contrary, the filter weights that account for the reaction delay depict variation among annotators throughout the

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502



(a) Delay window only (b) Adjustment window only

Fig. 6. Annotator-specific filter weights for (a) delay window only, and (b) adjustment window only cases for Gaussian approximation and selective initialization. The left graph is flipped to emulate the filter window on the left for delay window only cases.

considered 6 seconds. R1 shows a peak at 1 second and then the weight diminishes, while R2 and R4 have maximum weight after 3 seconds. Prior work suggests that annotator-specific reaction delay can be in a range of 2-6 seconds [32], [38]. Therefore, we have selected 4-6 second length for the delay window and 1-3 second length for the adjustment window. We have varied these window lengths to obtain results based on the fused ratings of stress for different configurations.

5.2.2 Evaluating the Baseline Methods

Results of the prediction experiment using fused ratings of perceived stress obtained from the baseline methods are presented in Table 5. The prediction performance is expressed as the combination of the mean and standard deviation of CCC metric over 20 repetitions for each method. The Mean label approach performed the best among the baseline methods, while DBA performed the worst. A one-way ANOVA test is conducted comparing the CCC values among the different baseline methods which reveals significant differences ($F(4, 95) = 74.488$, $p < 0.01$). The Tukey HSD test is performed as a post-hoc test for pairwise comparisons between the baseline methods. Results indicate that CCC for the Mean label approach is significantly higher than the other methods, while DBA shows significantly low CCC compared to the rest of the methods. There is no significant difference among CCC values obtained from EWE, RAAW, and Selective methods. Therefore, the proposed annotation fusion methods are compared against the performance of Mean label approach.

5.2.3 Evaluating the Proposed Method

Table 6 shows the result of the prediction experiment using ratings of perceived stress constructed by the proposed method. Different configurations are tested (i.e., Gaussian vs. Skew-Normal distribution approximation, Mean vs Selective initialization, and different window lengths for reaction delay and adjustment shift). Results indicate that CCC is comparatively lower when only the adjustment shift window is used (i.e., $W_d = 0$). This is expected as the annotator-specific reaction delays are not considered in this configuration. Meanwhile, the configuration that only accounts for the delay window (i.e., $W_a = 0$) usually performs better, as this configuration considers the reaction delay. This suggests that taking into account the reaction delay is more beneficial when fusing stress ratings from multiple annotators compared to the adjustment shift. Incorporating both the reaction delay and adjustment shift results in better predictive performance in the majority of cases. The best performance is obtained by the fused rating with a 6 second delay window (i.e., $W_d = 6$) and 2 second adjustment window (i.e., $W_a = 2$), resulting in a 9 second window (i.e., $W = 9$) for both Gaussian and Skew-Normal distribution assumptions and selective initialization.

TABLE 5
Prediction Performance Measured by CCC ($\mu \pm \sigma$) Obtained from Fused Ratings of Perceived Stress Using Baseline Methods.

Baseline method	CCC
Mean [28]	$\mathbf{0.2987 \pm 0.031}$
EWE [36]	0.2608 ± 0.020
DBA [29]	0.1610 ± 0.020
RAAW [29]	0.2659 ± 0.019
Selective [76]	0.2610 ± 0.038

We perform an independent *t*-test between the CCC values in the best-performing baseline method (i.e., Mean) and the best-performing configuration of the proposed method (i.e., $W_d = 6$, $W_a = 2$). Results indicate that the CCC values obtained from the proposed method are significantly higher than the baseline method ($t(38) = -3.46$, $p < 0.01$). This suggests that incorporating both reaction delay and adjustment shift in fusing ratings from multiple annotators helps in constructing reliable ratings instead of merely averaging the ratings. The delay window length of this configuration is similar to prior work [32], [68].

In general, the selective initialization performed better than the mean initialization. Instead of initializing the gold standard rating of perceived stress, $\boldsymbol{a}^s$ with the mean rating at the beginning of the EM optimization, the selective initialization provides a weighted mean of the most reliable annotators (i.e., the ones who depicted Spearman's $\rho > 0.4$ for at least one other annotator) and empirically performs better than the mean initialization. Although the selective method does not outperform the mean rating as a baseline method, the incorporation of the selection set during the initialization phase yields better predictive performance as the EM optimization might have been able to capture the comparative weights of different annotators. Finally, ratings obtained from the Skew-Normal approximation perform significantly worse than the mean rating and the ratings from the Gaussian approximation. However, they show significantly higher CCC than the worst baseline method (i.e., DBA). The Gaussian distribution assumption appears to exhibit better generalization in the test data compared to mean ratings, since the corresponding fused ratings obtained the best-performing configuration (i.e., $W_d = 6$, $W_a = 2$). Therefore, in the publicly available version of the VerBIO dataset [52], we provide this setting as the gold standard rating of perceived stress in addition to the ratings from the individual annotators.

5.2.4 Association of Gold Standard Rating with Self-reports and Physiological Signals

Next, we explore how the fused time-continuous ratings of stress are associated with self-reported anxiety scores (i.e., SAE score) and physiological indices (i.e., SCL features). For this purpose, we choose the fused ratings obtained by the

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

TABLE 6
Prediction Performance Measured by CCC ($\mu \pm \sigma$) Obtained From Fused Ratings of Perceived Stress for Different Configurations of the Proposed Method.

Distribution	W_d	W_a	CCC (Mean initialization)	CCC (Selective initialization)
Gaussian	4	0	0.3216 ± 0.023	0.3117 ± 0.020
	5	0	$\mathbf{0.3299 \pm 0.025}$	0.3231 ± 0.027
	6	0	0.3169 ± 0.023	0.3131 ± 0.034
	0	1	0.2978 ± 0.026	0.3003 ± 0.024
	0	2	0.2952 ± 0.019	0.2837 ± 0.038
	0	3	0.2834 ± 0.028	0.2766 ± 0.042
	4	1	0.3178 ± 0.022	0.3035 ± 0.030
	4	2	0.3180 ± 0.029	0.3113 ± 0.031
	4	3	0.3187 ± 0.027	0.3027 ± 0.031
	5	1	0.3168 ± 0.029	0.3179 ± 0.030
	5	2	0.3048 ± 0.024	0.3131 ± 0.028
	5	3	0.3130 ± 0.028	0.3089 ± 0.024
	6	1	0.3049 ± 0.026	0.3120 ± 0.028
	6	2	0.3084 ± 0.015	$\mathbf{0.3320 \pm 0.029}$
	6	3	0.3250 ± 0.025	0.3124 ± 0.020
Skew-Normal	4	0	0.2101 ± 0.023	0.2239 ± 0.025
	5	0	0.1944 ± 0.024	0.2242 ± 0.028
	6	0	0.2197 ± 0.034	0.2187 ± 0.036
	0	1	0.2251 ± 0.026	0.2332 ± 0.030
	0	2	0.2292 ± 0.028	0.2329 ± 0.024
	0	3	$\mathbf{0.2408 \pm 0.025}$	0.2284 ± 0.027
	4	1	0.2016 ± 0.027	0.2273 ± 0.033
	4	2	0.2181 ± 0.023	0.2263 ± 0.037
	4	3	0.2062 ± 0.029	0.2427 ± 0.030
	5	1	0.2350 ± 0.020	0.2341 ± 0.021
	5	2	0.2171 ± 0.026	0.2419 ± 0.036
	5	3	0.2312 ± 0.023	0.2375 ± 0.029
	6	1	0.1702 ± 0.022	0.2117 ± 0.027
	6	2	0.2255 ± 0.033	$\mathbf{0.2440 \pm 0.023}$
	6	3	0.1810 ± 0.029	0.2044 ± 0.026

best-performing configuration (i.e., Gaussian distribution, Selective initialization, $W_d = 6$, $W_a = 2$) described in Section 5.2.3. Table 7(a) presents the Pearson's correlation coefficient (r) between the mean value of the fused ratings and the SAE scores of the corresponding sessions. Time-continuous ratings obtained by the proposed annotation fusion method exhibits higher correlation (i.e., $r = 0.1905$) with self-reported SAE scores, compared to all baseline methods, except DBA ($r = 0.1948$) which has slightly higher correlation coefficient. This might have been caused by the loss of temporal information due to the aggregation of fused ratings into mean values for comparison with SAE scores. Nevertheless, the prediction performance of the fused ratings obtained by DBA is the worst among the baseline methods. Table 7(b) shows the association of fused time-continuous ratings with SCL features, in terms of the median value of Pearson's r over all sessions. Moment-to-moment ratings of stress generated by the proposed fusion method yield higher correlation ($r = 0.1929$) than ratings obtained from baseline feature-independent annotation fusion techniques. These results suggest that the proposed annotation fusion method is able to produce reliable time-continuous ratings of stress that are better correlated with other measures of stress, such as self-reported scores and biomarkers.

TABLE 7
Correlation between Fused Time-continuous Ratings of Stress and Different Indicators of Stress (e.g., Self-report, Physiology).

(a) Pearson's r between the Mean Value of the Fused Ratings and the Self-Reported State Anxiety Enthusiasm (SAE) Score.

Annotation Fusion Method	Pearson's r
Mean [28]	0.1858
EWE [36]	0.1806
DBA [29]	**0.1948**
RAAW [29]	0.1809
Selective [76]	0.1670
Proposed method (Gaussian distribution, Selective initialization, $W_d = 6$, $W_a = 2$)	0.1905

(b) Median Value of Pearson's r between Fused Ratings and Skin Conductance Level (SCL) Features.

Annotation Fusion Method	Pearson's r
Mean [28]	0.1318
EWE [36]	0.1291
DBA [29]	0.1037
RAAW [29]	0.1324
Selective [76]	0.1389
Proposed method (Gaussian distribution, Selective initialization, $W_d = 6$, $W_a = 2$)	**0.1929**

5.2.5 *Estimation of Individual Annotator Rating Instead of Annotation Fusion*

Table 8 exhibits the prediction performance of the multi-task learning model using the crowd layer based on acoustic measures. The model performs the best in estimating the ratings from R1 (i.e., CCC (R1) = 0.2880 ± 0.019), while the worst prediction performance is obtained for the ratings from R2 (i.e., CCC (R2) = 0.1886 ± 0.021). Results exhibit that the CCC obtained for ratings of R1 is significantly lower than the best-performing setting of the proposed method ($t(38) = -5.68$, $p < 0.01$). Therefore, fused gold stan-

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

TABLE 8
Prediction Performance Measured by CCC ($\mu \pm \sigma$) for Ratings from Individual Annotators Using Multi-task Learning Framework.

Evaluation Metric	CCC ($\mu \pm \sigma$)
CCC (R1)	0.2880 ± 0.019
CCC (R2)	0.1886 ± 0.021
CCC (R4)	0.2222 ± 0.016
CCC (R5)	0.2835 ± 0.019

dard ratings obtained from the proposed annotation fusion method can be better estimated from acoustic features by an LSTM model compared to the individual ratings estimated by the multi-task learning framework. This highlights the efficacy of the proposed annotation fusion method compared to an end-to-end learning model that uses ratings from all available raters during the training step by incorporating a multi-task framework and does not perform any annotation fusion during the process.

5.2.6 Effectiveness in Approximating Known Ground Truth

Table 9(a) presents the results of the analysis for Task A of the green intensity task data, while Table 9(b) shows the same for Task B data. Results indicate that the proposed feature-independent annotation fusion method has yielded lower MSE (i.e., 0.0186 for Task A, 0.0078 for Task B) and higher Pearson's r (i.e., 0.8389 for Task A, 0.9648 for Task B) than the feature-independent baseline methods for both Tasks A and B. Moreover, the proposed feature-independent method has outperformed the feature-dependent triplet embedding method by the work that introduced the green intensity dataset [41] in both tasks. For Task A, our method performs worse in terms of MSE and Peason's r compared to the other feature-dependent methods presented (i.e, EM [38], Triplet Comparison [42]). However, the proposed method has resulted in Pearson's r similar to these methods for the Task B data. This might have been caused by the difference in the nature of the ground truth data between the two tasks. Task B featured a time-series with slowly varying green intensity values that mimics the progression of affect rating. Meanwhile, Task A contained time interval with constant values and fast-changing peaks, that might be difficult to estimate for the proposed method as the fused annotation is constructed based on neighboring values. In contrast, the triplet comparison approach proposed by [42] can potentially better handle abrupt changes in values within the time-series, since it relies on relative comparisons, thus performing well for Task A. The proposed feature-independent method performs better than the baseline feature-independent methods and one of the feature-dependent methods (i.e., the triplet embedding method [41]) in constructing the time-continuous gold standard rating from multiple annotators, when evaluated using data with known ground truth. The performance of the proposed method is worse than the remaining feature-dependent annotation fusion methods presented, which might stem from the different annotation process (i.e., time-continuous annotation approach vs. triplet comparison annotation approach) and the type of annotation fusion method (i.e., feature-independent vs feature-dependent).

TABLE 9
Mean Squared Error (MSE) and Pearson's r between the Ground Truth and the Gold Standard Rating Constructed by Different Annotation Fusion Methods Using the Green Intensity Task Data [41], [53].

(a) Task A

Method Type	Annotation Fusion Method	MSE	Pearson's r
Feature-independent	Mean [28]	0.0338	0.7816
	EWE [36]	0.0339	0.7816
	DBA [29]	0.0187	0.8001
	RAAW [29]	0.0346	0.7776
	Selective [76]	0.0385	0.7723
Feature-dependent	Triplet Embedding [41]	0.0364	0.7762
	EM [38]	0.0049	0.903
	Triplet Comparison [42]	0.0013	0.975
Feature-independent	Proposed method (Gaussian distribution, Mean initialization, $W_d = 0$, $W_a = 2$)	0.0186	0.8389

(b) Task B

Method Type	Annotation Fusion Method	MSE	Pearson's r
Feature-independent	Mean [28]	0.0084	0.9540
	EWE [36]	0.0085	0.9542
	DBA [29]	0.0162	0.8778
	RAAW [29]	0.0090	0.9512
	Selective [76]	0.0103	0.9562
Feature-dependent	Triplet Embedding [41]	0.0101	0.9594
	EM [38]	0.0024	0.975
	Triplet Comparison [42]	0.0029	0.971
Feature-independent	Proposed method (Gaussian distribution Selective initialization, $W_d = 0$, $W_a = 1$)	0.0078	0.9648

6 Discussion

This paper investigates the process of obtaining moment-to-moment ratings of perception of stress from multiple annotators based on audio recordings, and proposes a feature-independent EM-based annotation fusion technique to aggregate these ratings to formulate gold standard ratings of perceived stress. For this purpose, we use the publicly available VerBIO dataset [52] which contains 22 hours of audio recordings of public speaking sessions from 53 participants. Recalling our contributions to the existing state of work (Section 2.3), we inspect how continuous time ratings of perception of stress can be collected from annotators who had access to the audio recordings of the public speaking sessions only, in the absence of video recording which is the most common modality being used in the annotation process [7], [27], [76]. Although there have been studies that used audio recordings to collect moment-to-moment ratings of affect dimensions, they mainly focused on the overall emotion in terms of valence and arousal [28], [39] instead of specific affect dimensions such as stress. We aim to bridge this gap in the current literature by collecting moment-to-moment ratings of perception of stress from four annotators who were asked to listen to the audio recording of 339 public speaking sessions.

We analyze multiple aspects of the inter-annotator agreement of this process which is measured by Spearman's ρ (Section 5.1.1). We find that 30% sessions have an average $\rho >= 0.3$ which is an indicator of moderate to strong inter-annotator agreement [82], [83]. However, this average ρ is

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

computed by aggregating ρ values from all possible pairs of annotators, and it might be affected by lower agreement between some pairs. To further investigate this issue, for each session, we construct a selective set of annotators who exhibited ρ value over a threshold with at least one other annotator. This approach of annotation quality analysis is similar to Metallinou *et al.* [57], [76]. Their empirical choice of threshold was 0.45 which resulted in ∼80% recordings having at least one annotator pair with higher than threshold agreement. Meanwhile, we have varied the threshold ρ to examine the distribution of sessions (Fig. 4), and findings from our work indicate that selecting $\rho = 0.4$ as the threshold value resulted in ∼60% sessions with at least one annotator pair with higher than threshold agreement. This suggests that our selected threshold agreement metric to divide sessions into high agreement and low agreement groups is similar to prior work [28], [76]. However, there is a slight difference in the percentage of sessions in the high agreement group between our work and Metallinou *et al.*'s research [57], [76] (i.e., 60% vs. 80%). This can be attributed to the difference in available modality between the two studies. In our work, the annotators had access to the audio recordings only in order to rate the perceived stress, while in [76], annotators rated the overall emotion while watching the audio-visual recordings of the sessions. This suggests that agreement between annotators might be affected when only audio modality is present or a specific affect dimension such as stress is being rated. Despite the lower percentage of sessions that is considered for the high agreement group, the average ρ for this group is 0.505, which is comparable to the agreement metric obtained in [28], [34], [76]. This indicates that the annotator agreement in our work based on only audio modality is on par with prior work where annotators relied on the audio-visual recording.

We examine the annotation characteristics (Section 5.1.2) and the acoustic properties of the audio (Section 5.1.3) between the high agreement and low agreement sessions. Annotators tend to exhibit increased agreement when they perceived a higher variation of stress during a session (i.e., $SD = 0.394 \pm 0.09$ in high agreement sessions) compared to the sessions with lower variation (i.e., $SD = 0.299 \pm 0.05$ in low agreement sessions). This iterates that annotators might find it difficult to perceive the subtle variation of stress. This observation is in line with prior work which points out that annotators are better at rating emotions in relative terms compared to absolute values [65], [85] and that higher variation of perceived emotions can help in more effectively recognizing one's affect state [67]. Similar differences are also found between the high agreement and low agreement sessions in terms of the acoustic features. High agreement sessions tend to contain speech with overall lower F1-F3 and jitter compared to the audios in low agreement sessions. However, an increase in jitter and an increase in F1-F3 are normally associated with an increase in psychological stress [21]. This discrepancy might have been caused by the fact that features were averaged over the entire duration of a public speaking session, therefore they might be also dependent on the overall vocal characteristics of a person and the annotators might not be able to perceive whether the speaker is under stress or whether the higher values of the features are due to the vocal characteristics. This sheds light on the complexity of the task of rating the perception of stress using the audio modality only.

Next, we propose a feature-independent, EM-based annotation fusion technique to obtain gold standard ratings of stress by introducing a modification in the methods presented in [38], [40]. We have incorporated the adjustment shift window along with the previously studied annotator-specific reaction delay. Based on the preliminary results (Section 5.2.1), the delay window length has been varied within a 4-6 seconds range, while 1-3 seconds is considered for adjustment window length. In addition, we experiment with the initialization (i.e., Mean vs. Selective) of the gold standard rating as well as the label distribution (i.e., Gaussian vs. Skew-Normal). Fused ratings are computed through the proposed technique by using different combinations of the aforementioned parameters. Unlike prior work in feature-independent annotation fusion [29], [36] that relied on empirical evaluation through visual inspection, we have conducted a formal evaluation through a prediction experiment. Our assumption is that the reliable gold-standard moment-to-moment ratings would maximally correlate with the input features (e.g., speech, physiology) that would maximize the prediction performance of the machine learning models [41], [50], [51]. We investigate the association between fused ratings and the corresponding modality used in the annotation process by measuring how good these ratings were in training a model that could estimate the fused ratings in the test set using the acoustic features. Among the baseline methods of annotation fusion (Section 5.2.2), Mean ratings exhibit the best predictive performance (i.e., $CCC = 0.2987 \pm 0.031$). This performance is surpassed by the proposed method with different combinations of design parameters (Section 5.2.3). The best-performing combination consisted of a 9 second window ($W_d = 6$, $W_a = 2$) with selective initialization and Gaussian distribution approximation that resulted in $CCC = 0.3320 \pm 0.029$, a significantly higher predictive performance than the baseline methods. This indicates that the proposed feature-independent annotation fusion technique can construct gold standard ratings that are better associated with the corresponding modality and are more generalizable compared to conventional methods [28], [29], [36].

Findings from the evaluation suggested that 6 second delay window (i.e., $W_d = 6$) and 2 second adjustment window (i.e., $W_a = 2$) constituted the best-performing configuration. Prior work indicates that annotators can have reaction delay between 2 and 6 seconds [32], [37], [38]. A delay window of a 6 second length would be able to account for this range of annotator-specific delay found in the literature. Although adjustment shift has not been previously incorporated in annotation fusion, several studies tried to align ratings from multiple annotators by shifting ratings in $[-2, 2]$ second range [28], [32], where the positive end of the range can be considered as a proxy of the adjustment shift proposed in our work. The choice of adjustment shift window length in our case is in line with the temporal range for the alignment of ratings.

Furthermore, we explore the degree of association between the fused time-continuous ratings and other measures of stress, such as self-reported anxiety scores and physiological indices (Section 5.2.4). Findings indicate that

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

the time-continuous ratings obtained by the proposed method exhibit higher correlation (i.e., $r = 0.1905$) than the majority of the baseline methods. The overall low range of the correlation might be attributed to the aggregation of ratings into mean values for comparison. However, such low correlation between self-reported and external observer affect ratings is common in prior work [63], [79]. In addition, our results indicate higher association between the fused ratings and SCL features from the EDA signal which is a commonly used biomarker of stress [8], [21], compared to baseline annotation fusion methods. Therefore, these findings suggest that the proposed annotation fusion method is able to construct reliable gold standard time-continuous ratings of stress that not only maximizes the predictive performance of machine learning models, but also exhibits better association with different measures of stress captured through self-reports and physiological signals compared to the baseline methods.

Next, we investigate whether the fused gold standard ratings obtained from the proposed annotation fusion method can be better estimated from acoustic features by an LSTM model, compared to the individual ratings estimated from the multi-task learning framework [59] that does not involve annotation fusion. Results suggest that the model is able to estimate the gold standard rating better than the individual ratings. This highlights the efficacy of the proposed annotation fusion method in generating a better gold standard rating compared to the individual ratings. Finally, we examine how efficient the proposed method is in approximating ground truth. For this analysis, we have used the green intensity task data, a synthetic dataset with known ground truth [41], [53], as this analysis would not be possible with the VerBIO dataset due to the lack of objective ground truth of stress ratings. The goodness of approximation of the known ground truth by the fused rating is evaluated using MSE and Pearson's r. Overall, the proposed method is able to approximate the ground truth better than the feature-independent baseline methods, as well as one of the feature-dependent methods, triplet embedding method [41]. This indicates that the proposed method is able to construct gold standard ratings that can better approximate ground truth, if one exists.

Despite the promising results, the work presented in this paper poses several limitations. Although the annotators are provided with detailed instructions before starting the annotation tasks, there is no frame-of-reference training [86] and annotators are not given examples of the different labels [35], [68]. This might have introduced more bias caused by subjectivity from the inter-individual differences of the annotators. In addition, the annotators are not chosen from medical professionals, to reduce the logistic load of the annotation process. The quality of annotation of stress might have been enhanced by incorporating medical professionals as they tend to have higher expertise compared to our chosen annotator pool. Next, the fused ratings constructed using Skew-Normal approximation perform worse than the mean rating and the ratings from the Gaussian approximation. A possible reason behind the poor performance of the Skew-Normal approximation might be the EM optimization scheme that is being used. Unlike the Gaussian approximation case, the Skew-Normal approximation does not have a closed-form solution [49] for the parameters that are being optimized. Although gradient descent optimizers have been employed in estimating the parameters of the Skew-Normal approximation, the choice of the optimization parameters has been done empirically. A more methodical and sophisticated approach might resolve this problem which we plan to address as part of our future work.

7 Conclusion

In this paper, we investigated the feasibility of collecting moment-to-moment ratings of perceived stress from multiple annotators using only audio signals as the behavioral marker and proposed an EM-based feature-independent annotation fusion technique to construct the gold standard rating. The proposed technique accounts for the annotator-specific reaction delay as well as the adjustment shift. Fused ratings obtained through different configurations of the design parameters are compared to the conventional methods (i.e., Mean, EWE). Instead of empirical evaluation, this comparison is performed by conducting a formal evaluation through a prediction experiment to estimate the ratings of stress using acoustic features. Results from the experiments conducted on the VerBIO dataset [52] indicate that moderate inter-annotator agreement can be achieved when multiple annotators are asked to provide continuous time ratings of stress based on audio recordings only. In addition, the proposed fusion technique using filter windows for both reaction delay and adjustment shift is capable of constructing better gold standard ratings in terms of association with mutimodal features, self-reports, and generalization, as they were estimated significantly better by the acoustic features compared to the ratings from baseline methods. These ratings are going to be augmented with the current version of the VerBIO dataset, which has the potential to be a useful resource for researchers in the domain of data-driven stress detection. This work lays the foundation for the development of a computational model to detect stress in a time-continuous manner using multimodal signals. Such models are indispensable in designing stress intervention modules to alleviate the negative effect of stress among the general population [23], [24].

Acknowledgement

The authors would like to acknowledge the Engineering Information Foundation (EiF18.02) and National Science Foundation (NSF #1956021) for funding this work.

References

[1] S. Cohen, R. C. Kessler, and L. U. Gordon, *Measuring stress: A guide for health and social scientists*. Oxford University Press on Demand, 1997.

[2] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press, 1991.

[3] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 291–298.

[4] H. Sarker, M. Tyburski, M. M. Rahman, K. Hovsepian, M. Sharmin, D. H. Epstein, K. L. Preston, C. D. Furr-Holden, A. Milam, I. Nahum-Shani *et al.*, "Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 4489–4501.

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

[5] A. Raij, P. Blitz, A. A. Ali, S. Fisk, B. French, S. Mitra, M. Nakajima, M. H. Nguyen, K. Plarre, M. Rahman *et al.*, "mstress: Supporting continuous collection of objective and subjective measures of psychosocial stress on mobile devices," *ACM Wireless Health 2010 San Diego, California USA*, 2010.

[6] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[7] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jaïdane, "Affectiveroad system and database to assess driver's attention," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018, pp. 800–803.

[8] M. Yadav, M. N. Sakib, K. Feng, T. Chaspari, and A. Behzadan, "Virtual reality interfaces and population-specific models to mitigate public speaking anxiety," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[9] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated analysis and prediction of job interview performance," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 191–204, 2016.

[10] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[11] D. O. Hebb, "Drives and the cns (conceptual nervous system)." *Psychological review*, vol. 62, no. 4, p. 243, 1955.

[12] P. Grossman, "Respiration, stress, and cardiovascular function," *Psychophysiology*, vol. 20, no. 3, pp. 284–300, 1983.

[13] G. S. Everly, Jr, J. M. Lating, G. S. Everly, and J. M. Lating, "The anatomy and physiology of the human stress response," *A clinical guide to the treatment of the human stress response*, pp. 19–56, 2019.

[14] M. Kalia, "Assessing the economic impact of stress [mdash] the modern day hidden epidemic," *Metabolism-clinical and experimental*, vol. 51, no. 6, pp. 49–53, 2002.

[15] D. Wainwright and M. Calnan, *Work stress: The making of a modern epidemic*. McGraw-Hill Education (UK), 2002.

[16] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," *Communication education*, vol. 59, no. 1, pp. 70–105, 2010.

[18] W. B. Cannon, "The wisdom of the body," 1939.

[19] S. Baltaci and D. Gokcay, "Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features," *International Journal of Human–Computer Interaction*, vol. 32, no. 12, pp. 956–966, 2016.

[20] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 201–216, 2001.

[21] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, 2019.

[22] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.

[23] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 351–360.

[24] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: towards a gold standard for continuous stress assessment in the mobile environment," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 493–504.

[25] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "MuSE: a multimodal dataset of stressed emotion," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 1499–1510.

[26] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[27] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 5–14.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "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, 2013, pp. 1–8.

[29] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigell, E. Cambria, and B. W. Schuller, "Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 75–82.

[30] J. R. Crawford and J. D. Henry, "The depression anxiety stress scales (dass): Normative data and latent structure in a large non-clinical sample," *British journal of clinical psychology*, vol. 42, no. 2, pp. 111–131, 2003.

[31] C. Barbosa-Leiker, M. Kostick, M. Lei, S. McPherson, V. Roper, T. Hoekstra, and B. Wright, "Measurement invariance of the perceived stress scale and latent mean differences across gender and time," *Stress and Health*, vol. 29, no. 3, pp. 253–260, 2013.

[32] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[33] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds." *Journal of machine learning research*, vol. 11, no. 4, 2010.

[34] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, S. Narayanan *et al.*, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[35] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech 2020*, 2020.

[36] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 381–385.

[37] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 85–90.

[38] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 76–89, 2018.

[39] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.

[40] A. Ramakrishna, R. Gupta, and S. Narayanan, "Joint multi-dimensional model for global and time-series annotations," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 473–484, 2022.

[41] B. M. Booth, K. Mundnich, and S. Narayanan, "Fusing annotations with majority vote triplet embeddings," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 83–89.

[42] K. Mundnich, B. M. Booth, B. Girault, and S. Narayanan, "Generating labels for regression of subjective constructs using triplet embeddings," *Pattern Recognition Letters*, vol. 128, pp. 385–392, 2019.

[43] Noldus, "The observer xt." [Online]. Available: https://www.noldus.com/observer-xt

[44] L. G. Jaimes, M. Llofriu, and A. Raij, "Preventer, a selection mechanism for just-in-time preventive interventions," *IEEE Transactions on Affective Computing*, vol. 7, no. 3, pp. 243–257, 2015.

[45] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "An audio-visual dataset of human–human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.

[46] M. Park, H. Oh, and K. Lee, "Security risk measurement for information leakage in iot-based smart homes from a situational awareness perspective," *Sensors*, vol. 19, no. 9, p. 2148, 2019.

[47] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[48] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian journal of statistics*, pp. 171–178, 1985.

[49] A. Azzalini and A. D. Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, no. 4, pp. 715–726, 1996.

[50] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Summary for avec 2018: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 2111–2112.

[51] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3435502

cca for analysis of affective behaviour," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*. Springer, 2012, pp. 98–111.
[52] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. H. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1168–1182, 2022.
[53] B. M. Booth, K. Mundnich, and S. S. Narayanan, "A novel method for human bias correction of continuous-time annotations," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3091–3095.
[54] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *2013 humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 709–710.
[55] J. M. Girard, "CARMA: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, p. e5, 2014.
[56] B. M. Booth and S. S. Narayanan, "Fifty shades of green: Towards a robust measure of inter-annotator agreement for continuous signals," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 204–212.
[57] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *2013 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2013, pp. 1–6.
[58] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, USA:, 1994, pp. 359–370.
[59] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Deep learning for continuous multiple time series annotations," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 91–98.
[60] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.
[61] C. D. Spielberger, "State-trait anxiety inventory for adults," 1983.
[62] S. Cohen, "Perceived stress in a probability sample of the united states." 1988.
[63] M. Pörhölä, "Trait anxiety, experience, and the public speaking state responses of finnish university students," *Communication research reports*, vol. 14, no. 3, pp. 367–384, 1997.
[64] "E4 wristband," https://www.empatica.com/en-gb/research/e4/, (Accessed on 04/04/2022).
[65] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2021.
[66] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 158–163.
[67] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, 2016.
[68] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.
[69] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[70] S. S. Yalowitz and K. Bronnenkant, "Timing and tracking: Unlocking visitor behavior," *Visitor Studies*, vol. 12, no. 1, pp. 47–64, 2009.
[71] A. Blackler, V. Popovic, and D. Mahar, "Studies of intuitive interaction employing observation and concurrent protocol," in *Proceedings of the Design 2004 8th International Design Conference*. Faculty of Mechanical Engineering and Naval Architecture, Zagreb, The Design . . . , 2004, pp. 135–143.
[72] N. C. Silver and W. P. Dunlap, "Averaging correlation coefficients: Should fisher's z transformation be used?" *Journal of Applied Psychology*, vol. 72, no. 1, p. 146, 1987.
[73] A. Azzalini and A. Capitanio, "Statistical applications of the multivariate skew normal distribution," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 579–602, 1999.
[74] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[75] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," *arXiv preprint arXiv:1806.10658*, 2018.
[76] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.
[77] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
[78] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
[79] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.
[80] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y
[81] E. H. Nirjhar, A. Behzadan, and T. Chaspari, "Exploring bio-behavioral signal trajectories of state anxiety during public speaking," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1294–1298.
[82] H. Akoglu, "User's guide to correlation coefficients," *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.
[83] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson education, 2007.
[84] G. Rupert Jr *et al.*, "Simultaneous statistical inference," 2012.
[85] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 574–580.
[86] G. R. VandenBos, *APA dictionary of psychology*. American Psychological Association, 2007.

Ehsanul Haque Nirjhar received his B.Sc. in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2015. He is pursuing his Ph.D. in Computer Science at Texas A&M University, under the supervision of Dr. Theodora Chaspari. His research interests are affective computing, signal processing, and machine learning. He is a student member of the IEEE and the ACM.

Theodora Chaspari (S'12, M'17) received her Ph.D (2017) and M.S. (2012) in Electrical Engineering from the University of Southern California, and diploma in Electrical & Computer Engineering from the National Technical University of Athens, Greece (2010). She is currently an Associate Professor in Computer Science and the Institute of Cognitive Science (ICS) at University of Colorado, Boulder and her research interests lie in the area of affective computing, machine learning, and signal processing. She is a recipient of the USC Annenberg Graduate Fellowship 2010, USC Women in Science and Engineering Merit Fellowship 2015, and the TAMU CSE Graduate Faculty Teaching Excellence Award 2019. Papers co-authored with her students have been nominated and won awards at the ACM BuildSys 2019, IEEE ACII 2019, ASCE i3CE 2019, and IEEE BSN 2018 conferences.

Authorized licensed use limited to: UNIVERSITY OF COLORADO. Downloaded on August 14,2024 at 14:40:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.