INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test

Alice Baird[1], *Shahin Amiriparian*[1], *Nicholas Cummins*[1], *Sarah Sturmbauer*[2], *Johanna Janson*[2], *Eva-Maria Messner*[3], *Harald Baumeister*[3], *Nicolas Rohleder*[2], *Björn Schuller*[1,4]

[1] ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Chair of Health Psychology, FAU Erlangen-Nuremberg, Germany
[3] Chair of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[4]GLAM – Group on Language, Audio and Music, Imperial College London, UK

alice.baird@informatik.uni-augsburg.de

Abstract

The effect of stress on the human body is substantial, potentially resulting in serious health implications. Furthermore, with modern stressors seemingly on the increase, there is an abundance of contributing factors which lead to a diagnosis of acute stress. However, observing biological stress reactions usually includes costly and time consuming sequential fluid-based samples to determine the degree of biological stress. On the contrary, a speech monitoring approach would allow for a non-invasive indication of stress. To evaluate the efficacy of the speech signal as a marker of stress, we explored, for the first time, the relationship between sequential cortisol samples and speech-based features. Utilising a novel corpus of 43 individuals undergoing a standardised Trier Social Stress Test (TSST), we extract a variety of feature sets and observe a correlation between speech and sequential cortisol measurements. For prediction of mean cortisol levels from speech, results show that for the entire TSST oral presentation, hand-crafted ComParE features achieve best results of 0.244 root mean square error [0;1] for the sample 20 minutes after the TSST. Correlation also increases at minute 20, with a Spearman's correlation coefficient of 0.421, and Cohen's d of 0.883 between the baseline and minute 20 cortisol predictions.

Index Terms: acoustic features, biological signals, cortisol, speech, Trier Social Stress Test, wellbeing.

1. Introduction

Stress can be a negative aspect of life affecting biological and mental states in various ways. Chronic states of stress can have an abundance of health-related consequences, e. g. , anxiety, weight gain and migraines, with occupational stress being a contributing factor to an employees overall task performance [1]. Objective markers of stress are therefore needed to help individuals recognise and appropriately respond to stressful situations.

One well-known marker of stress is *cortisol*, a hormone released by the body in response to stressful scenarios [2]. Cortisol is known to provoke a *fight-or-flight* response, which increases brain function and the availability of substances that repair tissues [3]. Through evolution, cortisol has manifested as a means of responding to physical threats [4]. In modern life, however, the adverse effects of cortisol can outweigh the positive, restricting nonessential functions, reducing the functionality of a series of bodily processes; e. g. , the immune, digestive, and reproductive systems [5]. Moreover, cortisol also impacts mood and motivation [6]. Common methods for measuring cortisol are typically invasive, requiring bodily fluid, e. g. , saliva- or blood-based measurements [7].

Since cortisol offers a well-established basis for understanding an individual's biological stress level, it is gathered as a standard practice during stress-inducing psychological paradigms [8]. These paradigms often include a public speech-based task e. g. , the *Stress Inducing Speech Task* (SIST) [9]. It has been observed that after the completion of such a speech task, there is a delayed response in cortisol levels, which can occur on average around 38 minutes after a stressor [10].

This finding has also been observed in the *Trier Social Stress Test* (TSST) [11]. The TSST is a well established and valid tool to induce an acute stress response [12]. The TSST is designed to exploit the vulnerable human response to stress during a social evaluation. Subjects are observed by interviewers, during three, 5 minute components which cause various stressful states: (1) anticipation, (2) presentation, and (3) mental arithmetic. Typically, blood and saliva samples are taken from the individuals sequentially, both before and after the main 15 minute oral presentation [13].

In this study, we utilise a novel dataset of audio recordings of 43 healthy individuals speaking during a TSST scenario. We investigate the efficacy of a variety of speech-based features (e. g. , ComParE , eGeMAPS and Deep Spectrum) to predict the sequentially measured samples of cortisol as a marker of biological stress. Previous research has demonstrated that speech features can be applied to physiological based signal predictions [14, 15]. However, to the best of our knowledge, this is the first time that cortisol has been the prediction target.

Similar topics relating to stress prediction from speech include cognitive and physical load analysis [16]. Speech can be reliably used to predict high and low states of both conditions [17]. Typical speech changes associated with stress include increased articulation rate and the number of filled pauses, as well a reduction in formant vowel space [18]. The effects of stress were also investigated on works conducted on the *Speech under Simulated and Actual Stress* (SUSAS) dataset [19]. The SUSAS data is focused on investigating the Lombard effect – an involuntary increase of speech volume – which has shown to occur for most individuals during states of stress [19].

This rest of this paper is organised as follows. In the proceeding section (Section 2), the FAU-TSST corpus used in our experiments is presented, including study procedure and data processing. We then describe our experimental settings for the task of cortisol prediction from speech in Section 3, followed by a discussion of results in Section 4. Finally, conclusions and future work plans are given in Section 5.

http://dx.doi.org/10.21437/Interspeech.2019-1352

2. FAU-TSST Corpus

To explore the relationship which may exist between speech and cortisol, we utilise the *Friedrich-Alexander-Universität Erlangen-Nürnberg, Trier Social Stress Test* (FAU-TSST) corpus. This corpus is a multimodal dataset of 43 individuals undergoing the renowned Trier Social Stress Test (TSST) [13].

2.1. Participants

Participants were recruited from the University campus and the community via print and multi-media advertising and received monetary compensation[1]. Before testing, eligibility was assessed by an online screening-questionnaire. Exclusion criteria were younger than 18 years of age, smoking, a body mass index (BMI) below 18 or above 30 kg/m^2, medication intake (e. g. , beta blocker, glucocorticoids, anti-depressants), with the exception of hormonal contraceptives in women, presence of physical or mental disorders, including clinically relevant depressive symptoms, and previous experience with the TSST protocol.

To exclude the effects of depression on stress responses [20], the *Allgemeine Depressionsskala* (ADS-L) scale [21] (Translation: General Depression Scale), was used during screening. Participants with ADS-L scores above 22 were excluded as scores above this cutoff indicate the presence of depressive symptomatology [21]. After data processing, the final sample used within this study consisted of 43 participants (μ age = 24.26 years, $\pm$ = 4.97 years, 65.1 % female).

2.2. Study Procedure

The laboratory testings took place on two consecutive days (although we only utilised data from first day), and the audio material used for the present analyses were captured on study day 1. On each study day, participants were scheduled between 1 :00 p.m. and 7 :00 p.m. to account for the influence of circadian cortisol variations [22]. Participants were instructed to refrain from exercising for 24 hours before the visit, and from smoking, brushing teeth, eating, and drinking anything except water for one hour before the visit.

After arrival at the laboratory, participants were accommodated in a comfortable armchair and received verbal and written instructions. This step was followed by a resting period of 45 minutes to ensure adequate recovery from travelling to the laboratory. During this period all participants provided verbal and written informed consent and a baseline saliva sample (-30 minutes) was collected as the participant's individual cortisol baseline (S0).

Participants were then introduced to a modified version of the original Trier Social Stress Test (TSST) [13] procedure, in the course of which they were guided to a test room and were introduced to a selection committee consisting of one female and one male observer wearing white lab coats. Participants were then instructed to take the role of a job applicant and to give a five-minute speech in order to present themselves as the best candidate for a vacant position (i. e. , their 'dream job'). They were given five minutes to prepare for their talk, after which they were instructed to stand in front of the panel and to begin their speech (hence forth named as the 'interview' scenario).

[1]Before commencement, the TSST study was approved by the Ethics Committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg Medical School. The study was carried out in accordance with the declaration of Helsinki, and informed consent was obtained from all participants at study entry.

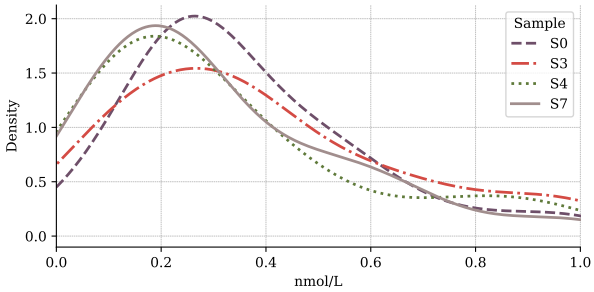

Figure 1: *Density distribution of normalised cortisol samples from selected measurements in the sequence (S0, S3, S4 and S7), across all speakers in the corpus.*

Table 1: *Speaker independent partitions, Train, (Val)idation, and Test. Including audio instances for each scenario.*

	Train	Val	Test	$\sum$
Speakers	15	15	14	43
Gender M:F	6:9	5:10	3:10	14:29
Interview	337	309	289	935
Arithmetic	363	311	401	1075
Combined	700	690	620	2010

Once their presentation was completed, participants were given a mental arithmetic task (hence forth named as the 'arithmetic' scenario), which took an additional five minutes, in which they were asked to serially subtract 17 from 2 043 as quickly as possible. In the case of any error, they were requested to start over again. The panel remained neutral during the entire task (only correcting the mistake). Immediately before at -1 minute (S1) as well as +1 (S2), +10 (S3), +20 (S4), +30 (S5), +45 (S6), and +60 (S7) minutes after the TSST, additional saliva samples were collected for the participants.

On the second day, participants were scheduled at approximately the same time. Procedures were similar to session one of the TSST procedures with a slightly modified version of the arithmetic scenario (i. e. , serially subtracting 13 from 2 011). At the end of the second laboratory session, participants were debriefed and dismissed.

2.3. Data Processing

The audio data which is utilised in this study, was captured during the first day of the TSST protocol, using a Sony HDR-CX240E video camera, placed approximately 3 meters from the subject. Given this recording scenario, the extracted audio data (44.1 kHz, 16 bit, stereo, WAV) required a series of additional processing steps in order to make it suitable for computational audio analysis.

The first step was extracting the audio channel from the video and converting the raw data to 16 kHz, 16bit, mono, WAV files. Then, we cut the first 2 minutes (interview) and the last 2 minutes (arithmetic) from the speech data to reduce the possibility of interviewer speech. This step was undertaken consistently for all participants.

Following this, we applied *voice activity detection*, utilising the Python implementation of the webrtcvad toolkit[2], and chunked the audio based on speech pauses. Given that the interviewer corrects any mistakes that the subject makes during

[2]`https://github.com/wiseman/py-webrtcvad`

the arithmetic task, we then manually removed chunks containing interviewer speech found in the arithmetic data. The data is then normalised to -1dB across the dataset. Post-processing, 01 h: 35 m: 49 s (μ 2.8 s, $\pm$ 2.5 s per instance) of audio data remained, of which 58 m: 28 s (μ 3.75 s, $\pm$ 3.24 s per instance) was from the interview scenario and 37 m: 22 s (μ 2.08 s, $\pm$ 1.20 s per instance) from the arithmetic scenario.

For each subject in the dataset, there are eight sequential saliva-based cortisol samples taken at consistent time intervals from S0 (baseline -30mins) to S7 (+60mins) (cf. Section 2.2). These have as raw values, range from 0.69–35.48 Nanomole/liter (nmol/L) across all subjects and measurements. Given the variance, we standardised the values to zero mean and unit standard deviation on a per-subject basis (cf. Figure 1). For the prediction task, we then applied normalisation across all subjects, resulting in a range of [0 ;1] nmol/L.

3. Experimental Setting

Saliva-based measurements were taken at 8 sequential time steps throughout the TSST process[3]. Given that the response of cortisol to a stressor can be delayed (by approximately 38mins [10]), we perform a series of speech-based regression experiments on each of the coritosl samples (S0–S7). The aim of this analysis is to ascertain if there are any changes in the level of correlation, between our speech features and the cortisol measurements, across the various time points. We also separate the different speech scenarios, namely: interview (1) and arithmetic (2) tasks, to explore the effect of speech type.

3.1. Feature sets

We extracted both hand-crafted speech-based features, namely the *Computational Paralinguistics challengE* (ComParE) feature set, and the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS). We also, as a state-of-the-art approach, extracted deep data representations from the speech signals (utilising the Deep Spectrum toolkit).

As a conventional approach, the 6 373 dimensional ComParE feature set [23] of hand-crafted speech-based features is used, given its effective brute-force approach for an abundance of similar paralinguistic tasks [24, 25, 26]. Additionally, we extract the 88 dimensional eGeMAPS feature set [27], and, much like the ComParE feature set, eGeMAPS has shown efficacy for tasks similar to the present study [28, 29]. From each instance, the ComParE and eGeMAPS acoustic features are extracted with the openSMILE toolkit [23]. Utilising the default parameter settings from openSMILE for the low level descriptors (LLDs) of each feature set, the higher level suprasegmental features were extracted at a rate of 1 s, using an overlapping window of 0.5 s.

Additionally, we extract a 4 096 dimensional feature set of deep data-representations using the Deep Spectrum toolkit [30][4]. Deep Spectrum has shown success for similar audio- and speech-based tasks [31], and extracts features from the audio data using pre-trained convolutional neural networks. For this study, we extract, viridis colour map, spectrograms using the default Deep Spectrum settings with a window size of 1 sec and a 0.5 sec overlap. The deep features are then extracted from the layer *fc7* of AlexNet [32].

[3](S0) baseline 30mins before first preparation, (S1) 1min before the first preparation, (S2) 1 after speech, (S3) 10mins after speech, (S4) 20mins after speech, (S5) 30mins after speech, (S6) 45mins after speech, and (S7) 60mins after speech.

[4]https://github.com/DeepSpectrum/DeepSpectrum

3.2. Training procedure

For our experiments, we use the epsilon-support vector regression (SVR) and a linear kernel, using the implementation from the open-source machine learning toolkit Scikit-Learn [33]. For training, the data is split into speaker independent sets: training, validation and test (cf. Table 1). During the development phase, we trained a series of SVR models, optimising the complexity parameters ($C \in 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1}, 1), evaluating their performance on the validation set. We re-trained the model with the concatenated train and validation set, and evaluate the performance on the test set. This method was repeated for each of the speaking scenarios, interview (1), arithmetic (2), and interview and arithmetic combined (1+2). Note, for consistency when comparing across the sample points (S0-S7) we report only $C = 10^{-3}$, chosen as this value acheived frequently strong test set results when evaluating ρ.

3.3. Evaluation metrics

To evaluate the results of our SVR experiments we utilise 2-core metrics, (1) *Spearman's correlation coefficient* (ρ) and (2) *Root Mean Square Error* (RMSE). Due to space limitations we do not display the development results for RMSE. Additionally, we utilise Cohen's d as a measure of effect size between the prediction of results of interest. Typically, a spearmans correlations above 0.4 is considered a moderate correlation, and a Cohen's d above 0.8 would be a large effect size [34].

4. Results and Discussion

When looking at the progression in results from S1 to S3, we can see an improvement in ρ across all feature sets and scenarios (Table 2). The highest ρ, 0.449, was achieved with eGeMAPS in scenario 2 (arithmetic) at sample point S3 (10 minutes after oral presentation). When extending this observation to S4 (20 minutes after oral presentation) we can see that ComParE features continue to improve, and so too for eGeMAPS in scenario 1 (Table 2); however, ρ drops for all other combinations at this point. From previous research into the delayed (ca. 38mins) biological response of cortisol [10], we would expect this correlation to be highest at this point (i. e. , from the start of test S3 25mins, and S4 45mins). Interestingly, from the distribution of data (Figure 1), we also observe that participants show the least amount of similarity at S3. We speculate that this may be one reason why we observe the improved correlation at this point. However this particular phenomena would require more fine-grained investigation

After the S4 sample, the results drop to a weak through to no correlation, although ComParE features retain a weak ρ, 0.149, at S6 in scenario 2 (RMSE: 0.294). This curve can be seen when plotting the data ρ scores across the sequential measurements (Figure 2). We also see the same trend although the inverse when observing the RMSE scores. Regarding, RMSE we see ComParE features achieving the best score, i. e. , the lowest, of 0.244 RMSE at scenario 1+2, S4.

The Deep Spectrum features consistently achieve the lowest ρ and highest RMSE scores across all scenarios, achieving at best ρ of 0.198 (RMSE: 0.286) at S3 scenario 1+2. The Deep Spectrum results between S0 and S3 (scenario 1+2) differ to a large extent, which is reflected by a large effect size of d = 1.25. Similarly, the best ComParE result, ρ 0.421 (RMSE: 0.245) at S4, also shows a large and consistent trend for effect size in comparison to the predictions achieved for S0 to S3 d = 1.02, and S0 to S4, d = 0.88.

Table 2: *SVR with linear kernal results for the FAU-TSST corpus cortisol prediction task. Reporting (dev)elopment and test partitions for Spearmans Correlation Coefficient (ρ), and test result for the Root Mean Square Error (RMSE). Normalised cortisol range [0, 1] nmol/L. Reporting $C = 10^{-3}$. Results are seperated for each speech (Sc)cenario (1) interview, (2) arithmetic, and (1+2) combined. Reporting all feature sets (Ft)* EGEMAPS *(EGE),* COMPARE *(COM) and* DEEP SPECTRUM *(DS), for measures S0 (baseline) to S7 (+60mins). Emphasied results discussed in Section 4.*

		S0 -30mins			S1 -1min			S2 +1mins			S3 +10mins			S4 +20mins			S5 +30mins			S6 +45mins			S7 +60mins		
		ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE	ρ	ρ	RMSE
Sc.	Ft.	dev	test	test	dev	test	test	dev	test	test	dev	test	test	dev	test	test	dev	test	test	dev	test	test	dev	test	test
1	EGE	.528	.162	.299	.050	.028	.272	.298	**.357**	.311	.044	**.315**	.273	.012	**.333**	.271	.013	.191	.268	.050	.037	.268	.050	.116	.254
	COM	.475	.043	.286	.138	.023	.270	.266	.267	.290	.088	**.363**	.263	.152	**.400**	**.250**	.211	**.363**	.251	.177	.093	.251	.179	.091	.269
	DS	.225	-.133	.306	-.037	-.156	.295	.010	.082	.293	-.095	.154	.287	-.088	.058	.293	-.138	-.038	.294	-.135	-.093	.294	-.184	-.104	.314
2	EGE	.464	.046	.309	.293	.039	.302	.044	**.365**	.281	.084	**.449**	.272	.113	**.328**	.283	.040	.277	.287	.052	.027	.287	.104	.067	.294
	COM	.399	.042	.313	.006	.021	.296	.014	.280	.264	.113	**.385**	.249	.101	**.412**	**.245**	.516	**.386**	.249	.066	.149	**.249**	.107	.085	.294
	DS	.391	.002	.301	.088	-.141	.277	-.047	.021	**.314**	-.145	.177	.287	-.198	.088	.286	-.248	.031	.282	-.262	.099	.282	-.208	-.001	.275
1+2	EGE	.516	.129	.314	.183	.090	.285	.234	**.353**	.303	.029	**.383**	.271	.029	.314	.274	.027	.177	.274	.130	.002	.274	.122	.082	.268
	COM	.455	.016	.304	.450	.035	.279	.196	.280	.282	.053	**.392**	.258	.110	**.421**	**.244**	.148	**.378**	**.249**	.094	.097	**.249**	.099	.077	.281
	DS	.390	-.043	.306	.078	-.158	.285	.002	.060	.307	.031	.198	.286	-.158	.119	.288	-.196	.027	.287	-.238	-.076	.287	-.237	-.049	.292

When considering all results, it can be observed that COMPARE is the consistently better-performing feature set for this task. Although when reporting ρ as the measure of correlation, results achieved using EGEMAPS features are not significantly different to COMPARE (e. g. , S4 EGEMAPS S4 COMPARE moderately positive relationship: Pearson r = 0.513 for s1+2 results). We speculate that the success for hand-crafted features comes from their tailored speech-based attributes i. e. , *leverage features*, e. g. , loudness, which have been discussed as being appropriate for other health-related speech tasks such as depression, or Parkinson's detection [35]. Additionally, we speculate that the weaker results of DEEP SPECTRUM features which, are extracted from spectral features only, could be due to the higher levels of ambient room noise present in the audio signal, as a consequence of the recording methods.

Finally, it would appear that there are minimal differences between the speaking scenarios. With a tendency toward better results across all feature sets and sample points, the scenario-2 (arithmetic) results are stronger, when observing ρ, μ ρ 0.151 from, S0 - S7 (μ ρ 0.123, 0.138, for scenario 1 and scenarios 1+2 respectively). However, for RMSE, a μ of 0.292 was achieved in Scenario-2 across all samples and features sets, compared with 0.281 and 0.283 for scenario 1 and scenario 1+2, respectively. This observation leaves us to speculate that perhaps there was some advantage when using scenario 2 speech features, as, from a qualitative observation of the data, the visible stress in the subject does seem to be elevated during the arithmetic task. This effect may be due to the lack of preparation time that the subject had for this task as compared to the interview. Gaining further insight in this regard would, however, require more in-depth analysis.

5. Conclusion and Future Outlook

In this study we have evaluated the ability for speech-based features to predict cortisol levels as a marker of biological stress. Findings have shown that speech-based features can achieve at best 0.244 RMSE (ρ 0.421) with COMPARE features at S4 (+20mins after speech), and in this same way correlation increases from samples taken at S0 to S3 and S4. Observing an improvement trend when utilising the features from stressed speech as cortisol samples reach the sample taken at the 10th minute (a finding which is supported by the literature in relation to delayed biological response to cortisol response [10]), this finding is consistent across all features sets, suggesting that

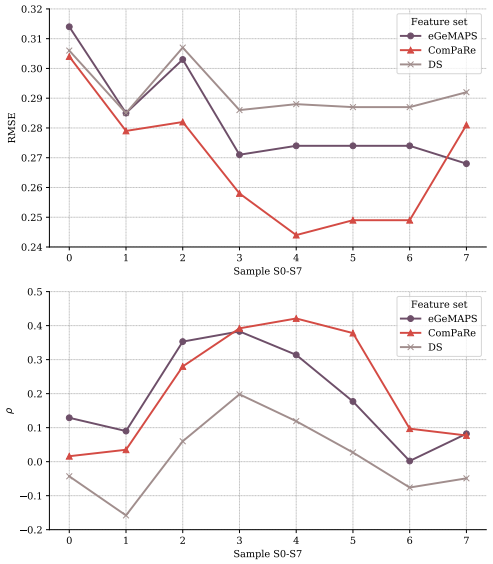

Figure 2: *Result for Root Mean Square Error (RMSE) (above) and Spearmans Correlation Coefficient (ρ) (below) on test for $C = 10^{-3}$, of scenario 1+2, across each feature sets* EGEMAPS , COMPARE *and* DEEP SPECTRUM *(DS), for measures S0 (baseline) to S7 (+60mins).*

an acoustic-based approach is suitable for the prediction of cortisol as a marker of stress.

With speech features from individuals in a stressed state showing promise for the task of cortisol prediction, particularly hand-crafted sets, future work would include feature selection, as a means of taking a closer look at the particular speech based features from such sets which are responsible for this. Additionally, given the varied speaking scenarios it may be of interest to evaluate the word-use within the FAU-TSST corpus more closely. From the analysis approach, it would also be an advantage to implement state of the art deep architectures, utilising a multimodal fusion approach, given the extensive modalities which are made available through the FAU-TSST data.

6. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

7. References

[1] S. J. Motowidlo, J. S. Packard, and M. R. Manning, "Occupational Stress: Its Causes and Consequences for Job Performance," *Journal of Applied Psychology*, vol. 71, no. 4, p. 618, 1986.

[2] M. Kivimäki, J. Head, J. E. Ferrie, M. J. Shipley, E. Brunner, J. Vahtera, and M. G. Marmot, "Work stress, weight gain and weight loss: evidence for bidirectional effects of job strain on body mass index in the Whitehall II study," *Journal of Obesity*, vol. 30, no. 6, p. 982, 2006.

[3] R. M. Sapolsky, L. M. Romero, and A. U. Munck, "How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions," *Endocrine reviews*, vol. 21, no. 1, pp. 55–89, 2000.

[4] R. M. Nesse and E. A. Young, "Evolutionary origins and functions of the stress response," *Encyclopedia of Stress*, vol. 2, pp. 79–84, 2000.

[5] P. A. Nepomnaschy, E. Sheiner, G. Mastorakos, and P. C. Arck, "Stress, Immune Function, and Women's Reproduction," *Annals of the New York Academy of Sciences*, vol. 1113, no. 1, pp. 350–364, 2007.

[6] M. A. Ellenbogen, A. E. Schwartzman, J. Stewart, and C.-D. Walker, "Stress and selective attention: The interplay of mood, cortisol levels, and emotional information processing," *Journal of Psychophysiology*, vol. 39, no. 6, pp. 723–732, 2002.

[7] R. F. Vining, R. A. McGinley, J. J. Maksvytis, and K. Y. Ho, "Salivary Cortisol: A Better Measure of Adrenal Cortical Function than Serum Cortisol," *Annals of Clinical Biochemistry*, vol. 20, no. 6, pp. 329–335, 1983.

[8] U. M. Nater, N. Rohleder, J. Gaab, S. Berger, A. Jud, C. Kirschbaum, and U. Ehlert, "Human salivary alpha-amylase reactivity in a psychosocial stress paradigm," *Journal of Psychophysiology*, vol. 55, no. 3, pp. 333–342, 2005.

[9] H. Steiner and S. Levine, "Acute stress response in anorexia nervosa a pilot study," *Child Psychiatry and Human Development*, vol. 18, no. 4, pp. 208–218, 1988.

[10] W. K. Goodman, J. Janson, and J. M. Wolf, "Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test," *Journal of Psychoneuroendocrinology*, vol. 80, pp. 26–35, 2017.

[11] R. Miller and C. Kirschbaum, "Trier Social Stress Test," *Journal of Encyclopedia of Behavioral Medicine*, pp. 2005–2008, 2013.

[12] S. S. Dickerson and M. E. Kemeny, "Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research," *Journal Psychological Bulletin*, vol. 130, no. 3, p. 355, 2004.

[13] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The Trier Social Stress Test–a tool for investigating psychobiological stress responses in a laboratory setting," *Journal of Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[14] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proc. ICME*. Hong Kong, P. R. China: IEEE, 2017, pp. 985–990.

[15] B. Schuller, F. Friedmann, and F. Eyben, "The Munich BioVoice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production," in *Proc. LREC*. Reykjavik, Iceland: ELRA, 2014, pp. 1506–1510.

[16] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Proc. INTERSPEECH*. Singapore, Singapore: ISCA, 2014, pp. 427–431.

[17] N. Cummins, A. Baird, and B. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and Opportunities," *Methods, Special Issue on on Translational data analytics and health informatics*, vol. 151, pp. 41–54, 2018.

[18] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Formant frequencies under cognitive load: Effects and classification," *Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 219253, 2011.

[19] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Proc. Eurospeech*, 1997, pp. 1743–1746.

[20] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, "Depression and cortisol responses to psychological stress: A meta-analysis," *Psychoneuroendocrinology*, vol. 30, no. 9, pp. 846–856, 2005.

[21] M. Hautzinger and M. Bailer, *ADS-Allgemeine Depressionsskala*. Weinheim, Germany: Beltz, 2003.

[22] N. Rohleder and U. M. Nater, "Determinants of salivary α-amylase in humans and methodological considerations," *Psychoneuroendocrinology*, vol. 34, no. 4, pp. 469–485, 2009.

[23] F. Eyben, F. Weninger, F. Gross *et al.*, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. 21st ACM Int. Conf. Multimedia, MM 2013*. Barcelona, Spain: ACM, Oct 2013, pp. 835–838.

[24] Y. Zhang and B. Schuller, "Towards Human-Like Holisitc Machine Perception of Speaker States and Traits," in *Proc. MI20-HLC*. Windsor, U. K.: Springer, 2016, 3 pages.

[25] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic Classification of Autistic Child Vocalisations: A Novel Database and Results," in *Proc. INTERSPEECH*. Stokholm, Sweden: ISCA, 2017, pp. 849–853.

[26] S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller, "Is deception emotional? an emotion-driven predictive approach," in *Proc. of INTERSPEECH 2016*. San Francisco, CA: ISCA, September 2016, pp. 2011–2015.

[27] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[28] G. Keren and B. Schuller, "Convolutional RNN: an enhanced model for extracting features from sequential data," in *Proc. of 2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, 2016, pp. 3412–3419.

[29] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. ACM MM*. Mountain View, CA: ACM, 2017, pp. 478–484.

[30] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. of INTERSPEECH 2017*. Stockholm, Sweden: ISCA, August 2017, pp. 3512–3516.

[31] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proc. of the 31st International Joint Conference on Neural Networks (IJCNN)*. Rio de Janeiro, Brazil: IEEE, July 2018, pp. 2419–2425.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] L. E. Kazis, J. J. Anderson, and R. F. Meenan, "Effect Sizes for Interpreting Changes in Health Status," *Journal Medical care*, pp. 178–189, 1989.

[35] A. Batliner and B. Möbius, "Prosody in Automatic Speech Processing," in *The Oxford Handbook of Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford, UK: Oxford University Press, 2019, ch. 42, p. 20 pages.