Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1499–1510
Marseille, 11–16 May 2020

MuSE: a Multimodal Dataset of Stressed Emotion

Mimansa Jaiswal[1], Cristian-Paul Bara[1], Yuanhang Luo[1], Mihai Burzo[2]
Rada Mihalcea[1], Emily Mower Provost[1]
[1]University of Michigan - Ann Arbor
[2]University of Michigan - Flint
{mimansa, cpbara, royluo, mburzo, mihalcea, emilykmp}@umich.edu

Abstract

Endowing automated agents with the ability to provide support, entertainment and interaction with human beings requires sensing of the users' affective state. These affective states are impacted by a combination of emotion inducers, current psychological state, and various contextual factors. Although emotion classification in both singular and dyadic settings is an established area, the effects of these additional factors on the production and perception of emotion is understudied. This paper presents a dataset, Multimodal Stressed Emotion (MuSE), to study the multimodal interplay between the presence of stress and expressions of affect. We describe the data collection protocol, the possible areas of use, and the annotations for the emotional content of the recordings. The paper also presents several baselines to measure the performance of multimodal features for emotion and stress classification.
Keywords: multimodal emotion, stressed emotion, natural language, spontaneus speech

1. Introduction

Virtual agents have become more integrated into our daily lives than ever before (Lucas et al., 2014). For example, Woebot is a chatbot developed to provide cognitive behavioral therapy to a user (Fitzpatrick et al., 2017). For this chatbot agent to be effective, it needs to respond differently when the user is stressed and upset versus when the user is calm and upset, which is a common strategy in counselor training (Thompson et al., 2013). While virtual agents have made successful strides in understanding the task-based intent of the user, social human-computer interaction can still benefit from further research (Clark et al., 2019). Successful integration of virtual agents into real-life social interaction requires machines to be emotionally intelligent (Bertero et al., 2016; Yuan, 2015).

But humans are complex in nature, and emotion is not expressed in isolation (Griffiths, 2003). Instead, it is affected by various external factors. These external factors lead to interleaved user states, which are a culmination of situational behavior, experienced emotions, psychological or physiological state, and personality traits. One of the external factors that affects psychological state is stress. Stress can affect everyday behavior and emotion, and in severe states, is associated with delusions, depression and anxiety due to impact on emotion regulation mechanisms (Kingston and Schuurmans-Stekhoven, 2016; Schlotz et al., 2011; Tull et al., 2007; Wang and Saudino, 2011). Virtual agents can respond in accordance to users' emotions only if the machine learning systems can recognize these complex user states and correctly perceive users' emotional intent. We introduce a dataset designed to elicit spontaneous emotional responses in the presence or absence of stress to observe and sample complex user states.

There has been a rich history of visual (You et al., 2016; Jiang et al., 2014), speech (Lotfian and Busso, 2017), linguistic (**?**), and multimodal emotion datasets (Busso et al., 2017; Busso et al., 2008; Ringeval et al., 2013). Vision datasets have focused both on facial movements (Jiang et al., 2014) and body movement (Lazarus and Cohen, 1977). Speech datasets have been recorded to capture both stress and emotion separately but do not account for their interdependence (Rothkrantz et al., 2004; Horvath, 1982; Kurniawan et al., 2013; Zuo and Fung, 2011). Stress datasets often include physiological data (Yaribeygi et al., 2017; Sun et al., 2012).

Existing datasets are limited because they are designed to elicit emotional behavior, while neither monitoring external psychological state factors nor minimizing their impact by relying on randomization. However, emotions produced by humans in the real world are complex. Further, our natural expressions are often influenced by multiple factors (e.g., happiness *and* stress) and do not occur in isolation, as typically assumed under laboratory conditions. The primary goal of this work is to collect a multimodal stress+emotion dataset – Multimodal Stressed Emotion (MuSE) – to promote the design of algorithms that can recognize complex user states.

The MuSE dataset consists of recordings of 28 University of Michigan college students, 9 female and 19 male, in two sessions: one in which they were exposed to an external stressor (final exams period at University of Michigan) and one during which the stressor was removed (after finals have concluded). Each recording is roughly 45-minutes. We expose each subject to a series of emotional stimuli, short-videos and emotionally evocative monologue questions. These stimuli are different across each session to avoid the effect of repetition, but capture the same emotion dimensions. At the start of each session, we record a short segment of the user in their natural stance without any stimuli, to establish a baseline. We record their behavior using four main recording modalities: 1) video camera, both close-up on the face and wide-angle to capture the upper body, 2) thermal camera, close-up on the face, 3) lapel microphone, 4) physiological measurements, in which we choose to measure heart rate, breathing rate, skin conductance and skin temperature (Figure 1). The data include self-report annotations for emotion and stress (Perceived Stress Scale, PSS) (Cohen, 1988; Cohen et al., 1994), as well as emotion annotations obtained from Amazon Mechanical Turk (AMT). To understand the influence of

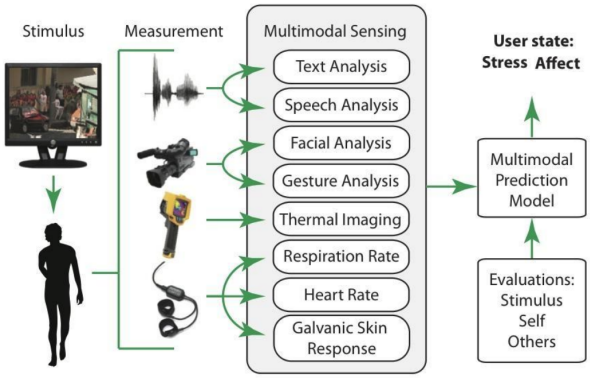

Figure 1: Broad visual overview of recordings

personality on the interaction of stress and emotion, we obtain Big-5 personality scores (Goldberg, 1992), which was filled by 18 of the participants, due to the participation being voluntary. The extracted features for each modality, and the anonymized dataset (other than video) will be released publicly along with all the corresponding data and labels. We present baseline results for recognizing both emotion and stress in the paper, in order to validate that the presence of these variables can be computationally extracted from the dataset, hence enabling further research.

2. Related Work

In the past years, there have been multiple emotional databases collected and curated to develop better emotion recognition systems. Table 1 shows the major corpora that are used for emotion recognition. However, some aspects of the datasets limit their applicability, including: a lack of naturalness, unbalanced emotion content, unmeasured confounding variables, small size, small number of speakers, and presence of background noise. These datasets are also limited in the number of modalities they use, usually relying on visual and acoustic/lexical information.

2.1. Recorded Modalities

As shown in Table 1, the most common modalities are video, acoustics, and text. In addition to these modalities, we chose to record two more modalities: thermal and physiological. Previous research has shown that thermal recordings perform well as non-invasive measurement of physiological markers like, cardiac pulse and skin temperature (Pavlidis et al., 2000; Pavlidis and Levine, 2002; Garbey et al., 2007). They have been shown to be correlated to stress symptoms, among other physiological measures. We used the physiological modality to measure stress responses (Yaribeygi et al., 2017; Sun et al., 2012) to psychological stressors. This modality has been previously noted in literature for measuring stress (Horvath, 1978), usually measured in polygraph tests. We perform baseline experiments to show that the modalities collected in the dataset are indeed informative for identifying stress and emotion.

2.2. Lack of Naturalness

A common data collection paradigm for emotion is to ask actors to portray particular emotions. These are usually either short snippets of information (Busso et al., 2008), a single sentence in a situation (Busso et al., 2017), or obtained from sitcoms and rehearsed broadcasts (Chen et al., 2018). A common problem with this approach is that the resulting emotion display is not natural (Jürgens et al., 2015). These are more exaggerated versions of singular emotion expression rather than the general, and messier, emotion expressions that are common in the real world (Audibert et al., 2010; Batliner et al., 1995; Fernández-Dols and Crivelli, 2013). Further, expressions in the real world are influenced by both conversation setting and psychological setting. While some datasets have also collected spontaneous data (Busso et al., 2008; Busso et al., 2017), these utterances, though emotionally situated, are often neutral in content when annotated. The usual way to get natural emotional data is to either collect data using specific triggers that have been known to elicit a certain kind of response or to completely rely on in-the wild data, which however often leads to unbalanced emotional content in the dataset (Ringeval et al., 2013).

2.3. Unbalanced Emotion Content

In-the-wild datasets are becoming more popular (Chen et al., 2018; Khorram et al., 2018; Li et al., 2016). The usual limitation to this methodology is that, firstly, for most people, many conversations are neutral in emotion expression. This leads to a considerable class imbalance (Ringeval et al., 2013). To counter this issue, MSP-Podcast (Lotfian and Busso, 2017) deals with unbalanced content by preselecting segments that are more likely to have emotional content. Secondly, data collected in particular settings, e.g., therapy (Nasir et al., 2017), or patients with clinical issues (Lassalle et al., 2019) comprise mostly of negative emotions because of the recruitment method used in the collection protocol.

2.4. Presence of Interactional Variables

The common way of inducing emotions involves either improvisation prompts or scripted scenarios. Emotion has been shown to vary with a lot of factors that are different from the intended induction (Siedlecka and Denson, 2019; Zhang et al., 2014; Mills and D'Mello, 2014). These factors in general can be classified into: (a) recording environment confounders and (b) collection confounders. Recording environment-based variables hamper the models' ability to to learn the emotion accurately. These can be environment noise (Banda and Robinson, 2011), placement of sensors or just ambient temperature (Bruno et al., 2017).

The data collection variations influence both the data generation and data annotation stages. The most common confounders are gender, i.e., ensuring an adequate mix of male vs female, and culture, i.e., having a representative sample to train a more general classifier. Another confounding factor includes personality traits (Zhao et al., 2018), which influence how a person both produces (Zhao et al., 2018) and perceives (Mitchell, 2006) emotion. Another confounder that can occur at the collection stage is the familiarity between the participants, like RECOLA (Ringeval et al., 2013), which led to most of the samples being mainly positive due to the colloquial interaction between the participants. They also do not account for the psychologi-

Table 1: Summary of some of the existing emotion corpora. Lexical modality is mentioned for manually transcribed datasets. A - Audio, L - Lexical, T- Thermal, V- Visual, P - Physiological.

	Corpus	Size	Speakers	Rec. Type	Language	Modality	Annotation Type
1.	IEMOCAP	12h26m	10	improv/acted	English	A, V, L	Ordinal, Categorical
2.	MSP-Improv	9h35m	12	improv/acted	English	A, V	Ordinal
3.	VAM	12h	47	spontaneous	German	A, V	Ordinal
4.	SEMAINE	6h21m	20	spontaneous	English	A, V	Ordinal, Categorical
5.	RECOLA	2h50m	46	spontaneous	French	A, V, P	Ordinal
6.	FAU-AIBO	9h12m	51	spontaneous	German	A, L	Categorical
7.	TUM AVIC	0h23m	21	spontaneous	English	A, V, L	Categorical
8.	Emotion Lines	30k samples	-	spont/scripted	English	A, L	Categorical
9.	OMG-Emotion	2.4k samples	-	spontaenous	English	A, V, L	Ordinal
10.	MSP-Podcast	27h42m	151	spontaenous	English	A	Ordinal, Categorical
11.	MuSE	10h	28	spontaneous	English	A, V, L, T, P	Ordinal (Random, Context)

cal state of the participant. Psychological factors such as stress (Lech and He, 2014), anxiety (Werner et al., 2011) and fatigue (Berger et al., 2012) have been shown previously to have significant impact on the display of emotion. But the relation between these psychological factors and the performance of models trained to classify emotions in these situations has not been studied.

The second set of confounders occurring from collection protocols are due to the way annotations are collected. Previous research has shown the difference between obtaining continuous vs single label per utterance (Jaiswal et al., 2019a). (Yannakakis et al., 2017) have also looked at the differences between ordinal vs categorical measures of emotion, showing that humans use anchors to evaluate the emotional state of a stimulus; suggesting again that ordinal labels are a more suitable way to represent emotions. Many of these collected emotion datasets rely on either expert evaluation or crowdsourced annotations (Busso et al., 2017). Previous work has looked at the trade-off between quality and quantity of the annotations received from crowdsourcing workers. Research in human- computer interaction and economics has also looked at the quality of annotations received as a function of hourly pay (Horton and Chilton, 2010), and of annotators psychological state (Paulmann et al., 2016) and found how increased pay leads to exponentially better annotations. Out of the datasets mentioned in Table 1, some of them present all information to the annotator to label the utterances while others just provide a single sentence. The authors have studied the effect of these labeling schemes on the annotations received and the performance of the machine learning algorithms trained on these annotations. The effect of these labelling schemes on the annotations received has been compared in terms of the annotations themselves and the change in classifiers' performance on the dataset (Jaiswal et al., 2019a). We collect a dataset that is indicative of how stress, a psychological confounding factor, interleaves with emotion production. We present baselines to verify that these outputs can be computationally extracted from the dataset.

3. MuSE Dataset

3.1. Experimental Protocol

We collect a dataset that we refer to as Multimodal Stressed Emotion (MuSE) to facilitate the learning of the interplay between stress and emotion. The protocol for data collection is shown in Figure 2. There were two sections in each recording: monologues and watching emotionally evocative videos. We measure the stress level at the beginning and end of each recording. The monologue questions and videos were specifically chosen to cover all categories of emotions. At the start of each recording, we also recorded a short one-minute clip without any additional stimuli to register the baseline state of the subject.

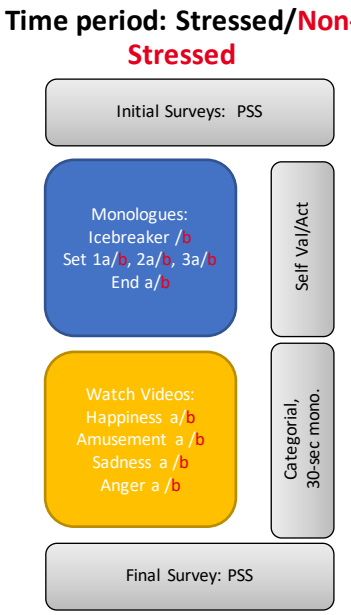

Figure 2: Experimental Protocol For Recording

Previous research has elicited situational stress such as public speaking (Kirschbaum et al., 1993; Giraud et al., 2013; Aguiar et al., 2014), mental arithmetic tasks (Liao and Carey, 2015) or use Stroop Word Test (Tulen et al., 1989). However, these types of stress are often momentary and fade rapidly in two minutes (Liao and Carey, 2015). We alleviate this concern by recording both during and after final exams (we anticipate that these periods of time are associated with high stress and low stress, respectively) in April 2018. We measure stress using Perceived Stress Scale (Cohen et al., 1994) for each participant. We measure their self-perception of the emotion using Self-Assessment Manikins (SAM) (Bradley and Lang, 1994). The recordings

and the survey measures were coordinated using Qualtrics[1] enabling us to ensure minimal intervention and limit the effect of the presence of another person on the emotion production.

Each monologue section comprised of five questions broken into sections meant to elicit a particular emotion (Table 2). These questions were shown to elicit thoughtful and emotional responses in their data pool to generate interpersonal closeness (Aron et al., 1997). We include an icebreaker and ending question to ensure cool off periods between change in recording section, i.e., from neutral to monologues, and from monologues to videos, hence decreasing the amount of carry-over emotion from the previous monologue to the next. Each subject was presented with a different set of questions over the two recordings to avoid repetition effect. We also shuffle the order of the other three questions to account for order effects (Lee et al., 2011). Each subject was asked to speak for a minimum of two minutes. After their response to each question, the subjects marked themselves on two emotion dimensions: activation and valence on a Likert Scale of one to nine using self-assessment manikins (Bradley and Lang, 1994).

For the second part of the recording, the subjects were asked to watch videos in each of the four quadrants i.e., the combination of *{low, high}* $\times$ *{activation, valence}* of emotion. These clips were selected from the corpus (Lichtenauer and Soleymani, 2011; Bartolini, 2011), which tested for the emotion elicited from the people when watching these clips (Table 3). The subjects were monitored for their reaction to the clips. After viewing a clip, subjects are asked to speak for thirty seconds about how the video made them feel. After their response, they marked a emotion category, e.g., angry, sad, etc. for the same clip. When switching videos, the subjects were asked to view a one-minute neutral clip to set their physiological and thermal measures back to the baseline (Samson et al., 2016).

The 28 participants were also asked to fill out an online survey used for personality measures on the big-five scale (Goldberg, 1992), participation being voluntary. This scale has been validated to measure five different dimensions named OCEAN (openness, conscientiousness, extraversion, agreeableness, and neuroticism) using fifty questions and has been found to correlate with passion (Dalpé et al., 2019), ambition (Barrick and Mount, 1991), and emotion mechanisms (Querengässer and Schindler, 2014). We received responses for this survey from 18 participants. These labels can be used in further work to evaluate how these personality measures interact with the affects of stress in emotion production, as previously studied in (Zhao et al., 2018).

3.2. Equipment Setup

The modalities considered in our setup are: thermal recordings of the subject's face, audio recordings of the subject, color video recording of the subject's face, a wide-angle color video recording the subject from the waist up and physiological sensors measuring skin conductance, breathing rate, heart rate and skin temperature. For these modalities we have set up the following equipment:

[1] umich.qualtrics.com

Table 2: Emotion elicitation questions.

Icebreaker
1. Given the choice of anyone in the world, whom would you want as a dinner guest?
2. Would you like to be famous? In what way?
Positive
1. For what in your life do you feel most grateful?
2. What is the greatest accomplishment of your life?
Negative
1. If you could change anything about the way you were raised, what would it be?
2. Share an embarrassing moment in your life.
Intensity
1. If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone?
2. Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?
Ending
1. If you were able to live to the age of 90 and retain either the mind or body of a 30-year old for the last 60 years of your life, which would you choose?
2. If you could wake up tomorrow having gained one quality or ability, what would it be?

Table 3: Emotion elicitation clips.

Movie	Description
Low Valence, Low Activation (Sad)	
City of Angels	Maggie dies in Seth's arms
Dangerous Minds	Students find that one of their classmates has died
Low Valence, High Activation (Anger)	
Sleepers	Sexual abuse of children
Schindler's List:	Killing of Jews during WWII
High Valence, Low Activation (Contentment)	
Wall-E	Two robots dance and fall in love
Love Actually	Surprise orchestra at the wedding
High Valence, High Activation (Amusement)	
Benny and Joone	Actor plays the fool in a coffee shop
Something About Mary	Ben Stiller fights with a dog
Neutral	
A display of zig-zag lines across the screen	
Screen-saver pattern of changing colors	

1. **FLIR Thermovision A40 thermal camera** for recording the close-up thermal recording of the subject's face. This camera provides a 640x512 image in the thermal infrared spectrum.
2. **Raspberry Pi with camera module V2 with wide-angle lens** used for the waist up shot of the subject. We have chosen Raspberry Pi's due to its low price

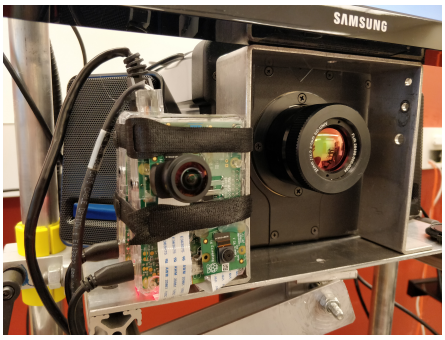

Figure 3: Close-up view of the thermal and video recording equipment.

and support for Linux OS, which integrates easily into a generic setup.

3. **Raspberry Pi with camera module V2** used to record the subject from the waist up.
4. **TASCAM DR-100 mk II** used to record audio. We chose this product for its high fidelity. It can record 24-bit audio at 48kHz.
5. **ProComp$^{\infty}$-8 channel biofeedback and neurofeedback system v6.0** used to measure blood volume pulse (BVP sensor), skin conductance (SC sensor), skin temperature (T sensor), and abdominal respiration (BR sensor)

The equipment operator started and marked the synchronization point between video and audio recordings using a clapper. Subsequent time stamps are recorded by the qualtrics survey using subject click timings.

3.3. Post-processing

Splitting of the Recordings. Each modality is split into neutral recordings of one-minute, five questions and four video recordings with associated monologues, resulting in fourteen recordings for emotional content, thus 28 recordings per subject. In total we have 784 distinct recordings over five modalities, 28 subjects and two stress states, for a total of 3920 recording events. Temperatures are clamped to between 0^oC and 50^oC. This helps reduce the size of the thermal recording files after being zipped.

Utterance Construction. The five monologues extracted above were divided into utterances. However, since the monologues are a form of spontaneous speech, there are no clear sentence boundaries marking end of utterance. We manually created utterances by identifying prosodic or linguistic boundaries in spontaneous speech as defined by (Kolář, 2008). The boundaries used for this work are: (a) clear ending like a full stop or exclamation, (b) a change in context after filler words or completely revising the sentence to change meaning, or (c) a very long pause in thought. This method has been previously shown to be effective in creating utterances that mostly maintain a single level of emotion (Khorram et al., 2018).

The dataset contains 2,648 utterances with a mean duration of 12.44 ± 6.72 seconds (Table 4). The mean length of stressed utterances (11.73 ± 5.77 seconds) is significantly different (using two-sample t-test) from that of the non-stressed utterances (13.30 ± 6.73 seconds). We remove utterances that are shorter than 3-seconds and longer than 35-seconds and end up retaining 97.2% of our dataset. This allows us to to avoid short segments that may not have enough information to capture emotion, and longer segments that can have variable emotion, as mentioned in (Khorram et al., 2018). Because our dataset is comprised of spontaneous utterances, the mean length of utterance is larger than those in a scripted dataset (Busso et al., 2017) due to more corrections and speech overflow.

Stress State Verification. We perform a paired t-test for subject wise PSS scores, and find that the mean scores are significantly different for both sets (16.11 vs 18.53) at $p < 0.05$. This implied that our hypothesis of exams eliciting persistently more stress than normal is often true. In our dataset, we also provide levels of stress which are binned into three categories based on weighted average (using questions for which the t-test score was significant).

4. Emotional Annotation

4.1. Crowdsourcing

Crowdsourcing has previously been shown to be an effective and inexpensive method for obtaining multiple annotations per segment (Hsueh et al., 2009; Burmania and Busso, 2017). We posted our experiments as Human Intelligence Tasks (HITs) on *Amazon Mechanical Turk* and used selection and training mechanisms to ensure quality (Jaiswal et al., 2019a). HITs were defined as sets of utterances in a monologue. The workers were presented with a single utterance and were asked to annotate the activation and valence values of that utterance using Self-Assessment Manikins (Bradley and Lang, 1994). Unlike the strategy adopted in (Chen et al., 2018), the workers could not go back and revise the previous estimate of the emotion. We did this to ensure similarity to how a human listening into the conversation might shift their perception of emotion in real time. These HITs were presented in either the contextual or the random presentation condition defined below.

In the contextual experiment, we posted each HIT as a collection of ordered utterances from each section of a subject's recording. Because each section's question was designed to elicit an emotion, to randomize the carry-over effect in perception, we posted the HITs in a random order over the sections from all the subjects in our recording. For example, a worker might see the first HIT as *Utterance 1...N from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1...M from Section 5 of Subject 10's non-stressed recording* where *N, M* are the number of utterances in those sections respectively. This ensures that the annotator adapts to the topic and fluctuations in speaking patterns over the monologue being annotated.

In the randomized presentation, each HIT is an utterance from any section, by any speaker, in random order. So, a worker might see the first HIT as *Utterance 11 from Section 2 of Subject 1's stressed recording monologue* and see the second HIT as *Utterance 1 from Section 5 of Subject 10's non-stressed monologue recording*. We use this method of randomization to ensure lack of adaptation to both speaker

Table 4: Data summary (R:random, C:context, F:female, M:male).

Monologue Subset	
Mean no. of utterances/monologue	9.69 ± 2.55
Mean duration of utterances	12.44 ± 6.72 seconds
Total no. of utterances	2,648
Selected no. of utterances	2,574
Gender distribution	19 (M) and 9 (F)
Total annotated speech duration	$\sim$ 10 hours
Crowdsourced Data	
Num of workers	160 (R) and 72 (C)
Blocked workers	8
Mean activation	3.62±0.91 (R) 3.69±0.81 (C)
Mean valence	5.26±0.95 (R) 5.37±1.00 (C)

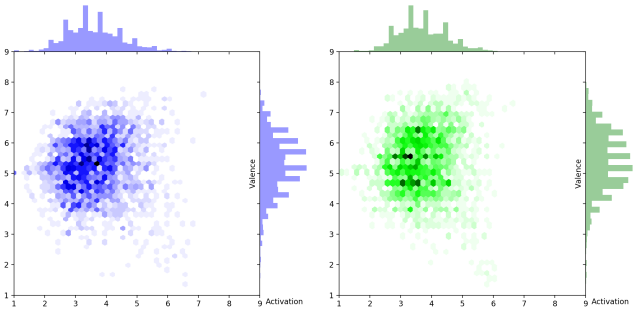

Figure 4: Distribution of the activation and valence ratings in random labeling scheme (on left) and contextual labeling scheme (on right).

specific style and the contextual information. Our previous work (Jaiswal et al., 2019a) showed that a mismatch between annotation and training conditions leads to poorer performance of the trained machine learning model. Hence, the per-utterance and the contextual labels can be used to train different machine learning models that are apt for either singular one-off instances or for holding multiple turn natural conversation, respectively.

4.2. Emotion Content Analysis

We show the distribution of the annotations received in both the random and contextual setting in Table 4 and Figure 4. The labels obtained for our dataset form a distribution that mostly covers negative and neutral levels of activation, and all but extremities for valence. This can also be seen in the data summary in Table 4. We performed a paired t-test between the labels obtained from random vs contextual presentation and found that these labels are significantly different (using paired t-test at $p < 0.05$ for both activation and valence for utterances in the non-stressed situation). Although the obtained labels are significantly different for valence in the stressed category using the same method as above, the same does not hold true for the activation annotations in this category.

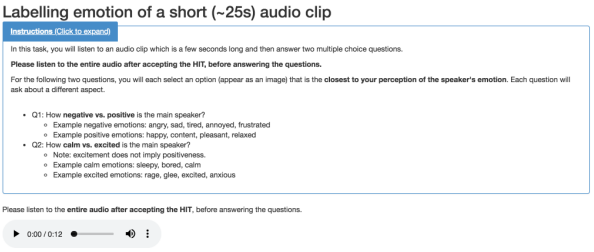

Figure 5: An overview of the instructions provided to the annotators for annotating an utterance.

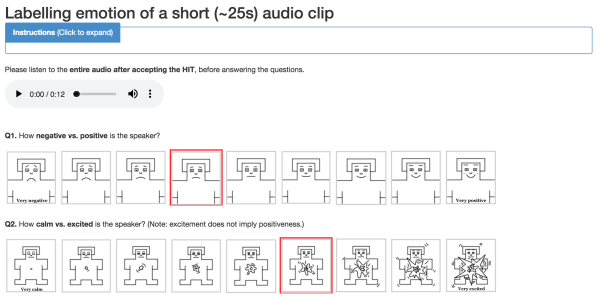

Figure 6: Annotation scale used by MTurk workers to annotate the emotional content of the corpus. They annotate valence and activation for each utterance.

5. Experiments

In this section, we describe our baseline experiments for predicting emotion and stress in the recorded modalities. We have a more granular marked annotation of emotion, i.e., over each utterance, as compared to stress over the complete monologue. Hence, we extract features for each modality over continuous one second frame intervals for predicting stress, and over the complete utterance for emotion. Audio and lexical features are still extracted over a complete utterance for stress due to higher interval of variation over time.

5.1. Evaluation of Emotion Recognition

We use the following set of features for our baseline models:

1. **Acoustic Features.** We extract acoustic features using OpenSmile (Eyben et al., 2010) with the eGeMAPS configuration (Eyben et al., 2016). The eGeMAPS feature set consists of 88 utterance-level statistics over the low-level descriptors of frequency, energy, spectral, and cepstral parameters. We perform speaker-level z-normalization on all features.
2. **Lexical Features.** We extract lexical features using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). These features have been shown to be indicative of stress, emotion, veracity and satisfaction (Golbeck et al., 2011; Monin et al., 2012; Newman et al., 2003). We normalize all the frequency counts by the total number of words in the sentence accounting for the variations due to utterance length.
3. **Thermal Features.** For each subject a set of four regions were selected in the thermal image: the forehead

area, the eyes, the nose and the upper lip as previously used in (Pavlidis and Levine, 2002; Garbey et al., 2007; Abouelenien et al., 2016). These regions were tracked for the whole recording and a 150-bin histogram of temperatures was extracted from the four regions per frame, i.e., 30 frames a second for thermal recordings. We further reduced the histograms to the first four measures of central tendency, e.g. Mean, Standard Deviation, Skewness and Kurtosis. We combined these features over the utterance using first delta measures (min, max, mean, SD) of all the sixteen extracted measures per frame, resulting in 48 measures in total.

4. **Close-up Video Features.** We use OpenFace (Baltrušaitis et al., 2016) to extract the subject's facial action units. The AUs used in OpenFace for this purpose are AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28 and AU25 comprising of eyebrows, eyes and mouth. These features have been previously shown to be indicative of emotion (Wegrzyn et al., 2017; Du et al., 2014) and have been shown to be useful for predicting deception (Jaiswal et al., 2016). We summarize all frames into a feature using summary statistics (maximum, minimum, mean, variance, quantiles) across the frames and across delta between the frames resulting in a total of 144 dimensions.

Network Setup. We train and evaluate multiple unimodal Deep Neural Networks (DNN) models for predicting valence and activation using Keras (Gulli and Pal, 2017). (Jaiswal et al., 2019a) have shown that a match between the context provided to the classifier and the annotator leads to better classification performance. Because we are performing single utterance classification, for all further experiments, we use the annotations obtained in a random manner as mentioned above. In all cases, we predict the continuous annotation using regression.

We also use an ensemble of these four networks (audio, lexical, visual and thermal) to measure multimodal performance. For each network setup, we follow a five-fold subject independent evaluation scheme and report the average RMSE across the folds. For each test-fold, we use the previous fold for hyper-parameter selection and early stopping. The hyper-parameters include: number of layers $\{2, 3, 4\}$ and layer width $\{64, 128, 256\}$. We use ReLU activation and train the networks with MSE loss using the Adam optimizer.

We train our networks for a maximum of 50 epochs and monitor the validation loss after each epoch. We perform early stopping if the loss doesn't decrease for 15 consecutive epochs. We save the weights that achieved the lowest validation performance during training. We train each network five times with different seeds and average the predictions to account for variations due to random initialization.

Results. We show our results in Table 5. We find that between acoustic and lexical modalities, the acoustic modality carries more information about activation and the lexical for valence. This is in line with previous research (Yang and Chen, 2011; Cambria et al., 2017). We also note that the visual modality significantly outperforms both the speech and lexical modalities for valence prediction.

Table 5: RMSE for emotion classification models using multiple modalities. Significance established at $p < 0.05$.

	Activation	**Valence**
Unimodal Models		
Acoustic (A)	**1.004***	1.122
Lexical (L)	1.343	0.980
Close Video (V)	1.111	**0.879****
Thermal (T)	2.012	1.565
Ensemble		
A+L	0.987	0.981
A+V	0.970	0.899
L+V	0.981	0.901
A+L+V	0.972	**0.856***
A+L+V+T (All)	**0.961***	0.868

When we merge these networks using late voting on each modality (decision fusion), we find that the combination of all modalities performs the best for predicting activation. But for predicting valence, the best performance is shown by the combination of acoustic, lexical, visual and thermal modalities. We believe this is true because previous work has shown that thermal features are mostly indicative of intensity and discomfort (Herborn et al., 2015) and hence improves performance on activation prediction, while the visual expressions are most informative about valence (Rubo and Gamer, 2018).

5.2. Evaluation of Presence of Stress

We use the following set of features for our baseline models. Given that stress vs non-stressed state is classified for the complete section (monologue or neutral recording), we extract visual features differently to use the the sequential information over the whole segment, i.e., a monologue. We also use physiological features for our network, since we found that even though they are highly variable over shorter segments (utterances), they are informative for recognizing physiological state on a whole section.

1. **Acoustic, Lexical, and Thermal Features.** We use the same features as extracted for predicting emotion.
2. **Wide-angle Video Features.** We extract the subject's pose using OpenPose (Cao et al., 2017; Simon et al., 2017; Wei et al., 2016) at 25 frames per second. For each frame, we extract 14 three-dimensional points representing anchor points for the upper body. For classification of each 3D point is interpolated over one second using a 5^{th} order spline (Oikonomopoulos et al., 2008; Huang and Cohen, 1993). The parameters of the splines are then used as features for classification.
3. **Close-up Video Features.** We use OpenFace to extract the subject's action units (Baltrušaitis et al., 2016). The features are extracted for every frame. In each frame, features include the gaze direction vectors, gaze angles, 2D eye region landmarks, head locations, rotation angles of the head, landmark locations, and facial action

Table 6: Baseline results for classifying stressed and non-stressed situations per time unit, unless specified otherwise. A - Accuracy, P - Precision, R - Recall.

Recording Parts	A	P	R	F_1
	Thermal			
Neutral	**0.61**	**0.67**	**0.62**	**0.64**
Questions	0.50	0.64	0.52	0.57
	Wide-angle Video			
Neutral	0.66	0.41	0.96	0.58
Questions	0.69	0.45	0.82	0.58
	Close-up Video			
Neutral	0.61	0.78	0.33	0.46
Questions	0.65	0.65	0.69	0.67
	Physiological			
Neutral	0.66	0.47	0.89	0.64
Questions	0.70	0.55	0.88	0.67
	Audio - Per utterance			
Questions	0.67	0.70	0.69	0.69
	Text - Per utterance			
Questions	0.60	0.74	0.61	0.67
	Late Fusion - Voting			
Questions	0.60	0.74	0.61	0.67

units. Landmarks locations offset by the nose location. We window the data into segments of one-second windows with 0.5 second overlap and calculate summary statistics (maximum, minimum, mean, variance). We retain the top 300 features based on the F values between the training features and corresponding labels (stressed vs non-stressed).

4. **Physiological Features.** While the physiological features varied greatly per second to be informative for emotion, they are informative for recognizing presence or absence of stress. We consider the raw measurements for heart rate, breathing rate, skin conductance and skin temperature and compute the first four measures of central tendency, e.g. mean, standard deviation, skewness, and kurtosis.

Network. We train a DNN to perform binary classification, i.e., to recognize stressed vs. non-stressed situation using ReLU as activation, with softmax as the classification method.The final layer uses a soft-max activation. We train six different networks for thermal, wide-angle video, close-up video, physiological, audio, and lexical modalities. Each network is trained in a subject-independent manner. We train network to recognize stress vs non-stress situation in both neutral recording,i.e., when the subject isn't speaking at the beginning of the recording, and during emotional monologue questions. To do so, we decide the final prediction by a majority vote over one-second predictions for the complete section of the recording. For the lexical and acoustic modality, we train the network for the question monologues, and decide the final prediction based on a majority vote over prediction for each utterance.

Results. We report our results for prediction of stress vs non-stress situation using various modalities in Table 6. We see that the captured modalities are indeed informative for recognizing stress vs non-stressed situations. We find that for recognizing this distinction when the subjects are speaking, audio and physiological features perform the best. This is in agreement with previous related work (Lazarus and Cohen, 1977; Yaribeygi et al., 2017; Horvath, 1978). Interestingly, we also find that the thermal and physiological modality is apt at recognizing differences in stress, even in the neutral recording, i.e., when the subject is not speaking. This advantage of thermal modality has been previously documented by researchers (Abouelenien et al., 2014; Pavlidis et al., 2000; Pavlidis and Levine, 2002; Garbey et al., 2007). We find that answering emotional monologue questions interferes with the recorded thermal modality, leading to a poorer performance at stress recognition.

6. Conclusions and Future Work

In this paper, we introduced a dataset that aims to capture the interplay between psychological factors such as stress and emotion. While various other datasets have explored the relationship between gender or personality measures and emotion production and perception, the relationship between psychological factors and emotion is understudied from a data collection point of view, and hence an automated modeling perspective.

We verified that the presence of emotion and stress can be detected in our dataset. Our baseline results for emotion classification using DNNs with acoustic, linguistic and visual features on our dataset are similar to reported results on other datasets such as IEMOCAP (Busso et al., 2008) and MSP-Improv (Busso et al., 2017). For classifying stressed vs non-stressed session, we observed that all modalities are discriminative, although at different levels of accuracy; for instance, visual and physiological modalities work best for recognizing stress under emotional influence.

Through our experiments, we found that the modalities that we use for the prediction of emotion are also good predictors of stress. The acoustic features are highly informative for both activation and stress, which is concurrent with previous research (Paulmann et al., 2016) that showed how speech patterns are heavily modulated in the presence of adversarial psychological factors. We know that the emotion representations obtained from our classification models would likely be interleaved with the distributions of stress (Chattopadhyay et al., 2019). Hence, the corpus is especially useful in developing models that effectively predict emotional states while accounting for the presence of these confounders (Jaiswal et al., 2019b). We anticipate that this will lead to robust emotion recognition models that generalize to emotion perception under varying conditions of production.

In the future, we plan to conduct additional annotation experiments using randomized blocks of utterances from single speakers (random order of utterances from random sections). This will better match the block of single-speaker tasks in context, where the annotator can adapt to the speaker but not to the topic of conversation. We will also explore the causal interaction between emotion production and the stress levels of the subjects using personalized and controlled modeling techniques. We are interested in uncovering how these stress measures affect each subject differently

and if this relates to personality. Through this dataset, we hope the community can make progress on understanding the correlation patterns between the distribution of emotion and stress, and how this can impact the performance of emotion and stress classifiers.

The MuSE dataset will be publicly available at: `http://lit.eecs.umich.edu/downloads.html`.

7. Acknowledgements

This material is based in part upon work supported by the Toyota Research Institute ("TRI") and by the National Science Foundation (NSF CAREER #1651740 and IIS #1815291). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, TRI, or any other Toyota entity.

8. References

Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., and Burzo, M. (2014). Deception detection using a multimodal approach. In Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, pages 58–65, New York, NY, USA. ACM.

Abouelenien, M., Mihalcea, R., and Burzo, M. (2016). Analyzing thermal and visual clues of deception for a non-contact deception detection approach. In Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, page 35. ACM.

Aguiar, A., Kaiseler, M., Cunha, M., Meinedo, H., Almeida, P. R., and Silva, J. (2014). Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments. In LREC 2014, Ninth International Conference on Language Resources and Evaluation.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., and Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23(4):363–377.

Audibert, N., Aubergé, V., and Rilliard, A. (2010). Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study. In Speech Prosody 2010-Fifth International Conference.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–10. IEEE.

Banda, N. and Robinson, P. (2011). Noise analysis in audio-visual emotion recognition. In Proceedings of the 11th International Conference on Multimodal Interaction (ICMI).

Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

Bartolini, E. E. (2011). Eliciting emotion with film: Development of a stimulus set.

Batliner, A., Kompe, R., Kießling, A., Nöth, E., and Niemann, H. (1995). Can you tell apart spontaneous and read speech if you just look at prosody? In *Speech Recognition and Coding*. Springer, pp. 321–324.

Berger, R. H., Miller, A. L., Seifer, R., Cares, S. R., and LeBourgeois, M. K. (2012). Acute sleep restriction effects on emotion responses in 30-to 36-month-old children. *Journal of sleep research*, 21(3):235–246.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1042–1047.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Bruno, P., Melnyk, V., and Völckner, F. (2017). Temperature and emotions: Effects of physical temperature on responses to emotional advertising. *International Journal of Research in Marketing*, 34(1):302–320.

Burmania, A. and Busso, C. (2017). A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors. In INTERSPEECH, pages 152–156.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2017). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Cambria, E., Hazarika, D., Poria, S., Hussain, A., and Subramanyam, R. (2017). Benchmarking multimodal sentiment analysis. In International Conference on Computational Linguistics and Intelligent Text Processing, pages 166–179. Springer.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In CVPR.

Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. N. (2019). Neural network attributions: A causal perspective. *arXiv preprint arXiv:1902.02302*.

Chen, S., Hsu, C., Kuo, C., Huang, T. K., and Ku, L. (2018). Emotionlines: An emotion corpus of multi-party conversations. *CoRR*, abs/1802.08379.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., et al. (2019). What makes a good conversation?: Challenges in designing truly conversational agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 475. ACM.

Cohen, S., Kamarck, T., Mermelstein, R., et al. (1994). Perceived stress scale. *Measuring stress: A guide for health and social scientists*, pages 235–283.

Cohen, S. (1988). Perceived stress in a probability sample of the united states.

Dalpé, J., Demers, M., Verner-Filion, J., and Vallerand, R. J. (2019). From personality to passion: The role of the big five factors. *Personality and Individual Differences*, 138:280–285.

Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pages 1459–1462. ACM.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Fernández-Dols, J.-M. and Crivelli, C. (2013). Emotion and expression: Naturalistic studies. *Emotion Review*, 5(1):24–29.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Garbey, M., Sun, N., Merla, A., and Pavlidis, I. (2007). Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE transactions on Biomedical Engineering*, 54(8):1418–1426.

Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Jauregui, D. A. G., Eyharabide, V., Filaire, E., Le Scanff, C., Devillers, L., et al. (2013). Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses. In Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, pages 149–156. IEEE.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Griffiths, P. E. (2003). Iii. basic emotions, complex emotions, machiavellian emotions 1. *Royal Institute of Philosophy Supplements*, 52:39–67.

Gulli, A. and Pal, S. (2017). Deep Learning with Keras. Packt Publishing Ltd.

Herborn, K. A., Graves, J. L., Jerem, P., Evans, N. P., Nager, R., McCafferty, D. J., and McKeegan, D. E. (2015). Skin temperature reveals the intensity of acute stress. *Physiology & behavior*, 152:225–230.

Horton, J. J. and Chilton, L. B. (2010). The labor economics of paid crowdsourcing. In Proceedings of the 11th ACM conference on Electronic commerce, pages 209–218. ACM.

Horvath, F. (1978). An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. *Journal of Applied Psychology*, 63(3):338.

Horvath, F. (1982). Detecting deception: the promise and the reality of voice stress analysis. *Journal of Forensic Science*, 27(2):340–351.

Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing, pages 27–35. Association for Computational Linguistics.

Huang, Z. and Cohen, F. S. (1993). 3-d motion estimation and object tracking using b-spline curve modeling. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 748–749, June.

Jaiswal, M., Tabibu, S., and Bajpai, R. (2016). The truth and nothing but the truth: Multimodal analysis for deception detection. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pages 938–943. IEEE.

Jaiswal, M., Aldeneh, Z., Bara, C., Luo, Y., Burzo, M., Mihalcea, R., and Provost, E. M. (2019a). Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. *CoRR*, abs/1903.11672.

Jaiswal, M., Aldeneh, Z., and Mower Provost, E. (2019b). Controlling for confounders in multimodal emotion classification via adversarial learning. In 2019 International Conference on Multimodal Interaction, pages 174–184. ACM.

Jiang, Y.-G., Xu, B., and Xue, X. (2014). Predicting emotions in user-generated videos. In Twenty-Eighth AAAI Conference on Artificial Intelligence.

Jürgens, R., Grass, A., Drolet, M., and Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of nonverbal behavior*, 39(3):195–214.

Khorram, S., Jaiswal, M., Gideon, J., McInnis, M., and Provost, E. M. (2018). The priori emotion dataset: Linking mood to emotion detected in-the-wild. *arXiv preprint arXiv:1806.10658*.

Kingston, C. and Schuurmans-Stekhoven, J. (2016). Life hassles and delusional ideation: Scoping the potential role of cognitive and affective mediators. *Psychology and Psychotherapy: Theory, Research and Practice*, 89(4):445–463.

Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. (1993). The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81.

Kolář, J. (2008). Automatic Segmentation of Speech into Sentence-like Units. Ph.D. thesis, University of West Bohemia in Pilsen.

Kurniawan, H., Maslov, A. V., and Pechenizkiy, M. (2013). Stress detection from speech and galvanic skin response signals. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pages 209–214. IEEE.

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., Elfström, S., Råde, A., Golan, O., Bölte, S., Baron-Cohen, S., and Lundqvist, D. (2019). The eu-emotion voice database. *Behavior Research Methods*, 51(2):493–506, Apr.

Lazarus, R. S. and Cohen, J. B. (1977). Environmental stress. In *Human behavior and environment*. Springer, pp. 89–127.

Lech, M. and He, L. (2014). Stress and emotion recognition using acoustic speech analysis. In *Mental Health Informatics*. Springer, pp. 163–184.

Lee, M., Bang, S., and Yang, G. (2011). Apparatus and method for inducing emotions, June 7. US Patent 7,955,259.

Li, W., Abtahi, F., Tsangouri, C., and Zhu, Z. (2016). Towards an "in-the-wild" emotion dataset using a game-based framework. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1526–1534. IEEE.

Liao, L.-M. and Carey, M. G. (2015). Laboratory-induced mental stress, cardiovascular response, and psychological characteristics. *Reviews in cardiovascular medicine*, 16(1):28–35.

Lichtenauer, J. and Soleymani, M. (2011). Mahnob-hci-tagging database.

Lotfian, R. and Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*.

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Mills, C. and D'Mello, S. (2014). On the validity of the autobiographical emotional memory task for emotion induction. *PloS one*, 9(4):e95837.

Mitchell, L. A. (2006). The relationship between emotional recognition and personality traits.

Monin, J. K., Schulz, R., Lemay Jr, E. P., and Cook, T. B. (2012). Linguistic markers of emotion regulation and cardiovascular reactivity among older caregiving spouses. *Psychology and aging*, 27(4):903.

Nasir, M., Baucom, B. R., Georgiou, P., and Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PloS one*, 12(9):e0185123.

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Oikonomopoulos, A., Pantic, M., and Patras, I. (2008). Human gesture recognition using sparse b-spline polynomial representations. In Proceedings of Belgium-Netherlands Conf. Artificial Intelligence (BNAIC 2008), Boekelo, The Netherlands, pages 193–200.

Paulmann, S., Furnes, D., Bøkenes, A. M., and Cozzolino, P. J. (2016). How psychological stress affects emotional prosody. *Plos one*, 11(11):e0165022.

Pavlidis, I. and Levine, J. (2002). Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6):56–64.

Pavlidis, I., Levine, J., and Baukol, P. (2000). Thermal imaging for anxiety detection. In Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No. PR00640), pages 104–109. IEEE.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Querengässer, J. and Schindler, S. (2014). Sad but true?-how induced emotional states differentially bias self-rated big five personality traits. *BMC Psychology*, 2(1):14.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pages 1–8. IEEE.

Rothkrantz, L. J., Wiggers, P., Van Wees, J.-W. A., and van Vark, R. J. (2004). Voice stress analysis. In International conference on text, speech and dialogue, pages 449–456. Springer.

Rubo, M. and Gamer, M. (2018). Social content and emotional valence modulate gaze fixations in dynamic scenes. *Scientific reports*, 8(1):3804.

Samson, A. C., Kreibig, S. D., Soderstrom, B., Wade, A. A., and Gross, J. J. (2016). Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and emotion*, 30(5):827–856.

Schlotz, W., Yim, I. S., Zoccola, P. M., Jansen, L., and Schulz, P. (2011). The perceived stress reactivity scale: Measurement invariance, stability, and validity in three countries. *Psychological assessment*, 23(1):80.

Siedlecka, E. and Denson, T. F. (2019). Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, 11(1):87–97.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In CVPR.

Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., and Griss, M. (2012). Activity-aware mental stress detection using physiological sensors. In Martin Gris et al., editors, Mobile Computing, Applications, and Services, pages 282–301, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thompson, E. H., Robertson, P., Curtis, R., and Frick, M. H. (2013). Students with anxiety: Implications for professional school counselors. *Professional School Counseling*, 16(4):2156759X150160402.

Tulen, J., Moleman, P., Van Steenis, H., and Boomsma, F. (1989). Characterization of stress reactions to the stroop color word test. *Pharmacology Biochemistry and Behavior*, 32(1):9–15.

Tull, M. T., Barrett, H. M., McMillan, E. S., and Roemer, L. (2007). A preliminary investigation of the relationship between emotion regulation difficulties and posttraumatic stress symptoms. *Behavior Therapy*, 38(3):303–313.

Wang, M. and Saudino, K. J. (2011). Emotion regulation and stress. *Journal of Adult Development*, 18(2):95–103.

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., and Kissler, J. (2017). Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, 12(5):e0177239.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In CVPR.

Werner, K. H., Goldin, P. R., Ball, T. M., Heimberg, R. G., and Gross, J. J. (2011). Assessing emotion regulation in social anxiety disorder: The emotion regulation interview. *Journal of Psychopathology and Behavioral Assessment*, 33(3):346–354.

Yang, Y.-H. and Chen, H. H. (2011). Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2184–2196.

Yannakakis, G. N., Cowie, R., and Busso, C. (2017). The ordinal nature of emotions. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 248–255. IEEE.

Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T. P., and Sahebkar, A. (2017). The impact of stress on body function: A review. *EXCLI journal*, 16:1057.

You, Q., Luo, J., Jin, H., and Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In Thirtieth AAAI Conference on Artificial Intelligence.

Yuan, X. (2015). An approach to integrating emotion in dialogue management. In International Conference in Swarm Intelligence, pages 297–308. Springer.

Zhang, X., Yu, H. W., and Barrett, L. F. (2014). How does this make you feel? a comparison of four affect induction procedures. *Frontiers in psychology*, 5:689.

Zhao, S., Ding, G., Han, J., and Gao, Y. (2018). Personality-aware personalized emotion recognition from physiological signals. In IJCAI, pages 1660–1667.

Zuo, X. and Fung, P. (2011). A cross gender and cross lingual study on acoustic features for stress recognition in speech. In ICPhS, pages 2336–2339.