A Physiologically-Adapted Gold Standard for Arousal during Stress

Alice Baird
EIHW, University of Augsburg
Augsburg, Germany

Lukas Stappen
EIHW, University of Augsburg
Augsburg, Germany

Lukas Christ
EIHW, University of Augsburg
Augsburg, Germany

Lea Schumann
EIHW, University of Augsburg
Augsburg, Germany

Eva-Maria Meßner
KPP, University of Ulm
Ulm, Germany

Björn W. Schuller
GLAM, Imperial College London
London, United Kingdom

ABSTRACT

Emotion is an inherently subjective psycho-physiological human state and to produce an agreed-upon representation (gold standard) for continuously perceived emotion requires time-consuming and costly training of multiple human annotators. With this in mind, there is strong evidence in the literature that physiological signals are an objective marker for states of emotion, particularly arousal. In this contribution, we utilise a multimodal dataset captured during a Trier Social Stress Test to explore the benefit of fusing physiological signals – Heartbeats per Minute (*BPM*), Electrodermal Activity (*EDA*), and Respiration-rate – for recognition of continuously perceived arousal utilising a Long Short-Term Memory, Recurrent Neural Network architecture, and various audio, video, and textual based features. We use the MuSe-Toolbox to create a gold standard that considers annotator delay and agreement weighting. An improvement in Concordance Correlation Coefficient (CCC) is seen across features sets when fusing *EDA* with arousal, compared to the arousal only gold standard results. Additionally, BERT -based textual features' results improved for arousal plus all physiological signals, obtaining up to .3344 CCC (.2118 CCC for arousal only). Multimodal fusion also improves CCC. Audio plus video features obtain up to .6157 CCC for arousal plus *EDA*, *BPM*.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Affective Computing; Stress; Multimodal Fusion

ACM Reference Format:
Alice Baird, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Meßner, and Björn W. Schuller. 2021. A Physiologically-Adapted Gold Standard for Arousal during Stress. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge (MuSe '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3475957.3484446

ACM ISBN 978-1-4503-8678-4/21/10…$15.00
https://doi.org/10.1145/3475957.3484446

1 INTRODUCTION

Physiological and emotional responses can coincide during a stressful situation [1], and the degree of correlation has shown to be dependent on factors including underlying psychological traits and states, e. g., social desirability, or physiological dispositions, e. g., brain morphology [2]. For research on the discrepancy between physiological and self-reported emotional states see [3]. During a stress-inducing situation, heart-rate, and breath become varied [4], along with the voice [5] (which is related strongly to perceived affect [6]). To this end, signals such as the Electrodermal Activity (*EDA*) – described as a psycho-physiological indication of emotional arousal [7] – correlate with an individual current perceived emotional state, specifically during high states of arousal, e. g., during a competitive video game [8].

Within the field of affective computing, recognition approaches to predict continuous states of emotion frequently utilise the two-dimensional Circumplex Model of Affect [9], observing the arousal (activation) and valence (positivity) of perceived emotion. However, as emotion is a subjective state of being, multiple raters must continuously annotate, which is time-consuming and costly. Further to this, the method to obtain a robust agreed-upon signal from multiple raters (gold standard) remains an ongoing research question, with several methods available. For example, given the likelihood of disagreement, weighting annotators based on level of agreement can be applied using the Evaluator Weighted Estimator (EWE) [10]. Furthermore, annotator delay is not consistent per annotator, and so aligning rating with consideration to peaks is needed and Canonical Time Warping (CTW) [11] can be applied in this case.

With this in mind, research into the fusion of physiological signals for use with perceived emotional signals is limited, and within this contribution we suggest that there are potentially two benefits to this (1) where agreement between raters is lower, replacing less reliable raters with a physiological signal may improve agreement (2) where only a small number of raters are available, adding a physiological signal to the gold standard may also be fruitful . Physiological signals have been utilised as features [12], or extracted during particular tasks, to better target arousal [13], however there has been minimal research on a combined physiological and perceived arousal gold standard. Recently, in the 2021 edition of the Multimodal Sentiment in-the-wild (MuSe) challenge, the signal of arousal was fused with *EDA* and used as a prediction target for the *MuSe-Physio* sub-challenge [14]. The baseline result from this was 0.3 CCC stronger than the arousal only *MuSe-Stress* sub-challenge when performing a late-fusion of audio and video-based

Table 1: Reported are the number (#) of speakers and total duration of the data splits across Train, (Devel)opment and Test partitions for the sub-set of the Ulm-TSST dataset.

	Train	Devel.	Test	∑
#	33	9	11	53
hh:mm:ss	2:45:29	0:45:32	0:55:33	4:26:36

features. Furthermore, the text-based features (typically less helpful for recognition of arousal) also improved through *EDA* fusion with arousal, showing promise that has encouraged the authors to investigate further. However, to the best of the authors' knowledge, this was the first time that arousal and *EDA* were fused in this manner, and there are no works that explore the fusion of arousal with other physiological signals such as respiration or heart rate (as *BPM*).

To explore this idea further, in this contribution, we utilise the same dataset available through the MuSe Challenge, the Ulm-Trier Social Stress dataset (Ulm-TSST), and explore the fusion of *EDA*, *BPM*, and respiration rate with arousal ratings. The Trier Social Stress Test (TSST), which the subjects of the Ulm-TSST dataset were undergoing, consists of a free-speech job interview scenario. Given this pseudo-professional setting, we specifically consider that utilisation of physiological signals (a more objective marker for arousal) will be of use here, as perceived arousal may be suppressed to make a better impression towards the interviewer [7]. For our experiments, we utilise the recently released MuSe-Toolbox[15][1], to apply a novel approach *Rater Aligned Annotation Weighting* (RAAW) for signal fusion which considers both weighting and alignment of ratings to create a gold standard. We extract several multimodal features from audio, video, and textual transcriptions and apply a Long Short-Term-Memory-Recurrent Neural Network (LSTM-RNN) as a regressor, following a similar training procedure as outlined in the MuSe 2021 challenge.

2 THE ULM-TSST DATASET

We make use of the Ulm-Trier Social Stress dataset (Ulm-TSST) for our experiments, a multimodal dataset first utilised as a part of the MuSe 2021 challenge [14]. Within the Ulm-TSST dataset, the subjects are undergoing a TSST, which is a standardised and renowned experiment to induce states of stress, allowing for a controlled setting with high-quality data. The full Ulm-TSST dataset consists of recordings from 110 German-speaking individuals (ca. 10 hours), which are annotated for the continuous dimensions of emotion (valence and arousal). The continuous emotion ratings are recorded at a sampling rate of 2 Hz and made by three annotators (obtaining an average inter-rater agreement for arousal of .173 Pearson's Correlation Coefficient). In addition, the modalities of audio, video, and text can be extracted from the dataset, as well as four captured physiological signals at a sampling rate of 1 kHz: Electrodermal Activity (*EDA*), Electrocardiogram (ECG), respiration rate (*RESP*) as chest displacement during breath [-10:+10], and heart rate as beats per minute (*BPM*). For our experiments, we utilise a sub-set of the dataset as presented in the MuSe 2021 challenge, which was further processed, and reduced to 53 speakers.

Table 2: The mean (μ) and standard deviation (±) for inter-rater agreement, as Pearson correlation coefficient (CC). Calculated during EWE after CTW alignment.

CC	μ	±
A_1,A_2,A_3	.173	.191
A_1, A_2, *EDA*	**.230**	.241
A_2,A_2 + *BPM*	.158	.187
A_2,A_2 + *RESP*	.108	.134
A_1,A_2,A_3, *EDA*,*BPM*	.119	.155
A_1,A_2,A_3, *EDA*,*RESP*	.088	.120
A_1,A_2,A_3, *BPM*,*RESP*	.070	.097
A_2, A_2, *EDA*, *BPM*, *RESP*	.127	.123
EDA, *BPM*, *RESP*	.197	.149

The data is in a speaker-independent train, development, and test partitioning, with balanced speaker demographics across the partitions, cf. Table 1. Before feature extraction, videos are cut from start to end of the TSST, and excluding participant names. We choose to use only *EDA*, *BPM* and *RESP* for the physiological signals, and each is down-sampled to 2 Hz (to match the arousal ratings) and smoothed, applying a Savitzky–Golay filter, to reduce irrelevant, fine-grained artefacts in the signal. We exclude the ECG signal, as *BPM* captures this activity at a higher level which is more optimal for the applied down-sampling.

3 EXPERIMENTAL SETTINGS

To evaluate the benefit of fusing physiological signals with perceived arousal, we primarily conduct a series of continuous recognition tasks utilising various combinations of the three perceived arousal ratings with the *EDA*, *BPM*, and *RESP* signals.

(1) A_1, A_2, A_3: Perceived arousal ratings only. From annotators one, two, and three.
(2) A_1, A_2, *EDA*: Arousal rater one (A_1) and Arousal rater two (A_2) plus EDA. A_1 and A_2 are chosen as correlation is slightly higher for these signals compared to A_3, as shown in Figure 1.
(3) A_1, A_2, *BPM*.
(4) A_1, A_2, *RESP*.
(5) A_1,A_2,A_3, *EDA*,*BPM*.
(6) A_1,A_2,A_3, *EDA*,*RESP*.
(7) A_1,A_2,A_3, *BPM*,*RESP*.
(8) A_1, A_2, A_3, *EDA*, *BPM*, *RESP*.
(9) *EDA*, *BPM*, *RESP*: Physiological signals only.

3.1 Label Fusion Strategy

We utilise a continuous annotator fusion technique *Rater Aligned Annotation Weighting* (RAAW), first presented in [14]. RAAW is aimed at two challenges for gold standard creation of continuous emotion, a) annotator delay, and b) rater disagreement. Annotator delay is a typical issue with continuous ratings [16] and can be reduced through an explicit time-shift or through methods that automatically time-shift based on factors such as rater agreement, and in this instance, Canonical Time Warping (CTW) is applied [11]. For weighting the annotators based on their agreement, the Evaluator

[1] https://github.com/lstappen/MuSe-Toolbox

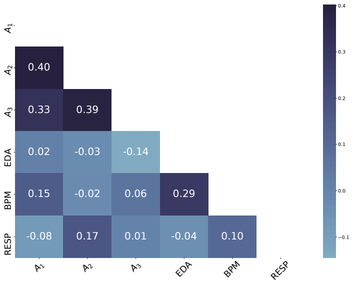

Figure 1: Correlation matrix between each individual signal across all speakers in the training set.

Weighted Estimator (EWE) [10] is commonly applied for emotional gold standard [17].

For our experiments, we create six gold standards as shown in Table 2 and above. For each, we are comparing to the perceived arousal only gold standard. There are two annotators (A_1, A_2) and a physiological signal. We explore the benefit of removing an annotator who is sub-optimal (in other words, the annotator with the lowest agreement). Where we apply two arousal raters with all physiological signals, we explore the advantage of using physiological signals to bring the rating in the gold standard up to five.

3.2 Features

For all features, we utilise the package provided from the MuSe 2021 challenge [14]. To reduce the scope of our experiments, we select the better performing features set from the baseline experiments. However, speech is strongly linked to perceived arousal, so we choose to use the two best performing feature sets. We offer a short description of critical points for the feature extraction process applied, for details cf. [14].

Audio: We apply a six-second window size for the acoustic features. As a first step, the entire audio sequence is extracted from a given video. This file is then converted from stereo to mono with a sampling of 16 kHz, 16 bit, and then normalised to -3 dB. For DeepSpectrum, we keep the default settings for extraction to obtain a 4 096-dimensional feature set. For VGGish, by aligning the frame and hop size to the annotation sample rate, we extract a 128-dimensional VGGish embedding vector every 0.5 s from the underlying log spectrograms.

Video: Given the human nature of this task, video-based features focus on the face, although it may be beneficial to explore gesture-based features more specifically in further research. The MTCNN [18] is used to extract faces. MTCNN is pre-trained on the datasets WIDER FACE [19] and CelebA [20]. The MTCNN extractions are used as inputs for VGGface. VGGface [21] is aimed at the extraction of general facial features and is based on detaching the top-layer of a pre-trained version of the deep CNN referred to as VGG16 [22]. This results in a 512 feature vector output.

Text: For extracting the text features from the transcripts, a pre-trained Transformer language model BERT [23], is used. We obtain word-level features from the sum of the last four BERT encoder layers resulting in a 768-dimensional feature vector for each word analogous to [24]. This data contains exclusively German speech. For this reason, the BERT pre-trained on German texts[2] is applied.

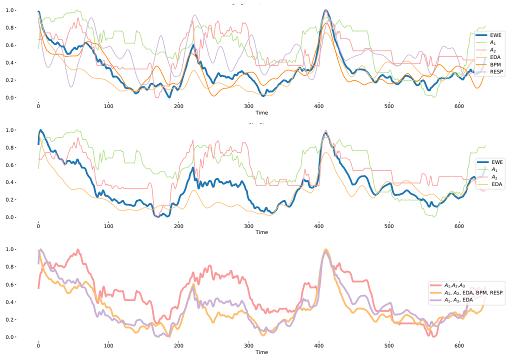

Figure 2: An example for subject # 9 for three of the gold standards used in these experiments. (Upper) A_1, A_2 + EDA, BPM and $RESP$ (±: 0.203). (Middle) A_1, A_2 + EDA (±: 0.217). (Lower) comparison of the above gold standards including arousal only (±: 0.241).

Alignment: The label-aligned features are made available from the MuSe challenge. These include the same frame rate as the provided label arousal labels. For the textual features of the Ulm-TSST, as they are based on manual transcripts of the videos, the Montreal Forced Aligner (MFA) [25] tool is applied to obtain word-level timestamps.

3.3 Regressor: LSTM-RNN

Given the time-dependent nature of this task, we utilise the same LSTM-RNN based architecture as applied for the baseline of the MuSe 2021 Challenge[3]. Extensive hyperparameter optimisation is applied, and the extracted feature sequences are input into a uni- and bi-directional LSTM-RNN with a hidden state dimensionality of $h = \{32, 64, 128\}$, to encode the feature vector sequences. We test different numbers of LSTM-RNN layers $n = \{1, 2, 4\}$, and search for a suitable learning rate $lr = \{0.0001, 0.001, 0.005\}$. For further detail of the architecture applied cf. [14]. In the training processes, the features and labels of every input video are further segmented via a windowing approach [24]. As in the MuSe Challenge, a window size of 300 steps (150 seconds) and a hop size of 50 steps (25 seconds) is used.

3.4 Feature Fusion

To observe the benefits of multimodal approaches, we apply a decision-level (late) fusion to evaluate the co-dependencies of the modalities. The experiments are restricted to the best performing features from each modality only. For decision-level fusion, separate models are trained individually for each modality. The predictions of these are fused by training an additional LSTM-RNN model as described above. For all tasks, we apply a unidirectional version with $lr = 0.001$, $h = 64$, and $n = 4$.

[2] https://deepset.ai/german-bert

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

Table 3: Reporting Concordance Correlation Coefficient (CCC) results for prediction of nine combinations of perceived arousal and physiological-arousal signals on the devel(opment) and test partitions. Utilising (V)ision: VGGFace, (A)udio: DeepSpectrum, VGGish, and (T)ext: BERT . Reporting the best result from hyperparameter optimisation, as well as reporting the mean (μ) across all feature sets for a given signal. Best test scores are emphasised.

Percieved	A_1,A_2,A_3		A_1,A_2		A_1,A_2		A_1,A_2		A_1,A_2,A_3		A_1,A_2,A_3		A_1,A_2,A_3		A_1,A_2			
Physiological			*EDA*		*BPM*		*RESP*		*EDA*,*BPM*		*EDA*,*RESP*		*BPM*,*RESP*		*EDA*,*BPM*,*RESP*		*EDA*,*BPM*,*RESP*	
CCC	Devel	Test	Devel	Test	Devel	Test	Devel	Test	Devel	Test	Devel	Test	Devel	Test	Devel	Test	Devel	Test
VGGFace	.3025	.3813	.3216	.3959	.4805	.3771	.1869	**.3745**	.3694	.4062	.3995	.3941	.3637	.4306	.4704	.4707	.5679	**.5838**
DeepSpectrum	.2826	.3060	.3366	.4031	.1649	.2327	.0382	.0977	.3089	.3861	.1841	.3807	.2527	.2046	.3683	.3832	.4189	.5157
VGGish	.2127	.2856	.3493	.4210	.3156	.3313	-.0079	.1716	.4851	.5164	.0901	.3985	.2689	.3649	.5161	.4712	.3197	.4613
BERT	.1341	.2118	.2431	.2402	.0567	.1037	.1063	.1802	.1999	.0542	.2733	.2393	.1210	.0922	.3568	.3344	.2909	.3842
Late-Fusion																		
A + V	.4638	**.5062**	.4506	**.5103**	.4640	.3889	.3196	.3108	.5666	**.6157**	.3630	**.3947**	.4722	**.4432**	.6674	.5025	.5030	.5728
A + T	.3240	.3841	.3821	.3470	.3044	.3205	.1396	.2032	.5089	.3677	.3249	.1777	.3295	.2817	.5570	.4357	.4175	.5586
V + T	.2526	.4668	.3442	.4213	.4735	**.4202**	.3443	.2871	.4839	.3783	.3836	.2301	.3738	.3881	.5916	**.5355**	.4386	.5594
A + V + T	.3476	.4965	.4186	.4987	.4458	.4104	.3811	.3036	.5895	.4596	.4028	.3470	.4086	.4230	.6669	.5055	.4623	.5639
μ of All	–	.3798	–	.4047	–	.3231	–	.2411	–	.3980	–	.3203	–	.3285	–	.4548	–	**.5250**

Table 4: Mean (μ) and standard deviation (±) for test results given in Table 3 which include either EDA, BPM, or RESP.

	A_1,A_2,A_3	**inc.** *EDA*		**inc.** *BPM*		**inc.** *RESP*	
		μ	±	μ	±	μ	±
VGGFace	.3813	.4167	.0364	.4212	.0396	.4175	.0424
DeepSpectrum	.3060	.3883	.0101	.3017	.0965	.2666	.1402
VGGish	.2856	.4518	.0527	.4210	.0872	.3516	.1279
BERT	.2118	.2170	.1174	.1461	.1273	.2115	.1018
Late-Fusion							
A + V	.5062	.5058	.0903	.4876	.0972	.4128	.0810
A + T	.3841	.3320	.1096	.3514	.0663	.2746	.1162
V + T	.4668	.3913	.1263	.4305	.0722	.3602	.1339
A + V + T	.4965	.4527	.0733	.4496	.0427	.3948	.0888

4 DISCUSSION OF RESULTS

To explore the benefit of physiological-based arousal and perceived arousal fusion, the extensive results for the computational prediction experiments conducted are given in Table 3, and Table 4. As an evaluation metric for these experiments, CCC is employed, as is typical for emotion recognition tasks, and to better compare to the initial baseline results obtained using the Ulm-TSST dataset [14].

For the results in Table 3, we see that the perceived arousal only (A_1-A_3) score is strong, particularly from a multimodal approach where at best 0.5062 CCC is achieved on the test set, from late-fusion of audio and video-based features. However, looking at the uni-modal approaches for A_1-A_3, as we expected given the pseudo-professional scenario of the TSST, audio-only features capture the perceived arousal to lesser degree compared to VGGFace. Furthermore, as is typical for arousal prediction tasks, the uni-modal textual features perform worst, obtaining 0.2118 CCC on the test set.

As we move along the table (cf. Table 3) to the right, we see in general a slight improvement across features when incorporating a physiological signal. Of interest, we see a ca. .3 CCC improvement for BERT features when utilising the *EDA* signal. Typically, perceived arousal is a challenging task for textual-based features, as seen from the perceived arousal baseline. However, at best for BERT features when predicting the combined A_1, A_2, *EDA*, *BPM*, *RESP* signal, we obtain .3344 CCC, which is .1 above the A_1-A_3 baseline. When observing the mean across experiments, including *EDA* Table 4, we confirm that *EDA* is the strongest physiological signal for the BERT features.

We also see the audio features obtain a more robust result than VGGFace when utilising *EDA*, suggesting that the behaviour of *EDA* is present in the voice, making this gold standard more attainable for the speech-based features. Similar behaviour for *BPM* and *RESP* fusion results are obtained; however, this is not as consistent as *EDA* results, as we can see through the more consistent mean results in Table 4. Furthermore, there are lower results for the A_1, A_2 *BPM*, and A_1, A_2 *RESP* results, compared to the A_1-A_3 baseline.

For audio features, all gold standard approaches, which include *EDA* report an improvement, and up to .4712 CCC is obtained by VGGish features, where all physiological signal are utilised. In Figure 2, we can see that the two physiological-adapted gold standards follow a similar trend to the arousal baseline gold standard. However, there is a slightly reduced standard deviation for this example, with 0.24 for A_1-A_3 compared to 0.22 for A_1, A_2, *EDA*, and 0.203 for the A_1, A_2, *EDA*, *BPM*, *RESP* signal. This may suggest that better results are obtained from a smoothing effect when the perceived arousal is fused with physiological signals.

To further analyse this smoothing effect, we see in Table 3 that the results are consistently higher than the arousal only baseline when utilising the physiological only (*EDA*-*RESP*) signal. With a standard deviation of around 0.157 for the same example in Figure 2, we do lean more toward this being a factor in results improvement. We additionally extract the mean absolute change (MAC), and skewness from each of the gold standard (Figure 3) across all speakers. Although further investigation should be done here, we see that there is an inherent difference in the MAC from A_1-A_3, and *EDA*-*RESP*, which is mirrored by the skew of the signals' distribution. Of promise, and perhaps opposing the smoothing effect, none of the physiological signal results obtains higher than the best result when fusing with perceived arousal, i. e., .6157 CCC from A_1, A_2, *EDA*, *BPM* with audio and video feature fusion. This leads us to consider that further investigation on this topic may be fruitful – particularly, as we do not see any reduction in results from physiological-adapted arousal fusion.

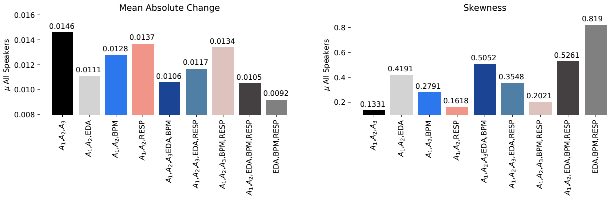

Figure 3: The mean absolute change and Skewness for the mean (μ) of all speaker from each gold standard signal.

5 CONCLUSION

To explore recognition of internal emotional states and improve the current state-quo for emotion-based gold standard creation, for the first time in this work, we explored the prediction of fused physical-based arousal with perceived arousal. We utilised the Ulm-TSST dataset, which offers a scenario in which stress is induced, and ultimately a testing bed for states of arousal. Findings have shown that in most cases, the *EDA* signal can improve recognition of arousal, specifically textual based features, and aid acoustic features, which the less aroused speech behaviours may have challenged. There was less of an improvement from *BPM* or *RESP* signals alone, however, when fused with *EDA*, various feature sets did see improvements, with the best score obtained from a fusion of perceived arousal with *BPM* and *EDA* of up to .6157 CCC with late-fusion of audio and video features. One observation consistent throughout the experiments was the reduction in the standard deviation for the gold standard with physiological signals. Results seem to indicate that this aided the learning process, and it would be of interest to explore this more deeply in future work.

As the original sampling rate available for the physiological signals was 1 kHz, it would also be interesting to explore this in more detail. For example, the Savitzky–Golay filtering and extreme down-sampling may have caused vital information loss. Furthermore, validation through other datasets would be optimal for exploring more deeply the combinations of signals and perhaps exploring a segmentation approach that is more suitable to the physiological signal. Another next step is to explore the valence dimension in this context. However, there is less literature supporting the manifestation of valence via physiological signals, so we have excluded it from our current research.

6 ACKNOWLEDGMENTS

This project has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

REFERENCES

[1] Stefanie Duijndam, Annemiek Karreman, Johan Denollet, and Nina Kupper. Physiological and emotional responses to evaluative stress in socially inhibited young adults. *Biological psychology*, 149:107811, 2020.

[2] Jana Campbell and Ulrike Ehlert. Acute psychosocial stress: does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology*, 37(8):1111–1134, 2012.

[3] Andreas Schwerdtfeger and Carl-Walter Kohlmann. Repressive coping style and the significance of verbal-autonomic response dissociations. *APA Psycnet*, 2004.

[4] Luciano Bernardi, Joanna Wdowczyk-Szulc, Cinzia Valenti, Stefano Castoldi, Claudio Passino, Giammario Spadacini, and Peter Sleight. Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology*, 35(6):1462–1469, 2000.

[5] Alice Baird, Shahin Amiriparian, Nicholas Cummins, Sarah Sturmbauer, Johanna Janson, Eva-Maria Messner, Harald Baumeister, Nicolas Rohleder, and Björn W. Schuller. Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test. In *Proc. Interspeech 2019*, pages 534–538, 2019.

[6] Nathaniel S Eckland, Teresa M Leyro, Wendy Berry Mendes, and Renee J Thompson. The role of physiology and voice in emotion perception during social stress. *Journal of Nonverbal Behavior*, 43(4):493–511, 2019.

[7] Delphine Caruelle, Anders Gustafsson, Poja Shams, and Line Lervik-Olsen. The use of electrodermal activity (eda) measurement to understand consumer emotions–a literature review and a call for action. *Journal of Business Research*, 104:146–160, 2019.

[8] Anders Drachen, Lennart E. Nacke, Georgios Yannakakis, and Anja Lee Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, page 49–54, New York, NY, USA, 2010. Association for Computing Machinery.

[9] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[10] Michael Grimm and Kristian Kroschel. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 381–385. IEEE, 2005.

[11] Feng Zhou and Fernando De la Torre. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, 2015.

[12] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*, pages 784–789, 2020.

[13] Alice Baird, Shahin Amiriparian, Manuel Milling, and Björn W. Schuller. Emotion recognition in public speaking scenarios utilising an lstm-rnn approach with attention. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 397–402, 2021.

[14] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. 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*, page [to appear], Chengdu, China, 2021. ACM.

[15] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigel, Erik Cambria, and Björn W. Schuller. Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox, 2021.

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

[17] Alice Baird, Shahin Amiriparian, Miriam Berschneider, Maximilian Schmitt, and Björn Schuller. Predicting biological signals from speech: introducing a novel multimodal dataset and results. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[18] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23, 04 2016.

[19] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. *CoRR*, abs/1511.06523, 2015.

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[21] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, September 2015.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[24] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 27–34, 2020.

[25] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of INTERSPEECH*, volume 2017, pages 498–502, 2017.