METHODS
published: 06 December 2021
doi: 10.3389/fcomp.2021.750284

An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress

Alice Baird[1]*, *Andreas Triantafyllopoulos*[1,2], *Sandra Zänkert*[3], *Sandra Ottl*[1], *Lukas Christ*[1], *Lukas Stappen*[1], *Julian Konzok*[3], *Sarah Sturmbauer*[4], *Eva-Maria Meßner*[5], *Brigitte M. Kudielka*[3], *Nicolas Rohleder*[4], *Harald Baumeister*[5] and *Björn W. Schuller*[1,2,6]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2]AudEERING GmbH, Gilching, Germany, [3]Institute of Psychology, University of Regensburg, Regensburg, Germany, [4]Chair of Health Psychology, FAU Erlangen-Nuremberg, Erlangen, Germany, [5]Chair of Clinical Psychology and Psychotherapy, University of Ulm, Ulm, Germany, [6]GLAM—Group on Language, Audio, and Music, Imperial College London, London, United Kingdom

OPEN ACCESS

Edited by:
Marwa Mahmoud,
University of Cambridge,
United Kingdom

Reviewed by:
Robertas Damasevicius,
Silesian University of Technology,
Poland
Ronald Böck,
Otto von Guericke University
Magdeburg, Germany

***Correspondence:**
Alice Baird
alicebaird@ieee.org

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *30 July 2021*
Accepted: *29 October 2021*
Published: *06 December 2021*

Citation:
Baird A, Triantafyllopoulos A, Zänkert S, Ottl S, Christ L, Stappen L, Konzok J, Sturmbauer S, Meßner E-M, Kudielka BM, Rohleder N, Baumeister H and Schuller BW (2021) An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress. Front. Comput. Sci. 3:750284. doi: 10.3389/fcomp.2021.750284

Life in modern societies is fast-paced and full of stress-inducing demands. The development of stress monitoring methods is a growing area of research due to the personal and economic advantages that timely detection provides. Studies have shown that speech-based features can be utilised to robustly predict several physiological markers of stress, including emotional state, continuous heart rate, and the stress hormone, cortisol. In this contribution, we extend previous works by the authors, utilising three German language corpora including more than 100 subjects undergoing a Trier Social Stress Test protocol. We present cross-corpus and transfer learning results which explore the efficacy of the speech signal to predict three physiological markers of stress—sequentially measured saliva-based cortisol, continuous heart rate as beats per minute (BPM), and continuous respiration. For this, we extract several features from audio as well as video and apply various machine learning architectures, including a temporal context-based Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). For the task of predicting cortisol levels from speech, deep learning improves on results obtained by conventional support vector regression—yielding a Spearman correlation coefficient (ρ) of 0.770 and 0.698 for cortisol measurements taken 10 and 20 min after the stress period for the two corpora applicable—showing that audio features alone are sufficient for predicting cortisol, with audiovisual fusion to an extent improving such results. We also obtain a Root Mean Square Error (RMSE) of 38 and 22 BPM for continuous heart rate prediction on the two corpora where this information is available, and a normalised RMSE (NRMSE) of 0.120 for respiration prediction (–10: 10). Both of these continuous physiological signals show to be highly effective markers of stress (based on cortisol grouping analysis), both when available as ground truth and when predicted using speech. This contribution opens up new avenues for future exploration of these signals as proxies for stress in naturalistic settings.

Keywords: affective computing, stress, computer audition, paralinguistics, multimodal

1 INTRODUCTION

Understanding how stress manifests in the human body has several meaningful use-cases, from improving safety during driving (Bianco et al., 2019) to early intervention of neurodegeneration (Zafar, 2020). Stress levels are globally on the rise, primarily due to the increased pressure from work and personal lifestyles (Sharma et al., 2021). Many individuals find themselves constantly dealing with several concurrent tasks, a feat known to put well-being off-balance, particularly during work (Pagán-Castaño et al., 2020). With this in mind, methods that can reduce levels of stress whilst still enabling the desired level of efficiency are highly desirable in workplace environments as they can be used to proactively prevent burnout, which is known to proceed consistent stress (Fendel et al., 2020). During a stress inducing situation, the adrenal glands begin to produce various hormones, of which cortisol (a stress hormone) is the most prominent (Leistner and Menke, 2020). The production of cortisol responds to the activation of the hypothalamic-pituitary-adrenal (HPA) axis, which begins to secrete the corticotropin-releasing hormone that causes the additional release of the adrenocorticotrophic hormone (ACTH) from the pituitary. The release of such hormones is known to alter other physiological responses, including heart rate (Gönülateş et al., 2017), which in turn affects face colouring (Niu et al., 2018) and speech, particularly during psychosocial stress (Brugnera et al., 2018). With this in mind, the speech signal can (non-intrusively) computationally monitor several states of wellbeing (Cummins et al., 2018). It has shown promise in recent studies to indicate physiological signals which are known to be markers of stress, e. g., correlation with saliva-based cortisol samples (Baird et al., 2019), states of emotional arousal (Stappen et al., 2021a), and co-occurring conditions including anxiety (Baird et al., 2020).

In this study, we extend previous works by the authors (Baird et al., 2019), by more deeply exploring the utility of speech for monitoring stress. We use three German corpora, the FAU, ULM- and REG-TSST which were all gathered with the renowned Trier Social Stress Test (TSST) protocol (Kirschbaum et al., 1993), and contain more than 100 subjects in total. In previous studies utilising the FAU-TSST dataset (Baird et al., 2019), speech derived features were found to strongly correlate with raw cortisol taken 10 and 20 min after the spoken task in the TSST, which supported the use of speech as a marker of stress, mainly as cortisol is known to peak between 10 and 20 min after a stress stimulus (Goodman et al., 2017). With this in mind, we aim to more closely explore the connection between spoken features and sequential cortisol samples extracted from saliva. To do this, we will perform a fundamental acoustic analysis of grouped signals and then, via a deep learning recognition paradigm, explore each corpus applying transfer learning to validate the efficacy of speech-based cortisol recognition on unseen data. Furthermore, as the ULM- and REG-TSST corpora both contain continuous heart rate as beats per minute (BPM), and ULM-TSST additionally includes respiration signals, we aim to recognise these signals and explore their relationship with the saliva-based cortisol samples to validate their use as markers of stress. There are also two speech scenarios within a TSST, the job interview and arithmetic task, separating these—we will also explore how the speech duration and activation in general effects recognition rates with a more fine-grained continuous analysis. Finally, we utilise visual-based features (where available) for multimodal recognition of relevant stress bio-markers and compare the performance of this to audio.

To summarise, the following analysis includes several insights and contributions. At the core, this work extends on previous results by the authors Baird et al. (2019), and for the first time, explores the task of sequentially sampled cortisol prediction from multimodal data within a deep learning-based architecture, validating the experimental paradigm via utility of a novel dataset. Therefore the areas explored through the utilisation of sequential cortisol as ground truth for stress are fourfold: 1) The utility of speech plus multimodal features for recognition of other physiological-derived signals is validated or not. 2) A fundamental acoustic analysis of speech under stress is conducted, utilising cortisol derived groupings as a ground truth for stress. 3) Multi-domain experiments are conducted to further validate previous works' findings with newly collected data.

This article is organised as follows; firstly, in **Section 2**, we provide a brief overview of related studies in the area of stress recognition. We then introduce the three corpora that have been used within the experiments in cf. **Section 3**, as well as offering detail of the TSST study procedure in general. This is followed by an acoustic analysis of each corpus in **Section 4**. We then outline the experimental set-up for the recognition tasks, in **Section 5**, and present our experimental results in **Section 6**, with a mention of study limitation in **Section 7**. Finally, we offer concluding remarks and future outlook in **Section 8**.

2 RELATED WORK

Stress recognition has been an active research area within the machine learning and affective computing communities for several years, thus making an extensive summary of this area of research beyond the scope of the current work. In this section, we discuss various approaches which have motivated aspects of our work and would suggest that the interested reader is directed to Panicker and Gayathri (2019) or Grzadzielewska (2021), for a survey on stress recognition in general, and to Garcia-Ceja et al. (2018), for mental health state recognition using machine learning.

As mentioned, speech as a marker of stress was explored in Baird et al. (2019), and sequentially measured cortisol samples were for the first time recognised in a traditional machine learning paradigm utilising a support vector regressor (SVR) with hand-crafted and image-based speech-derived features. Findings from this study showed that elevated cortisol levels—taken between 10 and 20 min after the TSST, i. e., the time of speech under stress—correlate to a substantial level (Spearman's correlation coefficient (ρ) of at best 0.421) with hand-crafted prosodic related feature sets performing best.

Aside from the work presented in Baird et al. (2019), there have been limited computational machine learning-based works which have looked at sequentially samples saliva-based cortisol. However, in Nath et al. (2021), the authors aim to provide a system for monitoring stress in older subjects and in this case instead of explicitly recognising the raw cortisol values, they utilise the samples to produce a ground truth for stress or no stress, in order to perform a binary classification on the subjects. Instead of speech-derived features, the authors perform various experiments based on the features extracted from wearable sensors, e. g., blood volume pressure and electrodermal activity. In this study, in particular, the authors find substantial improvement through the use of an LSTM-based deep learning architecture, obtaining an accuracy for the 2-class problem above 90% F^1 via the feature selection of physiological based signals. Such results show promise for the use of deep learning in the context of stress recognition.

There are several machine learning approaches that have explored physiologically derived markers in general for stress recognition (MacLaughlin et al., 2011; Dhama et al., 2019; Šalkevicius et al., 2019). From feature-based machine learning paradigms which classify various features extracted from wearable sensors, i. e., sleep quality, and percentage of screen time (Sano and Picard, 2013), or heart rate variability (HRV) (Dalmeida and Masala, 2021), and thermal-video recognition of the Initial Systolic Time Interval (Kumar S. et al., 2021), applying the state-of-the-art StressNet. StressNet consists of a Long Short-Term Memory (LSTM)-based architecture to harness spatial-temporal aspects of a continuous signal. Similarly, in a recent study, the DeepBreath system has been presented (Cho et al., 2017), a CNN-based architecture which was applied to small-scale datasets for stress recognition, and obtains up to 84.59% accuracy for a binary stress task and 56.52% for a 3-class problem. Kumar A. et al. (2021) present a hierarchical deep neural network that learns high-level feature representations for each type of physiological signal.

The use-cases associated with these approaches vary, with works in recent years being targeted at products including driver monitoring (Healey and Picard, 2005). However, a major limitation for such stress research is that stress can be potentially harmful to individuals, thus raising ethical concerns which make the collection of spontaneous and natural stress occurrences difficult in practice. With this in mind, the TSST is a standardised and common paradigm (Schmidt et al., 2018), which some stress targeted corpora have applied as it is known to induce moderate psychosocial stress to subjects (Dickerson and Kemeny, 2004; Plarre et al., 2011). Smaller-scale datasets following these established protocols and have been collected and used for machine learning-based stress recognition (Cuno et al., 2020). The SWELL dataset (Koldijk et al., 2014) (25 subjects, 8 female), is one where *time-pressure* and *interruptions* are integrated in the task which the subjects are asked to perform. In a similar way, cognitive load is another method for inducing stress, and in the renowned SUSAS (Hansen and Bou-Ghazale, 1997) corpora (aimed at robust speech processing from stressed and emotional speech), 32 subjects are perform various "tracking" tasks, which increase in their complexity.

Several studies based on these available datasets utilise classical machine learning methods to explore the relationship of multimodal features with stress. In Rodríguez-Arce et al. (2020), the authors apply a Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Random Forest and Logistic Regression (LogR) classifiers to analyse the accuracy of feature subsets based on various modalities, e. g., heart rate, respiration, and galvanic skin response. The limited available data makes deep learning approaches a challenge, however in the 2021 Multimodal Sentiment Analysis in Real-Life Media Challenge (MuSe) (Stappen et al., 2021a), the ULM-TSST corpus was presented and successfully utilised for emotion-based stress recognition during a TSST. The baseline for the *Multimodal Emotional Stress sub-challenge* (MUSE-STRESS) task (recognition of valence and arousal during stress) applies an LSTM-RNN with a late multimodal fusion of audio plus video-based features, obtaining a concordance correlation coefficient (CCC) of 0.509 (for combined arousal and valence). Audio features perform best for the uni-modal approaches in the MuSe paradigm, with EGEMAPS (Eyben et al., 2016) features yielding a CCC of 0.472, compared to a CCC of 0.305 for video-based VGGFACE features.

From this literature overview, it is clear that there is missing analysis in the literature, and need to explore more deeply the utility of markers of stress e. g., cortisol, in a machine learning paradigm. Computational understanding of cortisol is particularly meaningful, as it is known that, sustained levels of stress are substantial contributors to neurodegeneration (Zafar, 2020), with biological markers of this including fluctuations in neurotransmitters, e. g., dopamine or serotonin, and levels of stress hormones including cortisol, with Zafar (2020). More specifically, in Saitis and Kalimeri (2018) the authors use related bio-markers to automatically detect environments that are stressful for visually impaired persons which might help to improve accessibility within public spaces. This illustrates that successful monitoring of stress via such markers has a benefit beyond commercial applications.

Furthermore, as can be seen, current studies are largely based on smaller-scale corpora (ca. 30 subjects), with the current contribution attempting to go deeper by not only exploring across multiple corpora but in general including a more substantial number of speakers (+100) than is typically observed in the literature thus far. As well as this, applying deep learning, particularly an LSTM-RNN, appears to be a valid deep learning architecture for modelling states of continuous stress, and motivates us to explore the use of this in comparison to more robust models, e. g., the SVR. Finally, in Baird et al. (2019) no other modalities were explored for recognising the cortisol-derived markers of stress, neither in a uni- or multimodal manner, and so this strongly motivates the current work to explore how vision-based features perform in this setting.

3 CORPORA

For our experiments we utilise three corpora—the FAU-Trier Social Stress Test (FAU-TSST), the Regensburg-Trier Social Stress Test (REG-TSST), and the Ulm-Trier Social Stress Test (ULM-TSST)—which all include subjects undergoing the renowned and highly standardised Trier Social Stress Test (TSST) (Kirschbaum et al., 1993). All subjects were speaking in the German language and were recorded at Universities from southern German states (Bavaria and Baden-Württemberg). After processing, the total

TABLE 1 | An overview of each of the three corpora (FAU) (REG) and (ULM)-TSST used within this contribution. Including, number of subjects (#), distribution of gender (M)ale: (F) emale, Age in years (mean/standard deviation), continuous signals available for each—(A)udio, (V)ideo, heart rate as beats per minute (B)PM, (R)espiration, (C)ortisol and (E)motion (arousal and valence)—as well as, the speaker independent partitions, train, (dev)elopment and test, and the duration of audio data, after voice activity detection (VAD) and for each TSST task, (Inter)view, and (Arith)methic.

	# (M:F)	Age μ/±	Modes						Duration (hh: mm)				Partitions			
			A	V	B	R	C	E		VAD	Inter.	Arith.	Train	Dev	Test	∑
FAU	43 (14:29)	24.26/4.97	☒	☒	□	□	☒	□	7: 25	4: 20	2: 32	1: 48	15	15	13	43
REG	27 (13:14)	22.74/2.96	☒	□	☒	□	☒	□	4: 28	2: 26	1: 24	1: 02	10	9	8	27
ULM	69 (20:49)	25.06/4.48	☒	☒	☒	☒	□	☒	5: 47	2: 21	2: 21	–	41	14	14	69

amount of speakers is 134, including 50 males and 84 females. The FAU-TSST corpus was first introduced in Baird et al. (2019), and the ULM-TSST corpus in Stappen et al. (2021a). In **Table 1**, we provide an overview of all data available across each corpus. As can be seen, the only modality available in all corpora is audio. All three corpora have speech data from the job interview (interview) task (described in detail below), however the ULM-TSST corpus does not include the arithmetic task.

3.1 The Trier Social Stress Test Procedure

Each testing site obtained ethical approval from their respective university's ethics committee to perform the TSST study. In all cases, subjects were recruited from the university campus and the community via print and multi-media advertising and received monetary compensation. The study was carried out in accordance with the declaration of Helsinki, and informed consent was obtained from all subjects at study entry. For the REG-TSST eligible, subjects were then invited to a first laboratory session to conduct a structured clinical interview (Wittchen et al., 1997) for exclusion of acute or chronic psychiatric diseases. Further exclusion criteria applied to all corpora included; acute or chronic somatic diseases, psychotropic or glucocorticoid medication intake, BMI above 30 kg/m^2, drug abuse, and previous experience with the TSST procedure.

For all corpora, the participants did not know the details of the tasks and were given this information upon entering the TSST study room. The prior experience that subjects may have had with these styles of speaking tasks is unknown, although they were not informed of the task details prior to entering the test site. For the interview task, the participants were not restricted to a particular vacant position but rather considered it to be the interview for their 'dream roll'. Furthermore, it is unknown how many participants had a prior relationship with the panel, although there is likely some previous acquaintance-level relationship given the university location.

In **Figure 1** an timeline is given for the general TSST experiment. There was slight variance at each test site; however, we attempt to combine the description of procedure. The TSSTs were scheduled between 12: 00 p.m. and 7: 00 p.m. to account for the influence of circadian cortisol variations (Rohleder and Nater, 2009). Instructions for the subjects included instructions to refrain from exercising, smoking, teeth brushing, eating, and drinking anything except water before the arrival. Upon arrival, subjects received verbal and written instructions, followed by a resting period. During this time, for the FAU-TSST and REG-TSST a saliva sample (S0 30–45 min before TSST) was collected as the participant's cortisol baseline, and for the REG-TSST corpus, a sugary drink (chilled herbal tea with 75 g of glucose) was given to elevate blood glucose levels (Zänkert et al., 2020). One minute before the next stage, another saliva sample is taken (S1 1 min). The subjects are then introduced to the TSST procedure, and guided to a test room, and introduced to observers wearing white lab coats. Subjects were then instructed to take the role of a job applicant and give a 5-min speech to present themselves as the best candidate for a vacant position. This task is where continuous recording begins for the REG-TSST and ULM-TSST[1]. After this, in the FAU-TSST and REG-TSST corpora, subjects were given a mental arithmetic task, for a further 5 min, where they should serially subtract 17 from 2 043 as quickly as possible. In the case of any error, they were requested to start again. After completion of the TSST speaking tasks, six more saliva based samples are taken from the subjects (S2-S7).

3.2 Target Signals

As seen in **Table 1** overview, there are several signals available for each of the three corpora. As a core task, we focus on the recognition of sequential saliva-based cortisol measures S0 (45 min) to S7 (+60 min), measured in nanomoles per litre (nmol/L). For the FAU-TSST and REG-TSST corpora, saliva is collected at the same time-points, before and after the TSST, and stored at −20°C before extraction. However, for each corpus, the assay (i. e., biochemical analysis procedure) applied to extract cortisol varied, where FAU-TSST utilise a chemiluminescence immunoassay (CLIA), and REG-TSST a fluorescence-based immunoassay (DELFIA) meaning that the derived cortisol value is not completely comparable, for further detail on the difference in these procedures the interested reader is directed to Miller et al. (2013). With this in mind for the experiments in later sections the two corpora will only be utilised in a multi-domain manner, and not with a typical cross-corpus strategy, cf. **Section 6**.

Given this, we first want to analyse the variance in raw cortisol between the two corpora, and so we apply a repeated measures' analysis of variance (RM-ANOVA) with raw cortisol (S0-S7) as within-subject factor time and the between-subject factor corpora (FAU-TSSTvs REG-TSST). Due to lack of sphericity (pointing to unequal variances of within-subject measures) we report the Greenhouse-Geisser adjusted p-value. We find a significant main effect of the corpora [F (1, 67) = 4.02, p = 0.049, η^2 = 0.03] indicating that on average FAU-TSST raw cortisol is higher compared to REG-TSST raw cortisol. Further, we see a significant time × corpora interaction [F (1.76, 120.08) = 4.52, p = 0.016, η^2 = 0.017] with a slightly earlier and higher rise in raw

[1] For physiological signals, the REG-TSST corpus utilised the Polar RS800CX and V800 system, and the ULM-TSST corpus used the BIOPAC Systems, MP35.

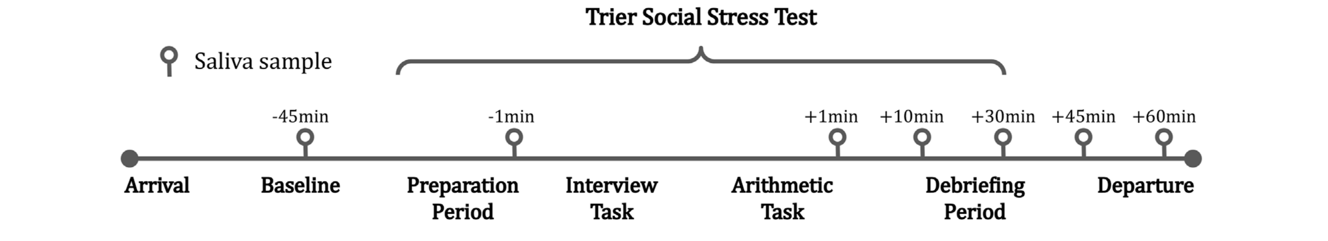

FIGURE 1 | An overview of the typical TSST paradigm, as applied in FAU-TSST and REG-TSST, where ULM-TSST excludes the arithmetic task.

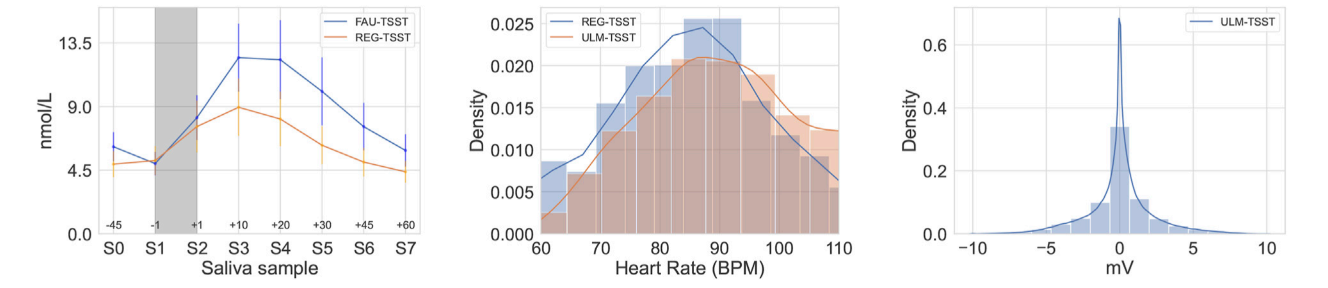

FIGURE 2 | Mean of raw cortisol as (nmol/L) for both FAU-TSST and REG-TSST, highlighting the stress period in grey, with annotations of sequential saliva (S), and sample time in minutes **(left)**. Distribution of continuous heart rate as BPM for the REG-TSST and ULM-TSST corpora **(middle)**. Distribution of the respiration-based signal as millivolts (mV) based on chest displacement **(right)**.

cortisol in FAU-TSSTcompared to REG-TSST. Also, testing the homogeneity of variances of S0—S7 with the Levene's Test reveals that for S0-S2, we can assume homogenous variances ($p > 0.1$) whereas for S3S7, we see inhomogeneous variances ($p < 0.05$). Whereas variances are comparable for S0-S2, for S3-S7 variances in the FAU-TSSTcorpora for raw cortisol are higher compared to REG-TSST. This suggests a large difference between both corpora regarding intra-individual cortisol trajectories in response to the TSST. For an overview of the raw cortisol in each corpus, cf. the left of **Figure 2**, as can be seen at points the variance in the subject's response becomes quite large, which is likely due to some subjects physiologically responding less to this type of stress than others, "non-responders", ? As the cortisol of the two corpora is derived with a different assay, and given these statistical differences, the two corpora will be treated individually unless otherwise stated.

For the ULM-TSST and REG-TSST corpora, we additionally explore the continuous physiological signals available cf. **Figure 2** (middle). We utilise heart rate as Beats per Minute (BPM) from the REG-TSST and ULM-TSST corpora, and for the ULM-TSST corpus, we also utilise the respiration signal provided cf. **Figure 2** (right), which is based on chest displacement at a range of -10 to +10 mV (mV), where negative indicates an exhalation and positive an inhalation. Both of these physiological signals are known to alter during stress stimuli (Bernardi et al., 2000). Of note, from **Figure 2** (centre) we see that the BPM signal for REG-TSST contains values below 50 BPM and above 180 BPM suggesting some noise in the signal, likely due to the equipment type[2]

The ULM-TSST corpus also includes continuous emotion ratings, which were rated by three annotators for the dimensions of arousal and valence, at a 2 Hz sampling rate. Arousal and valence are derived from Russell's circumplex for affect (Russell, 1980), and allow for dimensional interpretation of the strength (arousal) and positivity (valence) of an emotion. For these signals, a "gold standard" is obtained by the fusion of annotator ratings, utilising the RAAW method, implemented using the MuSe-Toolbox (Stappen et al., 2021c). The mean Pearson correlation inter-rater agreement for these fused signals are 0.186 (±0.230) for arousal, and 0.204 (±0.200) for valence.

3.3 Data Processing

For the FAU-TSST and ULM-TSST corpora, the audio data was extracted from the video camera, placed approximately 3 m from the subject. For the REG-TSST corpus, two channels of audio were captured, and for the experiments, we utilise the first channel, which was recorded using the AKG PW45 presenter set with a close-talk microphone. All audio was converted to 16 kHz, 16 bit, mono, WAV format and applying peak normalisation to 1 dB for each audio file, i. e., adjusting the loudness based on the maximum amplitude of the signal, before extracting features. We re-ran the processing procedure for the FAU-TSST corpus that was first presented in Baird et al. (2019) to include portions of non-speech, and match ULM-TSST and REG-TSST. For the audio of all corpora, we applied *voice activity detection* (VAD), utilising the LSTM-RNN approach described by Hagerer et al. (2017). This method utilises spectral and MFCC-based features to generate frame-level VAD decisions with a granularity of 20 ms. The model was trained in a multitask setting

[2] REG-TSST: Polar RS800CX and V800 system, and ULM-TSST: BIOPAC Systems, MP35.

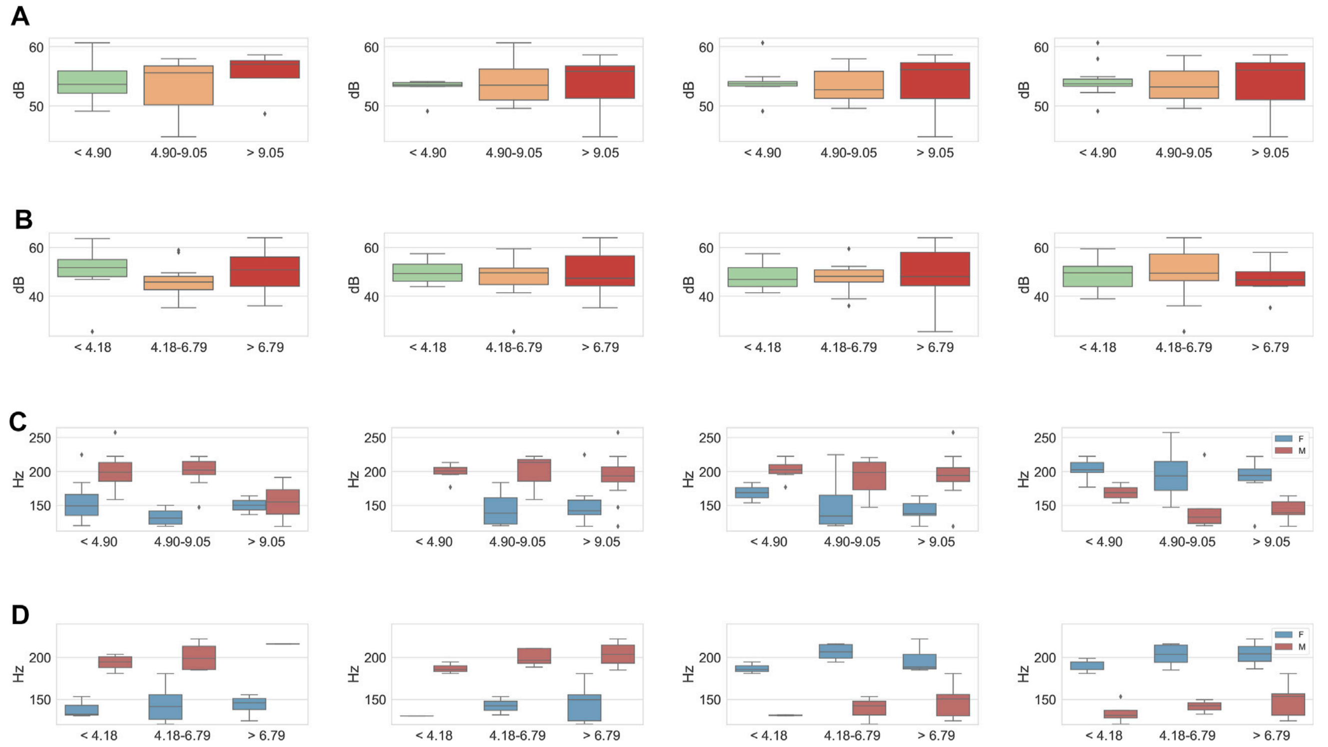

FIGURE 3 | Box plot representation of extracted intensity (dB) and pitch (F_0) for each speaker. Percentile grouping based on raw cortisol as nmol/L, showing measures taken at time-points S1, S3–S5, for FAU-TSST and REG-TSST corpora. **(A)** Intensity dB—FAU-TSST S1, S3—S5. **(B)** Intensity dB—Reg-TSST S1, S3, S3—S5. 3. **(C)** F_0 Hz—FAU-TSST S1, S3, S3—S5. **(D)** F_0 Hz—Reg-TSST S1, S3, S3—S5.

to jointly predict speech overlap, gender, and speech probability, achieving an overall performance of 93% F^1-score for speech detection. From this procedure in, cf. **Table 1**, it can be seen that the arithmetic task contains less speech, and in general, there appears to be substantial silence within the audio data, likely caused by the induced stress.

For all corpora, we create segments from the continuous signal. For FAU-TSST and REG-TSST, this is based on speech start (provided by the VAD), until the next utterance. To be comparable to the MuSe challenge, we do not alter the segmentation applied to ULM-TSST. As the text is also available for this corpus, the segmentation is based on aligned transcription (cf. Stappen et al. (2021a) for further detail). Each corpus is then partitioned in a speaker-independent manner into training, development, and test sets, cf. **Table 1**, where demographics including age and gender are balanced as best possible.

4 ACOUSTIC ANALYSIS

To further analyse the manifestation of stress in the human voice and explore each of the corpora utilised in our experiments more deeply, we extract the low-level acoustic features over the entire speech sample prior to segmentation for each speaker. We extract the fundamental frequency (F_0) and volume intensity (dB), as these are aspects of speech known to vary during stress (Protopapas and Lieberman, 1997; Giddens et al., 2013). For the F_0 extraction, we remove zero-values in other words, non-voiced parts to not skew the result based on segments of silence in the audio files; however, we consider the silence for intensity.

We first explore the acoustic behaviour in relation to the raw cortisol samples (nmol/L) from the FAU-TSST, and REG-TSST corpora, in groupings of 3-classes (lower 33rd, middle, and higher 66th percentile) at each sample time-point. It can be assumed that a higher feeling of stress leads to a higher cortisol response although with some delay Goodman et al. (2017). Given the variance in cortisol responses, as seen by the reasonably large standard deviation at each time-point **Figure 2**, these coarse groupings allow us to observe the behaviour of subjects with higher cortisol response against those with lower response to understand how if at all acoustic features relate to high states of stress. As the cortisol targets for each corpus were extracted with a different assay (cf. **Section 3**), we perform the grouping individually for each based on the percentile distribution. For FAU-TSST 33rd < 4.90 nmol/L, middle 4.90,—, 9.05 nmol/L, 66th > 9.05 nmol/L , and for REG-TSST 33rd < 4.18 nmol/L middle 4.18–6.79 nmol/L 66th > 6.79 nmol/L.

It is clear from plotting (cf. **Figure 3**) the classes that each corpus behaves similarly at each sequential time step. In general, speakers tend to have a more powerful intensity for the 66th

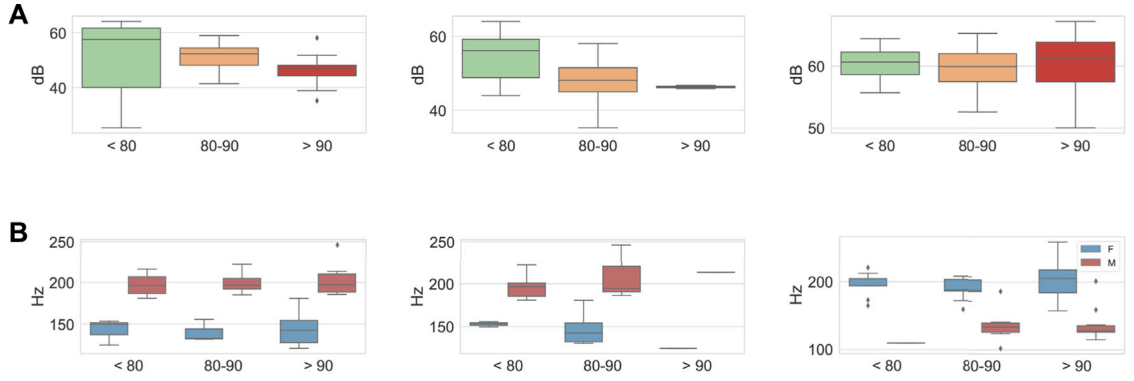

FIGURE 4 | Box plot representation of extracted intensity (dB), and pitch (F_0) of each speaker. Grouped based on μ Heart Rate as BPM. Including, REG-TSST interview task, REG-TSST arithmetic task, and ULM-TSST interview task. **(A)** Intensity dB—REG-TSST, Interview and Arithmetic, ULM-TSST Interview. **(B)** F_0 Hz—REG-TSST, Interview and Arithmetic, ULM-TSST Interview.

percentile cortisol groupings at S3-S4. In **Figure 3**, we see that at S1, those speakers in the lower 33rd percentile show a larger range in intensity, which reduces at S3 to S4. At S3 and S4, the mean intensity in dB also increases, particularly for those with a higher cortisol response; this intensity then decreases as the cortisol begins to lower at S5. In general, we can see from this analysis that those with higher levels of cortisol tend to have louder mean speech volume, and broader range in volume than other cortisol groupings, although this is in general less consistent for REG-TSST, potentially due to the differing microphone or the smaller population in the higher groupings of this corpus\enleadertwodots.

Interestingly, for F_0 (cf. **Figure 3**), we see similar behaviour concerning the cortisol groupings, particularly for the REG-TSST. In this case, the standard deviation of F_0 appears to increase with higher cortisol levels, and the same is true for FAU-TSST at S3-S4 although less prominent. As we split results by sex here (male and female) we see that the effect is not consistent for sex groupings, but it seems that at S3 both male and female groups do increase F_0 variance as cortisol response becomes higher, a finding which is consistent with related literature which states the F_0 mean increases as cortisol also increases Pisanski et al. (2016).

We also explore groupings of mean (μ) heart rate as beats per minute (BPM), Low μ < 80 BPM Middle μ 80,—, 90 BPM High μ > 90 BPM. These groupings were selected to balance the subjects in each group based on the distribution of the signal across both sets. This time, we plot the results for each of the TSST tasks, separately and all together. As with cortisol, we do see a relationship between the physiological BPM groupings and the acoustic features, cf. **Figure 4**. For the intensity of the REG-TSST corpus, there is a clear decline in volume as BPM increases for both tasks. This trend is not as clear for ULM-TSST, but the range does increase. When looking at F_0 in the same grouping for BPM cf. **Figure 4**, we see slightly more consistency, observing a slight increase in the range for F_0 as μ BPM increases. This finding is supported by other literature, which has shown that there is a relationship between heart rate and vocal quality (Kovalenko et al., 2019), showing that BPM can be considered an indication of stress as it pertains to cortisol. Furthermore, as with the cortisol groupings we also split by sex for F_0 analysis, and the mean F_0 for both sex does appear to increase with higher heart rate, although this is less consistent for the ULM-TSST corpus and also males as compared to females.

5 EXPERIMENTAL SETTINGS

We conduct four core experiments to explore further the benefits of speech features in the context of recognising markers of stress. As physiological markers are known to strongly affect the HPA axis, which is a factor that alters during a stressful situation, we recognise 1) sequential saliva-based samples of cortisol, utilising the FAU-TSST and REG-TSST corpora, where samples taken post-stress (S2+) with a strong correlation to the features would indicate an effective approach, 2) continuous emotion, as arousal and valence with ULM-TSST 3) continuous heartbeats per minute (BPM) utilising REG-TSST and ULM-TSST, and 4) continuous respiration, based on chest displacement, from ULM-TSST. Within these paradigms, we perform several cross-corpus (where possible) and transfer learning experiments (results discussed in **Section 6**) for each of these targets, exploring the efficacy of the machine learning approaches for entirely unlabelled data.

5.1 Features

We apply a feature-based machine learning approach, and we mainly focus on speech-driven audio features. However, we do include vision features to observe the potential benefit of fusion, and validate the advantage of speech features in this particular context.

Acoustic: From previous studies, we found that hand-crafted features appear to perform more robustly for the task of sequential cortisol prediction (Baird et al., 2019). However, as this was based on a single dataset, further validation was needed, and so for this study, we extract again both hand-crafted speech-based features, namely the *Computational Paralinguistics challengE* (COMPARE) feature set, and the *extended Geneva Minimalistic Acoustic Parameter Set* (EGEMAPS), as well as the deep learning spectrogram based approach utilising the DEEPSPECTRUM toolkit from the FAU-TSST and REG-TSST. From each audio instance, the COMPARE and EGEMAPS and DEEPSPECTRUM features are extracted at a rate of 1 s, using an overlapping window of 0.5 s. For the hand-crafted sets, we utilise the OPENSMILE toolkit to extract the 6 373 dimensional COMPARE feature set (Eyben et al., 2013), and 88 dimensional EGEMAPS feature set (Eyben et al., 2016). These features have shown to be effective for a number of similar wellbeing related tasks (Kim et al., 2019; ? ; Schuller et al., 2020), including detection of early stage dementia (Haider et al., 2019), and levels of anxiety (Baird et al., 2020). For the DEEPSPECTRUM features, we extract a 2,560 dimensional feature set of deep data-representations using the DEEPSPECTRUM toolkit (Amiriparian et al., 2017). DEEPSPECTRUM has shown success for various audio- and speech-based tasks (Mertes et al., 2020), and extracts features from the audio data using pre-trained convolutional neural networks. For this study, we extract features based on the viridis colour map, and the deep features are extracted from the layer *fc7* of AlexNet (Krizhevsky et al., 2012). We also explore the use of VGGISH functions Hershey et al. (2017) which are pre-trained on AudioSet (Gemmeke et al., 2017). From this, we extract a 128-dimensional VGGISH embedding vector from the underlying log spectrograms.

Visual: For the video-based features, we utilise the well-established VGGFACE set, and extract this from FAU-TSST and ULM-TSST excluding REG-TSST as no video data was available. The first step in this pipeline is to extract the faces as images, and to do this at the same frame-rate as the audio features (2 Hz), utilising the MTCNN (Zhang et al., 2016) which is pre-trained on the data sets WIDER FACE (Yang et al., 2015) and CelebA (Liu et al., 2015). We use the VGGFACE (version 1) (Parkhi et al., 2015), which is based on the pre-trained deep CNN VGGISH 16, which was introduced by the visual geometry group of Oxford (Simonyan and Zisserman, 2014). Detaching the top-layer of a pre-trained network results in a 512 feature vector output referred to as VGGFACE.

5.2 Regressors

For all the recognition tasks, we are performing regression experiments. To do this, we first validate the data itself by performing a series of arguably more robust Support Vector Regression (SVR) experiments for the cortisol targets only. This is then followed by a series of deep learning models based on an LSTM-RNN architecture to explore a more state-of-the-art approach, which may better observe the time-dependent nature of the observed signals.

SVR: For the initial experiments we use the epsilon-support vector regression (SVR) and a linear kernel implementation from the Scikit-Learn toolkit (Pedregosa et al., 2011). For training, the data is split into speaker-independent sets: During the development phase, we trained a series of SVR models, optimising the complexity parameters ($C \in 10^{-4}$–1), evaluating their performance on the development set. We re-trained the model with the concatenated train and development set and evaluated the test set performance.

LSTM-RNN: We utilise a similar LSTM-RNN based architecture to the one which was applied for the baseline of the MuSe 2021 Challenge[3] and similar tasks (Stappen et al., 2021b,c). In the training processes, the features and labels of every input are further segmented via a windowing approach (Sun et al., 2020), which may offer the network more context. We experimented with various window lengths, but as in the MuSe Challenge, a window size of 300 steps (150 s) was found to be optimal for all corpora. We tested n = (1, 2, 4)-layered uni and bidirectional networks with h = (50, 100, 200) hidden states and a learning rate of lr = (0.00005, 0.0001, 0.005, 0.001). Initial experiments showed that the best results were obtained with a 4-layered network, consisting of two LSTM and two fully-connected (FC) layers, with a hidden size of 50, and a learning rate of 0.00005 (cf. **Figure 5** for an overview). To reduce the computational overhead, we utilised these values in all experiments reported here.

Model evaluation: For some targets examined here, we have continuous frame-level labels available. This allows us to use the same formulation as in the MuSe Challenge, where we obtain frame-level predictions using an LSTM-RNN architecture and subsequently compare those to the frame-level target. This is not true for the cortisol task, as only one single target value is available per session. Moreover, each session lasts approximately 10 min, and stress may only manifest on short, intermittent segments throughout those recordings. To overcome these challenges, we opted to replicate the session-level labels on the frame and model them accordingly. During training, we use standard many-to-many training (Mousa and Schuller, 2016), where the networks (SVR and LSTM) are trained to predict the target on all frames. This formulation results in frame-level predictions during evaluation as well. However, as mentioned, we only have a single session-level target. Thus, to evaluate the performance of our models, we first aggregate (i. e., average) their predictions for each session before comparing them to the reference cortisol values.

As primary evaluation metrics for all models, we report either *Spearman's correlation coefficient* (ρ), *Root-Mean Square Error* (RMSE) or normalised RMSE (NRMSE). Reporting correlation as ρ is used for the sequential cortisol target, as we are interested in exploring trends in the data and how well the models can learn targets that are derived from a more ordinal value. When discussing specific results for ρ the *p-value* is also reported, to discuss the additionally significance of the correlation. In this case, as with any other p-values reported, significance can be consider at values of $p < 0.05$. RMSE, in contrast, is better suited to a more objective evaluation, which fits the case of time-continuous signals such as heart rate, and given the less intuitive range of the respiration signal, we report NRMSE in this case.

[3] https://github.com/lstappen/MuSe2021

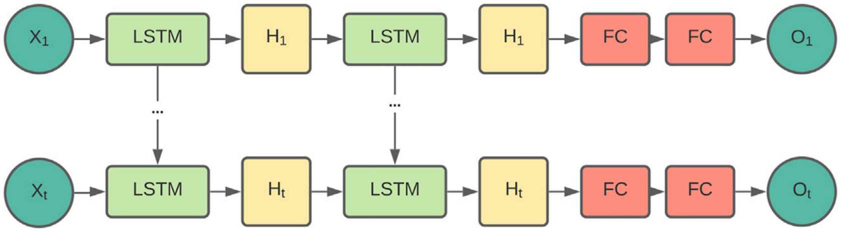

FIGURE 5 | LSTM-RNN model architecture. The input sequence $\{X_i, i \in [1, T]\}$ is first fed to two LSTM layers of hidden size 50. The intermediate representations $\{H_i', i \in [1, T]\}$ produced by the second LSTM layer are then processed by two FC layers to produce the output sequence $\{O_i, i \in [1, T]\}$.

6 RESULTS AND DISCUSSION

We provide a series of tables and plots to report various aspects of the results obtained by our experiments. For clarity of presentation, we will discuss the results obtained for each of the targets separately.

6.1 Sequential Cortisol Prediction

Our main source of truth for the degree of stress during the TSST setting is the saliva-based cortisol measurements obtained at differing time points. This information is only available for the FAU-TSST and REG-TSST datasets; therefore, we focus primarily on those two in this section. As discussed in **Section 3**, the only modality standard across those two datasets is audio, while for FAU-TSST, we additionally have video. For this reason, we primarily focus on the audio modality, which in Baird et al. (2019) was shown to be a strong predictor of cortisol-based stress.

Furthermore, as noted in **Section 3**, cortisol values were derived using different assays, thus making the two scales incompatible. This makes it incorrect to evaluate any trained models with a standard cross-corpus paradigm, whereby models are trained on one dataset and evaluated on another. Instead, the core focus of our experiments is to explore how well the methodology can be replicated on different datasets. Nevertheless, we additionally explore the direction of pooling the data from the two studies and learning a joint model. Pooling more data, which come from fundamentally different domains, e. g., acoustically and the cortisol assay used, might still benefit the training of neural networks, which typically require a lot of data to learn from. We thus train models in both single- and multi-domain settings, and always evaluate them on in-domain data separately for each dataset.

As discussed in **Section 3**, the subjects performed two tasks during the TSST; a speech interview and an arithmetic task. We hypothesise that subjects behaved differently during each task, and that stress manifested differently in the respective acoustic features. This hypothesis was validated by in the initial experiments of Baird et al. (2019), where models built on each task separately perform better than models built with both tasks. Thus, for these experiments we additionally differentiate between the interview and the arithmetic tasks, building separate models for each of them, and contrasting their performance to models built after pooling both tasks.

We first run a series of experiments with a traditional SVR algorithm and only acoustic features to explore if the REG-TSST dataset performs similarly to FAU-TSST, and if the study from Baird et al. (2019) can be replicated for FAU-TSST with a slightly adapted methodology (e. g., altered speech segmentation) for data processing. In **Figure 6**, we see that the FAU-TSST corpus behaves as expected, with correlation strongest after S4 (interview: S3, FAU-TSST EGEMAPS 0.200, $p < 0.05$; S4, FAU-TSST EGEMAPS 0.340 $p < 0.05$), slightly weaker for the arithmetic task compared to the interview, which could be caused by the reduced speech in the arithmetic task. For the REG-TSST corpus, the trend is less obvious for all feature sets, particularly for the interview task with COMPARE features where we see a strong decline from S1. The EGEMAPS features appear to perform consistently for both tasks of the REG-TSST, however, in this case the arithmetic task appears to have stronger correlations than the interview, peaking earlier at S3 than FAU-TSST for this task, which may indicate the above-mentioned difference in intra-individual stress response during the speech tasks of the two corpora. In general, from these experiments, we not only initially affirm the findings of Baird et al. (2019) that higher correlation is obtained post S2 (in general either S3 or S4) by validating this on an additional corpus, but we also affirm that hand-crafted features are more suited for this task. However, for the novel REG-TSST data, the smaller EGEMAPS set is performing more robustly, and more consistently overall. Given this, we will continue to use EGEMAPS as the main acoustic feature set for further experiments.

Results for the LSTM model are shown in **Table 2**. Again we see that, in line with Baird et al. (2019), speech-based models can predict cortisol levels samples taken at time points S2-S5 with a medium to strong correlation and a mean peak around S4 (+20 min after the TSST). This is consistent across both datasets and tasks. However, there are important and interesting differences across different settings.

In general, we observe that with the LSTM network, we can better predict cortisol from the arithmetic task of FAU-TSST, which slightly contradicts our SVR results and shows that this task can also yield good results if we consider the sequential nature of different frames. This indicates that, for this dataset, subjects either became more stressed during this part of the TSST or that the manifestation of stress in the speech was more pronounced. Based on our manual inspection of the dataset, the second hypothesis seems more plausible, as subjects who struggled

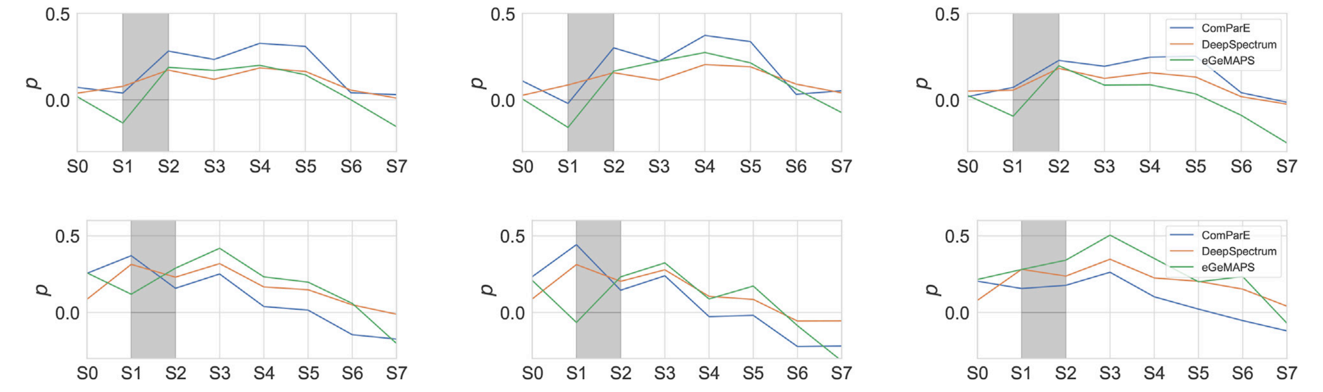

FIGURE 6 | SVR results for the FAU-TSST **(above)**, and REG-TSST **(below)**. Reporting Spearman's correlation coefficient (ρ) for all scenarios **(left)**, interview task **(middle)**, and arithmetic task **(right)**, for each sequential saliva (S) sample. SVR experiments were conducted with three different acoustic feature sets: EGEMAPS, DEEPSPECTRUM, and COMPARE.

TABLE 2 | Spearman's correlation coefficient (ρ) for session-based cortisol at each saliva (S)ample, from S0 45 min to S7 +60 mins. Utilising EGEMAPS features for FAU-TSST and REG-TSSTcorpora, for the (Inter)view and (Arith)metic tasks, as well as the mean (μ.) across all. Where emphasised results indicate a positive correlation above 0.2.

ρ	—	FAU-TSST							
Train	**Task**	**S0**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
FAU	Inter	0.104	0.016	**0.203**	0.000	**0.286**	-0.209	-0.352	-0.324
FAU	Arith	**0.302**	0.060	**0.236**	**0.385**	**0.396**	-0.165	-0.242	-0.225
FAU	Inter. and Arith	0.077	0.093	0.022	0.099	-0.176	-0.286	-0.555	-0.407
FAU and REG	Inter	0.154	0.055	-0.159	0.159	0.044	-0.341	0.016	-0.456
FAU and REG	Arith	**0.335**	**0.214**	**0.368**	**0.374**	**0.698**	**0.286**	-0.027	-0.214
FAU and REG	Inter. and Arith	0.126	**0.209**	-0.077	0.104	0.088	-0.220	-0.632	-0.456
μ	—	0.183	0.108	0.099	0.187	**0.223**	-0.156	-0.299	-0.347
	—	REG-TSST							
REG	Inter	**0.297**	**0.827**	**0.527**	**0.261**	**0.236**	-0.127	-0.527	-0.079
REG	Arith	0.091	**0.559**	-0.164	0.091	**0.455**	**0.333**	**0.552**	**0.406**
REG	Inter. and Arith	0.127	**0.474**	0.055	**0.285**	**0.248**	0.115	-0.273	-0.406
FAU and REG	Inter	-0.152	**0.559**	**0.467**	**0.200**	**0.261**	-0.018	-0.539	0.164
FAU and REG	Arith	-0.212	**0.267**	0.055	-0.042	**0.370**	**0.212**	0.188	0.091
FAU and REG	Inter. and Arith	0.006	**0.584**	**0.721**	**0.770**	**0.442**	0.176	-0.139	-0.042
μ	—	0.026	**0.545**	**0.279**	**0.261**	**0.335**	0.115	-0.123	0.022

during the interview tend to stay completely silent, whereas they would continuously produce utterances (although at short bursts) for the arithmetic task. Moreover, pooling data from both tasks resulted in worse performance when training on individual datasets, pointing towards a different expression of stress in each of them.

Overall, for both tasks, we observe higher correlations for times S3-S4, with the interview task tending to peak a bit earlier than the arithmetic one. Given the relative delay between the two tasks, this is in line with our previous research (Baird et al., 2019) showing that speech signals are more correlated with cortisol measurements taken approximately 10 min after initial stress. Interestingly, we also observe a high correlation for cortisol measures taken at S1 (1 min *before* the TSST) for REG-TSST (particularly for the interview task). When observing the mean score for REG-TSST we see that shows to be the highest peak. On the one hand, it could be considered that this is attributed to increased apprehension by the subjects, leading to more stressed behaviour during the early stage of the TSST; however, as we observed earlier there is lower variability across subjects for measurements at S1 (cf. **section 3**) which may have made this task easier to learn.

Finally, we observe that multi-domain models built by pooling both datasets perform consistently better, while additionally benefiting from the pooling of the interview and arithmetic tasks in the case of REG-TSST. This illustrates that, even though the cortisol measurements in the two datasets are based on fundamentally different scales, the relationship between relative cortisol values and acoustic features remains consistent, allowing the models to benefit from bigger and more diverse data and obtain better performance, as measured by Spearman's correlation.

Even though our quantitative evaluation is performed on the session level, it is interesting to investigate how stress manifests through the audio modality at different time points using our approach. **Figure 7** shows frame-wise predictions vs a selection of sequential cortisol values for two subjects, one from each corpus. For subject F-1 from FAU-TSST (top), we see a higher deviation from the cortisol ground truth, which settles more during

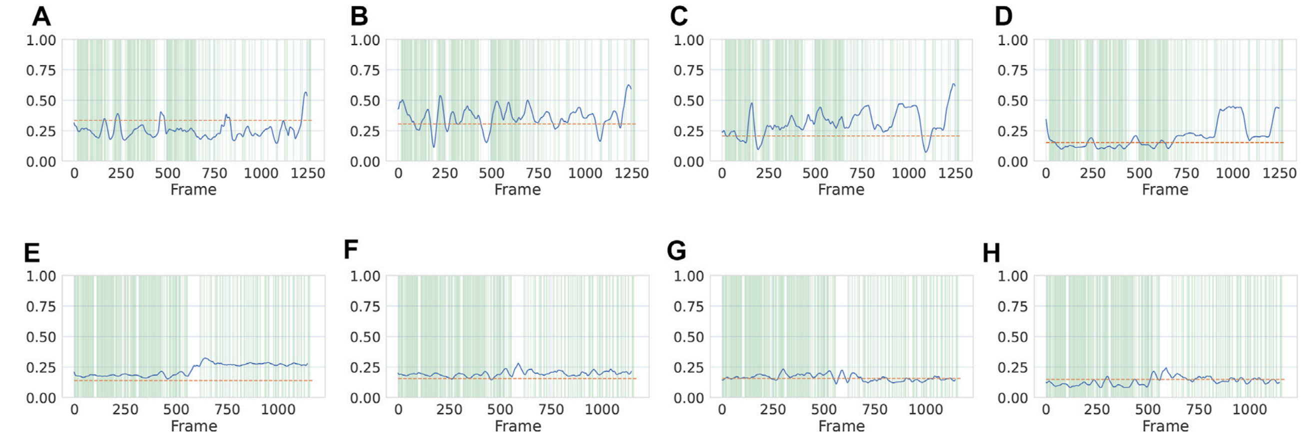

FIGURE 7 | Normalised frame-wise cortisol predictions (continuous) vs ground truth (dashed) sampled at times S2-S5 using EGEMAPS features for subjects F-1 from FAU-TSST and subject R-1 from REG-TSST. The shaded (green) area represents segments detected as having speech by the VAD. Predictions are smoothed with a moving average filter with a window size of 30 for visibility. **(A)** F-1 S2, **(B)** F-1 S3, **(C)** F-1 S4, **(D)** F-1 S5, **(E)** R-1 S2, **(F)** R-1 S3, **(G)** R-1 S4, **(H)** R-1 S5.

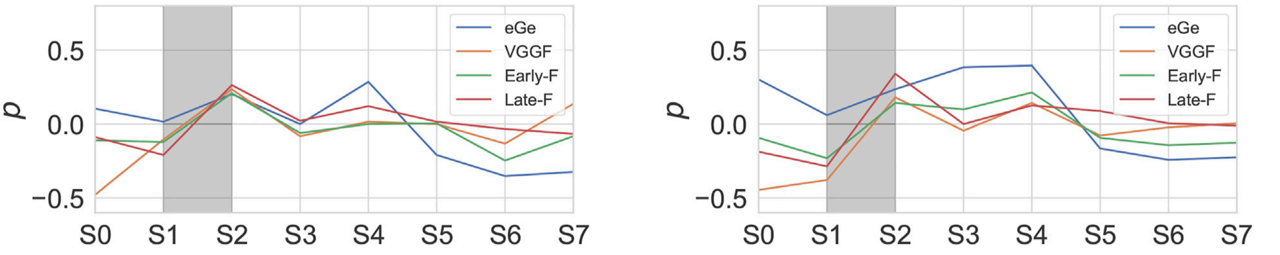

FIGURE 8 | Spearman's correlation coefficient for session-level cortisol prediction using VGGFACE features on the FAU-TSST dataset. We report unimodal visual-based results as well as multimodal ones utilising early and late fusion with EGEMAPS features (A + V). We also replicate the EGEMAPS results from **Table 2** for easier comparison.

segments of speech, as S3-S4 may be considered the true cortisol release at that time point. It is interesting that for subject R-1 for the REG-TSST in **Figure 7** (below) the prediction is more consistent. For the S2 time point, recognition is more accurate earlier in the speech session, i. e., the interview task. Counter to this, at S5 for the FAU-TSST plot; we see that the system struggles to recognise after the interview task, which would indicate that this sample of cortisol is less an indication of the stress response, affirming that the speech signal is a strong predictor for peaks in cortisol which occur due to stress. These individual differences across subjects suggest that speaker-adapted models, which have been shown to improve results for other affective computing tasks (Triantafyllopoulos et al., 2021), could improve the predictive accuracy of stress prediction models as well. We also see, when observing **Figure 7**, that the standard deviation is in general smaller for the REG-TSST corpus, possibly indicating the benefit of the close-talk recording method.

In addition, to compare the performance of audio, we investigate the effectiveness of video-based models for stress recognition on the FAU-TSST dataset, on which the video modality is available. Using an identical experimental protocol, and simply substituting EGEMAPS with VGGFACE features. Results are shown in **Figure 8**, and as can be seen, the vision features are much lower than those obtained with EGEMAPS features. This indicates that in the FAU-TSST dataset, the auditory modality is more appropriate for modelling stress, although there is still a similar behaviour where we see a peak in correlation after the point of stress (S2-S4). Moreover, we experiment with early and late multimodal fusion, where we either fuse (concatenate) the features and subsequently train a new model or fuse (average) the predictions of the existing unimodal models. As our acoustic experiments showed that task-specific models perform better, we did not fuse data from both TSST tasks for these experiments. We observe that multimodal fusion can lead to better performance in some cases, most notably for the prediction of cortisol at S2, suggesting that the interview task was more meaningful for these features, however, generally EGEMAPS features remain strong as a uni-modal approach.

Finally, we use the models built on FAU-TSST and REG-TSST to predict the likely cortisol levels on the ULM-TSST corpus, for which this information is not available. Although we do not have a ground truth here, we aim to see if the performance is similar

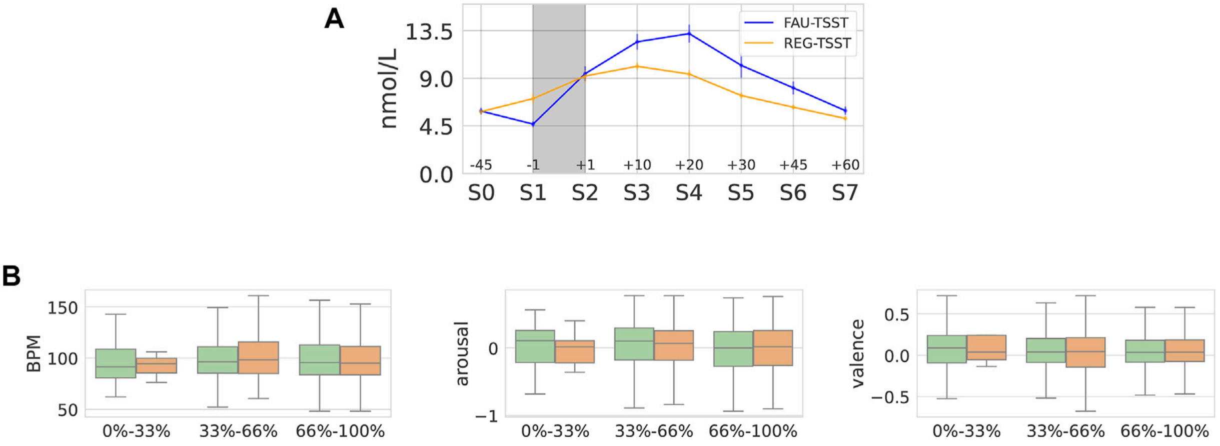

FIGURE 9 | The mean (μ) predicted cortisol value for each time step on the ULM-TSSTdataset using models built on FAU-TSSTand REG-TSST (above). Below, is the BPM, arousal and valence predictions grouped based on raw cortisol (nmol/L), obtained for time point S4 using models built on FAU-TSST **(left-green)** and REG-TSST **(right-orange)**. Low, mid, and high groupings defined separately for each dataset based on percentiles. **(A)** ULM-TSST μ cortsiol in nmol/L. **(B)** Cortisol groups for BPM, arousal, and valence predictions of FAU-TSST **(left-green)** and REG-TSST.

concerning peaking cortisol levels after the S2 time point. To perform these experiments, we use the models built separately on FAU-TSST and REG-TSST. These models were built on different scales, stemming from the fact that different assays were used to extract cortisol levels in the two datasets. We furthermore used the models built on data from the interview task alone, as this is the only task available for ULM-TSST. **Figure 9** shows the mean predicted cortisol levels from an entire ULM-TSST session; similar to FAU-TSST and REG-TSST, we observe a peak in (predicted) cortisol levels at times S3 and S4. The FAU-TSST model is returning higher cortisol values; this is consistent with the dataset overview presented in **Section 3** which shows that cortisol levels are higher for FAU-TSST.

In addition, box plots of grouped cortisol levels, and the other biomarkers available to the ULM-TSST corpus show that higher (predicted) cortisol levels correspond to slightly higher BPM and arousal, and slightly lower (negative) valence. Moreover, we observe some noticeable differences between the predictions obtained by the two models. For example, the model built on REG-TSSTdata shows its lowest cortisol predictions for very narrow beats per minute (BPM), arousal and valence ranges, which is less narrow for FAU-TSST at those targets, and for valence the lower percentile shows a broader range for valence than all other groupings. These differences further demonstrate that models trained on different corpora, with differences in the acoustic conditions and the way cortisol levels were measured, can result in models that behave in different ways on out-of-domain data. However, in general, behaviours appears to be consistent.

In summary, our results demonstrate that it is possible to predict cortisol levels taken 10–20 min (common time frame for the post-stress cortisol peak (Goodman et al., 2017)) after a stressful event using speech as well as video features, with the former performing better in this context. Stratifying the data concerning the task that the subjects were performing additionally reveals an interesting trend; we see a general trend that we are able to better predict cortisol levels from the arithmetic task of, FAU-TSST but from the interview task of REG-TSST. This may point to underlying differences in the way subjects experienced and expressed stress in the two data collection procedures; there is overall much fewer speech data in the REG-TSST arithmetic task, which may be another reason for this.

As mentioned, cortisol levels constitute our primary source of truth for an individual's stress level. However, these measurements are not easily collected and readily available, e.g., for the ULM-TSST corpus they are missing, and learning from a single value from each session, is a challenge for any machine learning architecture. With this in mind, in the following sections, we further investigate continuous physiological markers of stress which are more readily available and offer a more fine-grained view of stress responses, particularly if combined with a cortisol ground truth.

6.2 Emotional Dimensions

We begin our discussion of alternative markers for stress with the emotional dimensions of arousal and valence (Russell, 1980). These dimensions are known to be related to stress (Johnson and Anderson, 1990). The ULM-TSST dataset is the only one of the three datasets examined here, which contains annotations for arousal and valence. These dimensions form the targets for the 2021 MUSE-STRESS sub-challenge (Stappen et al., 2021a). As there are no available annotations for FAU-TSST and REG-TSST, we proceed to predict emotional values on the interview task for both, using models built on the ULM-TSST dataset. As audio is our core focus and is the only modality commonly shared across all three datasets, we use the EGEMAPS -based models developed and released as part of the challenge baseline[4]. Both emotion models show strong performance on the ULM-TSST test set, with

[4] https://github.com/lstappen/MuSe2021

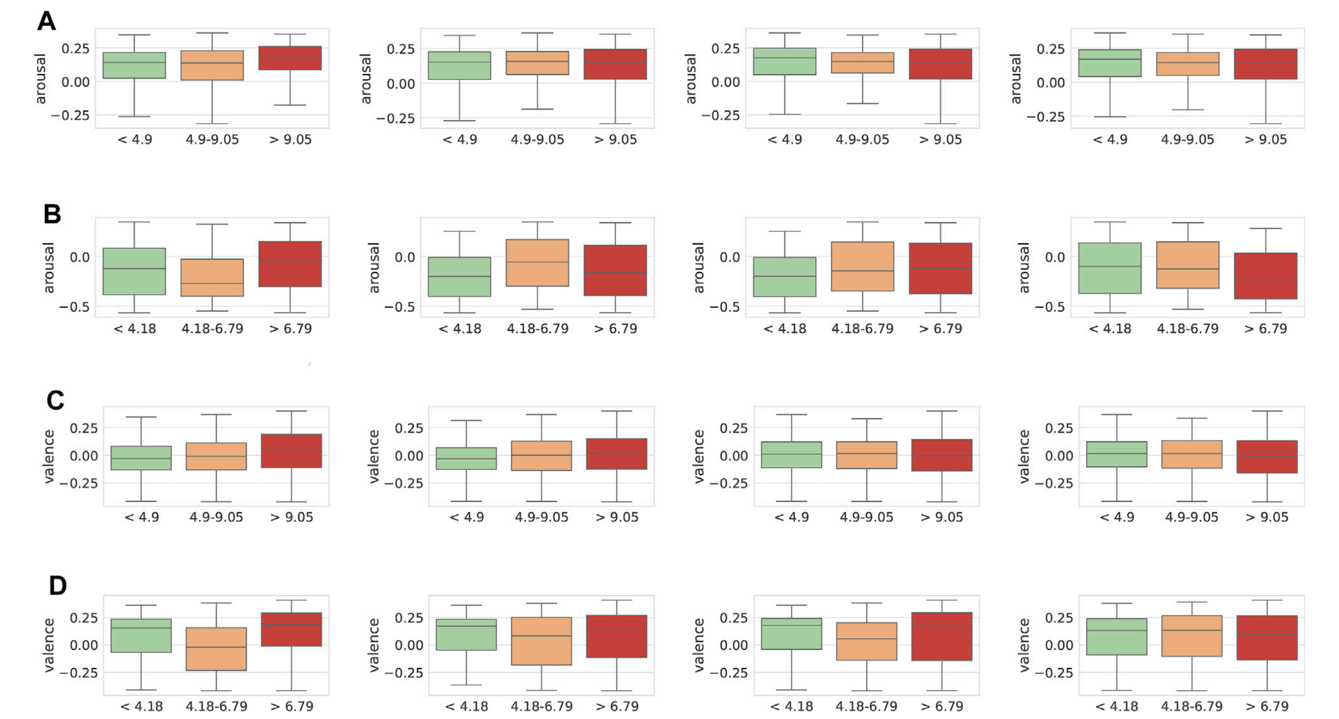

FIGURE 10 | Box plots of predicted arousal valence values for the FAU-TSST and REG-TSST. Grouped based on raw cortisol (nmol/L) measures taken at time points S2-S5 **(left to right)**. **(A)** Arousal with cortisol groups at S2–S5 for FAU-TSST. **(B)** Arousal with cortisol groups at S2–S5 for REG-TSST. **(C)** Valence with cortisol groups at S2–S5 for FAU-TSST. **(D)** Valence with cortisol groups at S2–S5 for REG-TSST.

the arousal model achieving a CCC of 0.4415, and the valence model one of 0.5019. Moreover, as the ULM-TSST corpus only contains the interview task, we only predict those dimensions at the respective functions for FAU-TSST and REG-TSST.

In **Figure 10**, we show distribution plots of the arousal and valence predictions for FAU-TSST and REG-TSST vs the cortisol measures taken at different time-points. The cortisol values have been grouped in the same way as **Section 4**, i. e., to their low, medium, and high based on the 33 and 66% percentiles derived from the raw cortisol values for the different datasets. As previously, we do observe different trends across the two datasets. FAU-TSST is generally showing positive values for arousal, whereas REG-TSST is showing negative ones. Although these results are based on model predictions and are thus not as reliable as human annotations, they nevertheless shed light on potential differences across the two datasets. Interestingly, subjects in REG-TSST appear generally less aroused compared to those in FAU-TSST, which once again points to underlying differences in how subjects reacted during the TSST in the two settings. For the high percentile grouping at S2 (+1 min after the TSST), we generally observe higher arousal values for both datasets, whereas we observe that lower arousal values are predicted for subjects in the lower cortisol percentile for FAU-TSST, as measured at S3 (+10 min after the TSST).

We additionally used a two-sample independent *t*-test to test the differences in predicted arousal and valence values for all groups and datasets. Of note, we did not conduct a normality test, the *t*-test is know to be robust to deviations from normality larger sample sizes Sawilowsky and Blair (1992). All differences were found to be statistically significant at the $p < 0.05$, except arousal in lower vs middle percentile-cortisol percentiles measured at times S0 and S2 for FAU-TSST and S5 for REG-TSST, mid vs high percentiles measured at S4 for REG-TSST, and low vs high percentiles measured at S3 for FAU-TSST. For valence, the only non-significant results were those between the lower vs middle percentiles measured at S5 for both FAU-TSSTand REG-TSST, and the middle vs high percentiles measured at S4 and S6 for FAU-TSSTand REG-TSST, respectively. This shows that, even though we lack ground truth values for FAU-TSST and REG-TSST, we could use a model trained on a related but different dataset to predict them and obtain strong predictors of stress.

6.3 Continuous Heart Rate

Stress is known to impact heart rate (HR) (Berntson and Cacioppo, 2004; Taelman et al., 2009) through its activation of the sympathetic (Goldstein, 1987) and suppression of the parasympathetic branch of the autonomic nervous system (Akselrod et al., 1981). HR can therefore serve as a vital indicator of stress in modern affective computing applications. As discussed in **Section 3**, however, only one of the three datasets examined here, the REG-TSST dataset, has both HR and cortisol measurements, whereas the FAU-TSST dataset has only cortisol measures and ULM-TSST only HR ones. Thus, the only dataset

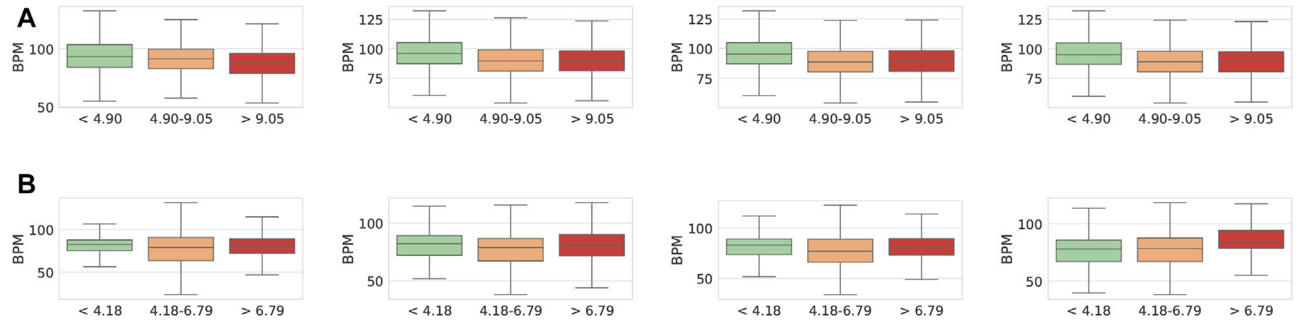

FIGURE 11 | Box plots of BPM value. Showing the ground truth REG-TSST and predicted for FAU-TSST. Grouped based on raw cortisol (nmol/L) measures taken at time-points S2-S5. **(A)** Predicted BPM with cortisol groups at S2–S5 for FAU-TSST. **(B)** Ground truth, BPM with cortisol groups at S2–S5 for REG-TSST.

where we can precisely evaluate the relationship of HR with stress is REG-TSST.

Figure 11 shows the distribution of ground truth HR values for the REG-TSST dataset vs low, mid, and high cortisol levels taken at different time points. Two-sample independent t-tests show that all results are significant at the $p < 0.05$ level, except the low vs high percentiles at time S0 and the low vs middle percentiles at time S5. Overall, we observe a rising trend for BPM values as the cortisol levels increase; this is consistent with our expectations and prior work (Berntson and Cacioppo, 2004; Taelman et al., 2009). This trend is particularly pronounced for S5 (+20 min after the TSST) showing that higher cortisol values obtained during that time were highly correlated with high BPMs during the TSST. In general, this trend differs to what was observed for the acoustic signals (cf. **Section 4**), indicating that different modalities may be better at predicting cortisol levels measured at different times.

As the other dataset used in this study with cortisol measurements, FAU-TSST, does not have available HR measures, we attempt to predict BPMs using models built on the other two datasets. Specifically, we use the speech modality of the REG-TSST and ULM-TSST datasets to build a model, which we then use to predict BPMs on the FAU-TSST dataset. This is motivated by audio being the only common modality across the three corpora, and also that the effect of HR on the voice has long been established by previous research (Orlikoff and Baken, 1989). Several prior works have attempted to model HR from voice signals, either as a classification (Schuller et al., 2013) or a regression task (Smith et al., 2017; Jati et al., 2018). Jati et al. (2018) use EGEMAPS to predict BPM from speech on the segment level, and achieve an root mean squared error (RMSE) of 12 BPM.

Inspired by these past findings, we attempt to predict HR in the form of BPMs using speech signals. In line with our previous results for cortisol, we use an long short-term memory (LSTM) architecture on EGEMAPS features. As all three datasets were recorded in different locations with potentially different acoustic conditions, we are faced with the well-understood domain mismatch problem (Ben-David et al., 2010), where models trained on data from one domain might not generalise well to different domains. Moreover, as discussed in **Section 3**, the two datasets cover non-overlapping ranges of the BPM range, with subjects in REG-TSST having a generally lower BPM than subjects in ULM-TSST, and are also recorded in different conditions, with ULM-TSSTconsisting of far-field and REG-TSSTof near-field recordings. To address this issue, we first train two single-domain models using both available datasets in isolation and then train a multi-domain model using data from both datasets. In all cases, we evaluate and report model performance separately for each dataset.

TABLE 3 | RMSE as BPM for single- and multi-domain results for BPM prediction on the REG-TSST and ULM-TSST corpora using EGEMAPS and the LSTM-based architecture.

ρ	REG-TSST		ULM-TSST	
Train	**Dev**	**Test**	**Dev**	**Test**
REG-TSST	**39.90**	**38.57**	20.98	22.96
ULM-TSST	36.53	40.80	**19.32**	**22.70**
REG-TSST and ULM-TSST	36.23	38.96	23.07	23.05

Emphasised results indicate strongest performance on given evaluation set.

RMSE results are shown in **Table 3**. Our initial observation is that all models perform better on the ULM-TSST dataset, and that in-domain models perform better than their cross-domain counterparts. Moreover, the multi-domain model does not bring any improvements compared to the single-domain ones. The limited overlap partially explains this in the BPM range for the two datasets; combining the data does not lead to considerable benefits since the target is different. The best performing combination is obtained when training and testing on the ULM-TSST dataset, and achieves an RMSE of 19 BPMs. This is lower than the results reported by previous works (Jati et al., 2018), which were, however, performed on different data and are thus not directly comparable to ours. Moreover, as discussed above, potential movements of the subjects lead to more unreliable measurements, which make the target more of a challenge to learn.

In general, predicting HR from free speech signals is a challenging task and is especially hampered by the lack of information whenever subjects remained silent. This is illustrated in **Figure 12**, where we present frame-wise BPM predictions vs ground truth signals for three subjects coming from the REG-TSST and ULM-TSST datasets. As seen in particular for

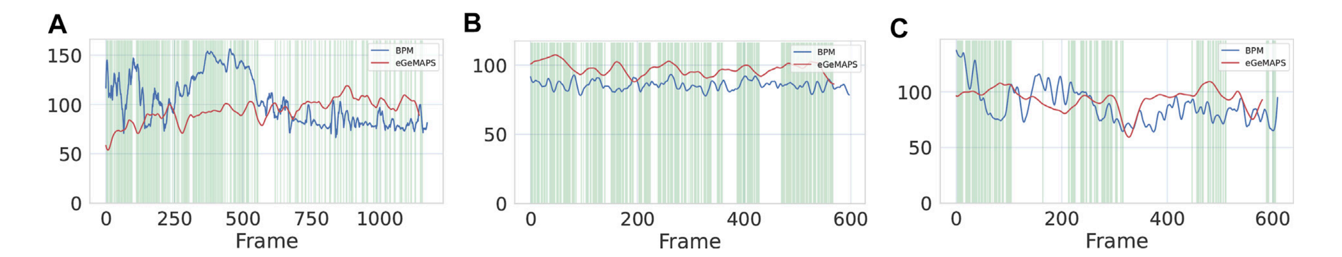

FIGURE 12 | Frame-wise BPM predictions vs ground truth for subject R-1 from the REG-TSST dataset and subjects U-39 and U-64 from the ULM-TSST dataset. Shaded (green) area represents segments detected as having speech by the VAD. Predictions have been smoothed with a moving average filter with a window size of 30 for visibility. **(A)** R-1 BPM. **(B)** U-39 BPM. **(C)** U-64 BPM.

subject U-64 (right), there may be periods of prolonged silence, where the audio modality is unavoidably a bad predictor of HR. Interestingly, even though we found silence periods occurring whenever subjects struggled with finding something to say during the interview task, we do not necessarily see an accompanying rise in BPMs, as seen for subject U-64.

Despite the relatively low performance obtained by our speech-to-BPM models, we still use them to obtain BPM predictions on the FAU-TSST dataset, as we are primarily interested in the usefulness of predicted BPM values for stress modelling. In **Figure 11**, we show the distribution of predicted BPM values for cortisol measurements obtained at different time points. Surprisingly, we observe a downward trend for BPMs as the stress level increases. This counterintuitive finding can be explained as follows: when subjects move a lot, the BPM monitoring devices may lead to erroneous measurements. Therefore, rather than these low measurements implying that stress leads to a lower BPM, we interpret them as a demonstration that BPM signals, though theoretically well justified as predictors of stress, are nevertheless a challenge to collect in practice. Thus, BPM alone may be inferior to signals like voice that are easier to manage and provide richer information for evaluation. However, the trend is not what we expect. We still see a separation between different cortisol levels, indicating that predicting HR from speech signals can be a useful proxy for stress prediction. Two-sample independent t-tests show that all differences are significant at the $p < 0.05$ level except the middle vs high percentiles as measured at S4.

6.4 Respiration

The final biological signal we examine here is respiration derived from chest displacement with a range of (−10:+10), which, similarly to the emotional dimensions, is only available for ULM-TSST. Based on previous research (Suess et al., 1980), we expect this signal to have a solid connection to stress. Although this physiological signal has strong potential for several affective applications (Wu et al., 2010; Ishii et al., 2016; Zhang et al., 2017), to the best of our knowledge, there has been little work on predicting it from other modalities. As we have both audio and video signals available for ULM-TSST, we attempt to use both to model respiration. However, similar to the other biomarkers, we only use the audio modality when predicting this signal for the other two datasets, as this is the only modality shared among all. Given that process of breathing, and vocalising shares related anatomy, we naturally expect the audio modality to be a strong predictor of respiration. Similarly to the emotional dimensions, cf. **Section 6.2**, we only predict respiration on the interview task of FAU-TSST and REG-TSST, as this was the only task available for ULM-TSST.

TABLE 4 | normalised root mean squared error (NRMSE) results for unimodal and multimodal Audio + Video (A + V) respiration prediction range [−10:10] on the development test sets of the ULM-TSST corpus utilising an LSTMs. For the multimodal results, we perform both early and late fusion.

NRMSE	Dev	Test
EGEMAPS	0.118	0.122
VGGFACE	0.146	0.139
A + V (Early)	0.142	0.143
A + V (Late)	**0.120**	**0.120**

Emphasised results indicate strongest performance on given evaluation set.

In **Table 4**, we show multimodal results for the recognition of respiration rate in the ULM-TSST corpus. As the signal is measured in arbitrary units, we report NRMSE, which is equivalent to the standard RMSE normalised by the target range. Late fusion of the two modalities brings the strongest results. However, unimodal EGEMAPS features appear to be only slightly lower than the best multimodal result, indicating that they can be used to predict respiration in isolation. This is not too surprising, as the speech and respiration are likely highly correlated, and artefacts from breath will inherently remain with the audio features. This may be to a lesser degree for the VGGFACE features, mainly due to possible occlusions, which may not observe mouth movement related to deeper breath.

From the box plots of **Figure 13**, we can observe that the respiration signals predicted appear to behave in an expected way for such cortisol groupings, for both the FAU-TSST and REG-TSST corpora. For example, we observe a rise in respiration levels as cortisol increases for REG-TSST; this trend also manifests for FAU-TSST but less pronounced. Two-sample independent t-tests show that all differences are significant at the $p < 0.05$ level, except the mid vs high percentiles sampled at S1, S3, S6, and S7 for FAU-TSST, the high vs low percentiles measured at S5 for FAU-TSST, and the low vs mid-percentiles sampled at S4 and S5 for REG-TSST. This shows that predicted respiration signals can be valuable biomarkers of stress. Our results show that respiration can be successfully recognised from both speech and other audio, and that the predicted signals are used for identifying

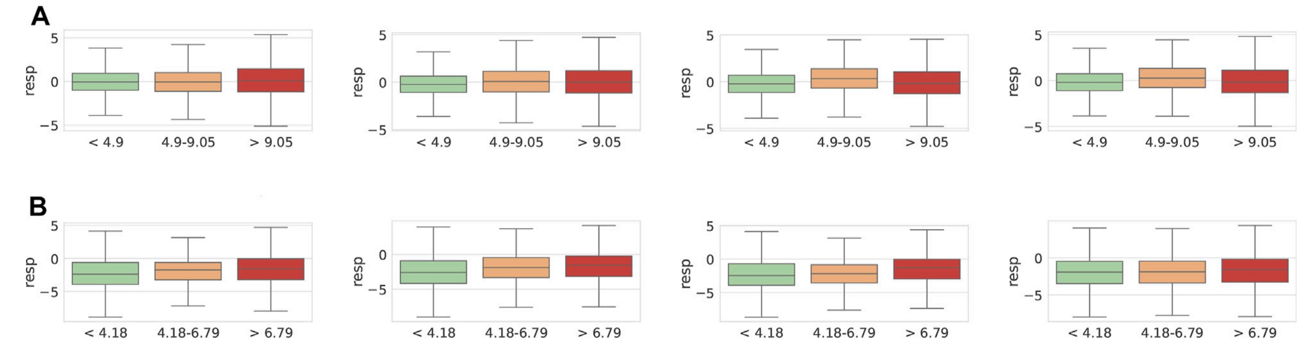

FIGURE 13 | Box plots of predicted respiration values for the FAU-TSST **(A)** and REG-TSST **(B)** datasets. Grouped based on raw cortisol (nmol/L) measures taken at time points S2-S5.

speaker states. As respiration prediction from other modalities remains an underexplored topic, our findings warrant a closer investigation in follow-up work.

7 LIMITATIONS

Dealing with human naturalistic data brings several challenges from a machine learning perspective, and from the analysis we have performed on the three corpora of interest, we see that variance in the display of physiological signals is one such challenge. In this current work, our approach was somewhat "brute-force" in nature, in that we did not condition the models or "correct" the targets with consideration to any specific subject or corpora variance. This can be more limiting when it comes to variation due to the assay applied for cortisol extraction. A transformation from raw cortisol values derived from different assays to cortisol factor scores for better comparison has been suggested by Miller et al. (2013), but this approach needs replication to ascertain its reliability and validity. It would be of interest to explore the benefit of this correction in future work, as well as other personalised training methods which may allow for a more robust result which in turn is more globally generalisable (Triantafyllopoulos et al., 2021).

Further to this, within the corpora themselves, there is a heavy gender bias, which it should be noted may have an implicate effect on the results obtained. For the FAU-TSSTand ULM-TSSTsets, this is particularly prominent. Although this is considered in the acoustic analysis conducted, the manifestation of stress is generally known to vary across genders. In further work, personalised training strategies would aid in exploring this potential bias. Similarly, regarding demographics, the mean age across all corpora is 24.02 years, with a reasonably small standard deviation of 4.13 years. This of course limits the current work as being only applicable to this age range, due to the inherent variance that stress is known to have throughout a lifetime, from factors including hormonal changes and overall life satisfaction, without deep experiments analysis these results should not be taken to be fully generalisable to a larger and more diverse population.

CONCLUSION

In the current contribution, we explored several markers of stress, learning from various modalities, with a core focus on the advantage of speech-based features. We processed and unified three different corpora collected under the well-known TSST, and we could verify our previous finding from (Baird et al., 2019) that audio features are best able to predict cortisol measurements taken approximately 15 min after the stress event. This effect was validated by a similar behaviour found on unlabelled data. This research establishes that audio can be utilised as a real-time guide for an individuals' current state of stress. Furthermore, a similar effect was found when using video-derived features from the face, meaning that a multimodal approach may provide further confidence, particularly given the potential periods of silence during stressful situations. Moreover, we have shown that emotion, heart rate, and respiration can be reliably recognised from speech during stress and have a strong relation to cortisol levels. This is found even when these physiological markers are not available during the data collection process but are predicted using other available modalities, mainly audio.

Our extensive analysis primarily shows that audio is suitable for the recognition of several physiological markers of stress. However, we do see, that as with many states of wellbeing, there is a large variance in stress manifestation in an individual, which makes generalisation a challenge. Given this, one needs to explore in follow-up work the potential for personalised machine learning strategies for this domain.

DATA AVAILABILITY STATEMENT

The datasets analysed for this study are not available in the public research domain, unless explicit consent is given via direct contact with the data owners. Namely, FAU-TSST; the Chair of Health Psychology, FAU Erlangen-Nuremberg, Germany—REG-TSST; the Institute of Psychology, University of

Regensburg—ULM-TSST; the Chair of Clinical Psychology and Psychology and Psychotherapy, University of Ulm, Germany.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the data owner's university ethics committee: Institute of Psychology, University of Regensburg, Germany—Chair of Health Psychology. FAU Erlangen-Nuremberg, Germany—Chair of Clinical Psychology and Psychotherapy, University of Ulm, Germany. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AB: literature analysis, data preparation, experimental design, computational analysis, manuscript drafting and preparation. AT: experimental design, computational analysis, manuscript drafting and preparation. SZ: data acquisition, literature analysis, computational analysis, manuscript editing. SO: data preparation, literature analysis, manuscript editing. LC: data preparation, literature analysis, manuscript editing. LS: literature analysis, manuscript editing. JK: data acquisition. SS: data acquisition. E-MM: data acquisition. BK: manuscript editing. NR: manuscript editing. HB: manuscript editing. BS: technical guidance and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

FUNDING

This project received funding from the German Research Foundation's (DFG) Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), the Zentrales Innovationsprogramm Mittelstand (ZIM) under grant agreement No. 16KN069455 (KIRun), and the DFG grant number KU 1401/6-1 (BM).

REFERENCES

Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Berger, A. C., and Cohen, R. J. (1981). Power Spectrum Analysis of Heart Rate Fluctuation: a Quantitative Probe of Beat-To-Beat Cardiovascular Control. *Science* 213, 220–222. doi:10.1126/science.6166045

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., et al. (2017). "Snore Sound Classification Using Image-Based Deep Spectrum Features," in *Proc. Interspeech* (Stockholm, Sweden, 2017, 3512–3516. doi:10.21437/interspeech.2017-434

Baird, A., Amiriparian, S., Cummins, N., Sturmbauer, S., Janson, J., Meßner, E.-M., et al. (2019). "Using Speech to Predict Sequentially Measured Cortisol Levels during a Trier Social Stress Test," in *Proc. Interspeech 2019* (India: Hyderabad), 534–538. doi:10.21437/interspeech.2019-1352

Baird, A., Cummins, N., Schnieder, S., and Schuller, B. W. (2020). "An Evaluation of the Effect of Anxiety on Speech–Computational Prediction of Anxiety from Sustained Vowels," in *Proc. INTERSPEECH 2020* (Shanghai, China: ISCA), 4951–4955.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A Theory of Learning from Different Domains. *Mach Learn.* 79, 151–175. doi:10.1007/s10994-009-5152-4

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., and Spadacini, G. (2000). Effects of Controlled Breathing, Mental Activity and Mental Stress with or without Verbalization on Heart Rate Variability. *J. Am. Coll. Cardiol.* 35, 1462–1469. doi:10.1016/s0735-1097(00)00595-7

Berntson, G. G., and Cacioppo, J. T. (2004). Heart Rate Variability: Stress and Psychiatric Conditions. *Dynamic Electrocardiography*, 57–64.

Bianco, S., Napoletano, P., and Schettini, R. (2019). "Multimodal Car Driver Stress Recognition," in Proc. International Conference on Pervasive Computing Technologies for Healthcare, 302–307. doi:10.1145/3329189.3329221

Brugnera, A., Zarbo, C., Tarvainen, M. P., Marchettini, P., Adorni, R., and Compare, A. (2018). Heart Rate Variability during Acute Psychosocial Stress: A Randomized Cross-Over Trial of Verbal and Non-verbal Laboratory Stressors. *Int. J. Psychophysiology* 127, 17–25. doi:10.1016/j.ijpsycho.2018.02.016

Cho, Y., Bianchi-Berthouze, N., and Julier, S. J. (2017). "Deepbreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition Using Low-Cost thermal Imaging in Unconstrained Settings," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (San Antonio, TX: ACII), 456–463. doi:10.1109/acii.2017.8273639

Cummins, N., Baird, A., and Schuller, B. (2018). The Increasing Impact of Deep Learning on Speech Analysis for Health: Challenges and Opportunities. *Methods Spec. Issue. Translational Data analytics Health Inform.* 151, 41–54. doi:10.1016/j.ymeth.2018.07.007

Cuno, A., Condori-Fernandez, N., Mendoza, A., and Lovón, W. R. (2020). "A Fair Evaluation of Public Datasets for Stress Detection Systems," in 2020 39th International Conference of the Chilean Computer Science (Coquimbo, Chile: Society SCCC), 1–8. doi:10.1109/SCCC51225.2020.9281274

Dalmeida, K. M., and Masala, G. L. (2021). Hrv Features as Viable Physiological Markers for Stress Detection Using Wearable Devices. *Sensors* 21, 2873. doi:10.3390/s21082873

Dhama, K., Latheef, S. K., Dadar, M., Samad, H. A., Munjal, A., Khandia, R., et al. (2019). Biomarkers in Stress Related Diseases/disorders: Diagnostic, Prognostic, and Therapeutic Values. *Front. Mol. Biosciences* 6. doi:10.3389/fmolb.2019.00091

Dickerson, S. S., and Kemeny, M. E. (2004). Acute Stressors and Cortisol Responses: a Theoretical Integration and Synthesis of Laboratory Research. *Psychol. Bull.* 130, 355. doi:10.1037/0033-2909.130.3.355

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affective Comput.* 7, 190–202. doi:10.1109/taffc.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in Proc. International Conference Multimedia (Barcelona, Spain: ACM), 835–838. doi:10.1145/2502081.2502224

Fendel, J. C., Bürkle, J. J., and Göritz, A. S. (2020). Mindfulness-based Interventions to Reduce Burnout and Stress in Physicians: a Systematic Review and Meta-Analysis. *Acad. Med.* 96, 751–764.

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., and Tørresen, J. (2018). Mental Health Monitoring with Multimodal Sensing and Machine Learning: A Survey. *Pervasive Mobile Comput.* 51, 1–26. doi:10.1016/j.pmcj.2018.09.003

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), 776–780. doi:10.1109/icassp.2017.7952261

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., and Winter, A. S. (2013). Vocal Indices of Stress: a Review. *J. voice* 27, 390–e21. doi:10.1016/j.jvoice.2012.12.010

Goldstein, D. S. (1987). Stress-induced Activation of the Sympathetic Nervous System. *Bailliere's Clin. Endocrinol. Metab.* 1, 253–278. doi:10.1016/s0950-351x(87)80063-0

Gönülateş, S., Tetik, S., Dündar, U., Tansu, Y., and Dündar, K. (2017). Analyzing the before and after Effects of Endurance Training on Acth Hormone. *Int. J. Sport Cult. Sci.* 5, 340–346.

Goodman, W. K., Janson, J., and Wolf, J. M. (2017). Meta-analytical Assessment of the Effects of Protocol Variations on Cortisol Responses to the Trier Social Stress Test. *J. Psychoneuroendocrinology* 80, 26–35. doi:10.1016/j.psyneuen.2017.02.030

Grzadzielewska, M. (2021). Using Machine Learning in Burnout Prediction: A Survey. *Child. Adolesc. Soc. Work J.* 38, 175–180.

Hagerer, G., Pandit, V., Eyben, F., and Schuller, B. (2017). "Enhancing Lstm Rnn-Based Speech Overlap Detection by Artificially Mixed Data," in Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio. no pagination.

Haider, F., De La Fuente, S., and Luz, S. (2019). An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Selected Top. Signal Process.* 14, 272–281.

Hansen, J. H., and Bou-Ghazale, S. E. (1997). "Getting Started with Susas: A Speech under Simulated and Actual Stress Database," in *Proc. Eurospeech.*, 1743–1746.

Healey, J., and Picard, R. (2005). Detecting Stress during Real-World Driving Tasks Using Physiological Sensors. *IEEE Trans. Intell. Transportation Syst.* 6, 156–166. doi:10.1109/TITS.2005.848368

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "Cnn Architectures for Large-Scale Audio Classification," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), 131–135. doi:10.1109/icassp.2017.7952132

Ishii, R., Otsuka, K., Kumano, S., and Yamato, J. (2016). Using Respiration to Predict Who Will Speak Next and when in Multiparty Meetings. *ACM Trans. Interactive Intell. Syst. (Tiis)* 6, 1–20. doi:10.1145/2946838

Jati, A., Williams, P. G., Baucom, B., and Georgiou, P. (2018). "Towards Predicting Physiology from Speech during Stressful Conversations: Heart Rate and Respiratory Sinus Arrhythmia," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), 4944–4948. doi:10.1109/icassp.2018.8461500

Johnson, A. K., and Anderson, E. A. (1990). *Stress and Arousal*. Ithaca, NY: APA PsycNET.

Kim, S., Kwon, N., and O'Connell, H. (2019). *Toward Estimating Personal Well-Being Using Voice*.arXiv preprint arXiv:1910.10082

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. *J. Neuropsychobiology* 28, 76–81. doi:10.1159/000119004

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., and Kraaij, W. (2014). "The Swell Knowledge Work Dataset for Stress and User Modeling Research," in Proc. of the 16th international conference on multimodal interaction, 291–298. doi:10.1145/2663204.2663257

Kovalenko, A., Kastyro, I., Torshin, V., Guhschina, Y., Doroginskaya, E., and Kamanina, N. (2019). "Comparison of Immediate Effects of Vocal Breathing Exercises and Physical Exercises on Heart Rate Variability in Healthy Students," in Proc. Models and Analysis of Vocal Emissions for BioMedical Applications: International Workshop (Firenze, Italy: Firenze University Press), 245.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25. Editors F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 1097–1105.

Kumar, A., Sharma, K., and Sharma, A. (2021a). Hierarchical Deep Neural Network for Mental Stress State Detection Using Iot Based Biomarkers. *Pattern Recognition Lett.* 145, 81–87. doi:10.1016/j.patrec.2021.01.030

Kumar, S., Iftekhar, A., Goebel, M., Bullock, T., MacLean, M. H., Miller, M. B., et al. (2021b). "Stressnet: Detecting Stress in thermal Videos," in Proc. of International Conference on Applications of Computer Vision, 999–1009. doi:10.1109/wacv48630.2021.00104

Leistner, C., and Menke, A. (2020). "Hypothalamic–pituitary–adrenal axis and Stress," in *Handbook of Clinical Neurology* (Elsevier), 175, 55–64. doi:10.1016/b978-0-444-64123-6.00004-7

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep Learning Face Attributes in the Wild," in Proc. of International Conference on Computer Vision. no pagination. doi:10.1109/iccv.2015.425

MacLaughlin, B. W., Wang, D., Noone, A.-M., Liu, N., Harazduk, N., Lumpkin, M., et al. (2011). Stress Biomarkers in Medical Students Participating in a Mind Body Medicine Skills Program. *Evidence-Based Complement. Altern. Med.* 2011. doi:10.1093/ecam/neq039

Mertes, S., Baird, A., Schiller, D., Schuller, B. W., and André, E. (2020). "An Evolutionary-Based Generative Approach for Audio Data Augmentation," in Proc. 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland (IEEE), 1–6. doi:10.1109/mmsp48831.2020.9287156

Miller, R., Plessow, F., Rauh, M., Gröschl, M., and Kirschbaum, C. (2013). Comparison of Salivary Cortisol as Measured by Different Immunoassays and Tandem Mass Spectrometry. *Psychoneuroendocrinology* 38, 50–57. doi:10.1016/j.psyneuen.2012.04.019

Mousa, A. E.-D., and Schuller, B. W. (2016). "Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks for Grapheme-To-Phoneme Conversion Utilizing Complex many-to-many Alignments," in *Interspeech*, 2836–2840. doi:10.21437/interspeech.2016-1229

Nath, R. K., Thapliyal, H., and Caban-Holt, A. (2021). Machine Learning Based Stress Monitoring in Older Adults Using Wearable Sensors and Cortisol as Stress Biomarker. *J. Signal Process. Syst.*, 1–13. doi:10.1007/s11265-020-01611-5

Niu, X., Han, H., Shan, S., and Chen, X. (2018). "Synrhythm: Learning a Deep Heart Rate Estimator from General to Specific," in 2018 24th International Conference on Pattern Recognition (ICPR) (IEEE), 3580–3585. doi:10.1109/icpr.2018.8546321

Orlikoff, R. F., and Baken, R. (1989). The Effect of the Heartbeat on Vocal Fundamental Frequency Perturbation. *J. Speech, Lang. Hearing Res.* 32, 576–582. doi:10.1044/jshr.3203.576

Pagán-Castaño, E., Maseda-Moreno, A., and Santos-Rojo, C. (2020). Wellbeing in Work Environments. *J. Business Res.* 115, 469–474. doi:10.1016/j.jbusres.2019.12.007

Panicker, S. S., and Gayathri, P. (2019). A Survey of Machine Learning Techniques in Physiology Based Mental Stress Detection Systems. *Biocybernetics Biomed. Eng.* 39, 444–469. doi:10.1016/j.bbe.2019.01.004

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep Face Recognition," in Proc. of the British Machine Vision Conference, 1–41. 12. doi:10.5244/c.29.4141

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Pisanski, K., Nowak, J., and Sorokowski, P. (2016). Individual Differences in Cortisol Stress Response Predict Increases in Voice Pitch during Exam Stress. *Physiol. Behav.* 163, 234–238. doi:10.1016/j.physbeh.2016.05.018

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'Absi, M., et al. (2011). "Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment," in Proc. of the 10th ACM/IEEE international conference on information processing in sensor networks (IEEE), 97–108.

Protopapas, A., and Lieberman, P. (1997). Fundamental Frequency of Phonation and Perceived Emotional Stress. *The J. Acoust. Soc. America* 101, 2267–2277. doi:10.1121/1.418247

Rodríguez-Arce, J., Lara-Flores, L., Portillo-Rodríguez, O., and Martínez-Méndez, R. (2020). Towards an Anxiety and Stress Recognition System for Academic Environments Based on Physiological Features. *Comp. Methods Programs Biomed.* 190, 105408. doi:10.1016/j.cmpb.2020.105408

Rohleder, N., and Nater, U. M. (2009). Determinants of Salivary α-amylase in Humans and Methodological Considerations. *Psychoneuroendocrinology* 34, 469–485. doi:10.1016/j.psyneuen.2008.12.004

Russell, J. A. (1980). A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* 39, 1161–1178. doi:10.1037/h0077714

Saitis, C., and Kalimeri, K. (2018). Multimodal Classification of Stressful Environments in Visually Impaired Mobility Using Eeg and Peripheral Biosignals. *IEEE Trans. Affective Comput.* 12, 203–214.

Šalkevicius, J., Damaševičius, R., Maskeliunas, R., and Laukienė, I. (2019). Anxiety Level Recognition for Virtual Reality Therapy System Using Physiological Signals. *Electronics* 8, 1039.

Sano, A., and Picard, R. W. (2013). "Stress Recognition Using Wearable Sensors and mobile Phones," in 2013 Humaine association conference on affective computing and intelligent interaction (IEEE), 671–676. doi:10.1109/acii.2013.117

Sawilowsky, S. S., and Blair, R. C. (1992). A More Realistic Look at the Robustness and Type Ii Error Properties of the T Test to Departures from Population Normality. *Psychol. Bull.* 111, 352. doi:10.1037/0033-2909.111.2.352

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. (2018). “Introducing Wesad, a Multimodal Dataset for Wearable Stress and Affect Detection,” in Proc. of International Conference on Multimodal Interaction (Boulder, CO, USA: Association for Computing Machinery), 400–408. doi:10.1145/3242969.3242985

Schuller, B., Friedmann, F., and Eyben, F. (2013). “Automatic Recognition of Physiological Parameters in the Human Voice: Heart Rate and Skin Conductance,” in Proc. International Conference on Acoustics, Speech and Signal Processing (IEEE), 7219–7223. doi:10.1109/icassp.2013.6639064

Schuller, B. W., Batliner, A., Bergler, C., Meßner, E.-M., Hamilton, A., Amiriparian, S., et al. (2020). “The Interspeech 2020 Computational Paralinguistics challenge: Elderly Emotion, Breathing & Masks,” in Proc. Interspeech 2020 (Shanghai, China: ISCA), 2042–2046. doi:10.21437/interspeech.2020-32

Sharma, S., Singh, G., and Sharma, M. (2021). A Comprehensive Review and Analysis of Supervised-Learning and Soft Computing Techniques for Stress Diagnosis in Humans. *Comput. Biol. Med.* 134, 104450. doi:10.1016/j.compbiomed.2021.104450

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv preprint arXiv:1409.1556.

Smith, J., Tsiartas, A., Shriberg, E., Kathol, A., Willoughby, A., and de Zambotti, M. (2017). “Analysis and Prediction of Heart Rate Using Speech Features from Natural Speech,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), 989–993. doi:10.1109/icassp.2017.7952304

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Meßner, E., et al. (2021a). “The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress,” in Proc. 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (Chengdu, China: ACM). [to appear].

Stappen, L., Baird, A., Schumann, L., and Schuller, B. (2021b). “The Multimodal Sentiment Analysis in Car Reviews (Muse-car) Dataset: Collection, Insights and Improvements,” in *IEEE Transactions on Affective Computing*. doi:10.1109/taffc.2021.3097002

Stappen, L., Schumann, L., Sertolli, B., Baird, A., Weigel, B., Cambria, E., et al. (2021c). “Muse-toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox,” in Proc. 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (Chengdu, China: ACM). [to appear].

Suess, W. M., Alexander, A. B., Smith, D. D., Sweeney, H. W., and Marion, R. J. (1980). The Effects of Psychological Stress on Respiration: a Preliminary Study of Anxiety and Hyperventilation. *Psychophysiology* 17, 535–540. doi:10.1111/j.1469-8986.1980.tb02293.x

Sun, L., Lian, Z., Tao, J., Liu, B., and Niu, M. (2020). “Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism,” in Proc. Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, 27–34. doi:10.1145/3423327.3423672

Taelman, J., Vandeput, S., Spaepen, A., and Van Huffel, S. (2009). “Influence of Mental Stress on Heart Rate and Heart Rate Variability,” in 4th European conference of the international federation for medical and biological engineering (Springer), 1366–1369. doi:10.1007/978-3-540-89208-3_324

Triantafyllopoulos, A., Liu, S., and Schuller, B. W. (2021). “Deep Speaker Conditioning for Speech Emotion Recognition,” in 2021 IEEE International Conference on Multimedia and Expo (ICME) (IEEE), 1–6. doi:10.1109/icme51207.2021.9428217

Wittchen, H.-U., Zaudig, M., and Fydrich, T. (1997). *SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II*. Handanweisung. (Hogrefe).

Wu, D., Courtney, C. G., Lance, B. J., Narayanan, S. S., Dawson, M. E., Oie, K. S., et al. (2010). Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals: Virtual Reality Stroop Task. *IEEE Trans. Affective Comput.* 1, 109–118. doi:10.1109/t-affc.2010.12

Yang, S., Luo, P., Loy, C. C., and Tang, X. (2015). WIDER FACE: A Face Detection Benchmark. *CoRR abs*/1511, 06523.

Zafar, T. (2020). Potential Biomarkers of Emotional Stress Induced Neurodegeneration. *eNeurologicalSci* 21, 100292. doi:10.1016/j.ensci.2020.100292

Zänkert, S., Kudielka, B. M., and Wuest, S. (2020). Effect of Sugar Administration on Cortisol Responses to Acute Psychosocial Stress. *Psychoneuroendocrinology* 115, 104607. doi:10.1016/j.psyneuen.2020.104607

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal. Process. Lett.* 23. doi:10.1109/lsp.2016.2603342

Zhang, Q., Chen, X., Zhan, Q., Yang, T., and Xia, S. (2017). Respiration-based Emotion Recognition with Deep Learning. *Comput. Industry* 92, 84–90. doi:10.1016/j.compind.2017.04.005

Conflict of Interest: AT and BS were employed by AudEERING GmbH

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.