VoStress – Voice-based Detection of Acute Psychosocial Stress

Marie Oesten[1*], Robert Richer[1], Luca Abel[1], Nicolas Rohleder[2], and Bjoern M. Eskofier[1]

***Abstract*—Current stress assessment methods include self-reports and biomarkers which are evaluated in often complex, laboratory procedures. Due to that investigating new indicators for acute stress is crucial for the development of automatic stress detection systems. A promising extension might be provided by investigating speech, which has been shown to be affected by negative emotions and threatening situations. For that reason, we extracted verbal acoustics from audio data collected during a study where N=21 participants underwent the Trier Social Stress Test (TSST), the gold standard for laboratory stress induction, and a stress-free control condition (friendly-TSST) while concurrently collecting cortisol via saliva samples to assess the biological response to stress. Our results show that acute stress leads to significant ($p < 0.05$) alterations of acoustic features. A stepwise backward multiple linear regression model explained 58.8 % of the variance of the maximum cortisol increase. In addition to that, we performed classification experiments that distinguished stress from non-stress situations with an accuracy of 80.0 ± 12.7 %. While further research is needed to validate our approach, we are convinced that the information extracted from speech can be a valuable indicator for automatic stress detection systems and can even predict the biological response to stress situations.**

***Index Terms*—Acute stress, speech, acoustic features, machine learning, cortisol**

I. Introduction

Psychosocial stress is a prevalent phenomenon that we encounter in our everyday lives dealing with work, personal relationships, or daily life demands [1]. It encloses the psychological and social challenges that arise in demanding situations, often leading to physiological and psychological responses [2]. One of the key responses to this form of stress involves activation of the hypothalamic-pituitary-adrenal (HPA) axis, which triggers the release of the stress hormone cortisol [3]. While this reaction is considered crucial for preparing the body to adequately handle the upcoming stressful situation, a maladaptive HPA axis reaction can, if not effectively managed, transition into chronic stress, leading to detrimental changes in a person's physical and mental health [4].

To investigate acute psychosocial stress and measure the physiological responses, researchers have developed standardized laboratory stress induction procedures. Among these, the Trier Social Stress Test (TSST), initially proposed by Kirschbaum et al. [5], has emerged as the gold standard for acute psychosocial stress induction as it reliably activates the HPA axis [2]. The body's stress response is typically measured using self-reports for assessing the subjective stress experience, saliva- or blood-based samples to extract neuroendocrine markers like cortisol or alpha-amylase, or electrophysiological data like electrocardiography (ECG) or electrodermal activity (EDA). With stress being a risk factor for human health [4], the demand for automatic, unobtrusive, possibly even contactless stress markers to prevent the transition to chronic stress is high. Since the evaluation of self-reports and saliva samples are not feasible in a large-scale, real-world setting, other digital biomarkers are required to be investigated.

Speech as a stress detection modality has been widely researched, revealing observable changes like an increasing fundamental frequency in response to stress exposure [6]. Leveraging similar effects, Baird et al. examined the effectiveness of speech-based features in predicting sequential cortisol measurements [7]. Their results show a moderate correlation between speech and cortisol data in the TSST. In a follow-up work, they predicted physiological parameters such as heart rate and respiration using the same feature set and a deep learning-based architecture [8]. Norden et al. explored the impact of different stress definitions for stress detection, observing better results for external annotations than for cortisol stress levels [9]. However, one drawback of all previous approaches is that their findings are only based on acoustic data collected during the TSST, without having a comparable stress-free control condition. Thus, their findings are limited to the efficacy of verbal acoustics for the prediction of sequential physiological responses. To determine whether verbal acoustics can be used as potential biomarkers to distinguish between stressed and non-stressed conditions, it is necessary to evaluate participants in the same experimental setting with and without the presence of acute stress.

For that reason, we present an approach for detecting acute psychosocial stress with verbal acoustics based on data from participants that were exposed to acute psychosocial stress via the TSST, and the friendly-TSST, a stress-free control condition [10], on two consecutive days. To the best of our knowledge, our work is the first to assess the efficacy of vocal acoustics for stress detection in a within-subjects design.

II. Methods

A. Data Acquisition

The data used in this work was collected as part of an experiment to examine the influence of acute psychosocial stress on body posture and movements [11]. For this study, we recruited N=21 young healthy individuals (85.7 % women) aged 22.6 ± 4.0 years. The exclusion criteria for study participation are in line with the recommended guidelines for assessing HPA axis activity [12] and explained in detail in our previous work [11]. The study was approved by FAU's

*Responsible author; Contact: marie.oesten@fau.de

[1]Machine Learning and Data Analytics Lab (MaD Lab), Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany;

[2]Chair of Health Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054 Erlangen, Germany

ethics committee (protocol #493_20 B) and was conducted in compliance with relevant guidelines and regulations. All participants gave written consent before starting the experiment.

Each participant underwent the procedure shown in Figure 1. The Trier Social Stress Test (TSST) [5] and the friendly Trier Social Stress Test (f-TSST) [10] were conducted in randomized order on two consecutive days. Both conditions were carried out in a quiet, separate room at similar times to control for the diurnal cortisol rhythm [13].

The TSST was conducted following the protocol initially proposed by Kirschbaum et al. [5]. The TSST is designed as a fictional job interview with social-evaluative elements and consists of three different phases (*Preparation*, *Interview*, *Mental Arithmetics*) lasting 5 min each. All tasks are performed in front of an evaluation panel wearing lab coats. The panel was trained to maintain complete neutrality, minimize emotions, and not engage in any interaction with the participants. As a stress-free control condition that does not activate the HPA axis, we used the f-TSST as proposed by Wiemers et al. [10]. The f-TSST has a structure similar to the TSST, but all social-evaluative elements were removed. Additionally, it did not contain a *Mental Arithmetics* phase but only consisted of a 10 min *Interview* phase.

Before and after (f-)TSST, we collected six saliva samples (S0-S5) per day at defined time points to assess HPA axis activity from salivary cortisol (Figure 1). In addition, participants were video-recorded during the *Interview* and/or *Mental Arithmetic* phases using a camera placed in front of them.

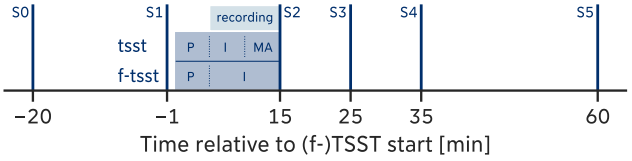

Fig. 1: Study protocol of (f-)TSST; P: *Preparation*; I: *Interview*; MA: *Mental Arithmetic*

B. Data Processing

1) Cortisol Measures: To assess HPA axis activity, we extracted raw cortisol concentrations using a chemiluminescence immunoassay (CLIA, IBL, Hamburg, Germany) as described in more detail in previous publications (e.g. [14]). From the raw cortisol levels, we computed the maximum increase (Δc_{max}) as response to the (f-)TSST between S1 and S2-S5 as a measure for HPA axis reactivity.

2) Audio Processing: To process audio data, we extracted the audio channels from the video recordings using *ffmpeg*[1] and stored them as *.wav* files. As the audio data recorded during the (f-)TSST contained conversation shares of both panel and participant, we first segmented the audio data into the individual speech segments by applying a speaker diarization algorithm implemented in the *pyannote.audio* Python package [15]. Based on this output, we discarded all speech segments shorter than 0.3 s. Afterwards, we identified the participant as the speaker with the highest conversation share, while the remaining identified speakers were considered as the panel. If multiple speakers were detected simultaneously, regardless of whether the participant was speaking, we discarded this segment to avoid overlap. Subsequently, we concatenated consecutive segments of either the participant or panel speaking. This resulted in a segmentation where parts of the panel speaking can be distinguished from parts that contain the participant's verbal acoustics or periods of silence before, between, or after the speech.

3) Feature Extraction: For feature extraction, we used the *OpenDBM* v2.0 library that allows the computation of digital acoustic biomarkers[2]. *OpenDBM* extracts a set of *raw variables* using Parselmouth [16], such as audio intensity (I), fundamental frequency (F_0) and formant frequencies ($F_{1...4}$), vocal jitter and shimmer, and mel frequency cepstral coefficients ($MFCC_{1...12}$). They serve as the basis for computing *derived variables*, which aggregate the time-series raw variables by computing basic signal characteristics. Since the derived variables were computed over the whole audio input and, therefore, also included parts where the panel was speaking, we re-implemented the derived variable extraction methods to compute the derived variables for each speech segment of participants individually. We then aggregated the derived variables per segment by computing mean (μ) and standard deviation (σ) as well as mean and standard deviation weighted by the segment lengths ($\hat{\mu}$ and $\hat{\sigma}$, respectively). An overview of extracted features is provided in the OpenDBM documentation[3].

C. Evaluation

To investigate the association between acute psychosocial stress and changes in verbal acoustics, we only considered features computed over the *Interview* phases of the (f-)TSST to allow better comparability between both conditions.

1) Statistical Analysis: We performed non-parametric Wilcoxon signed-rank tests on all extracted audio features with *condition* as between-variable since all participants were exposed to both TSST and f-TSST and applied Bonferroni corrections across all tests to control for multiple comparisons. We set a significance level of $\alpha = 0.05$ and report effect sizes as Hedge's g. In all Figures and Tables, statistical significance was denoted using the following notation: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

2) Stepwise Backward Multiple Linear Regression: To investigate the association between cortisol and verbal acoustics we employed a stepwise backward multiple linear regression (SBMLR) approach by computing the differences of acoustic features and Δc_{max} between TSST and f-TSST to quantify the stress-induced voice alteration. For the features, we only used statistically significant features from the previous step as input

[1] https://ffmpeg.org/

[2] https://github.com/AiCure/open_dbm

[3] https://aicure.github.io/open_dbm/docs/verbal-acoustic

and standardized them via z-score normalization. SBMLR was performed iteratively with acoustic features as predictors and Δc_{max} as dependent variable. In each iteration, the predictor with the highest p-value was removed until all p-values were below α. The final selection of the best-fitting model was based on the highest adjusted R^2 value.

3) Classification Experiments: To assess the possibility of detecting acute stress situations based on acoustic features, we conducted a series of classification experiments, evaluating different combinations of pre-processing, feature selection, and classification algorithms. After removing all features with zero variance in the pre-processing, we scaled features with a Standard Scaler or Min-Max Scaler. In the feature selection step, we applied either Select-k-Best (SkB) or Recursive Feature Elimination (RFE). As classifiers, we evaluated Naïve Bayes (NB), k-Nearest-Neigbors (kNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perception (MLP), and AdaBoost (Ada). All different pipeline combinations were evaluated using a five-fold nested cross-validation (CV). The outer CV was used for model evaluation and assessing the generalizability of our approach, the inner CV was used for hyperparameter optimization using randomized-search with 40 000 iterations (RF) or grid-search (all others). Each CV split was performed on a participant level, ensuring that all data from a participant was either in the training or the test set. In the inner CV, we retrained the classification pipelines with the best-performing hyperparameters and made predictions on the unseen test set of the model evaluation CV folds. We used the accuracy as metric for optimization and additionally computed the F1-score.

All evaluation experiments were conducted using the Python libraries *BioPsyKit* (v0.9.0) [17], *tpcp* (v0.13.0) [18], and *scikit-learn* (v1.2.2) [19].

III. Results

We excluded data from one participant due to failed speaker segmentation, resulting in a final sample size of $N = 20$ participants. In our previous work, we demonstrated that the TSST successfully induced acute psychosocial stress while the f-TSST did not, indicated by a significantly higher cortisol increase Δc_{max} as response to the TSST, $W = 26.0, p = 0.004, g = 0.746$ [11].

A. Statistical Analysis

After statistical testing, nine out of 48 acoustic features showed significant differences between TSST and f-TSST (Table I, Figure 2). Acute psychosocial stress leads to increased audio intensity variation $\sigma(I)$. Additionally, all four formants $\mu(F_{1...4})$ increased, on average, during the TSST. In contrast, $\mu(MFCC_1)$ and μ(Shimmer) decreased during the TSST.

B. Stepwise Backward Multiple Linear Regression

Before fitting the SBMLR model, $\hat{\mu}$(Shimmer) was omitted due to high multicollinearity (> 0.8). Based on the resulting eight features, the best-performing regression model based on five features explained 58.8 % of the variance (indicated by adj. R^2) in cortisol increase Δc_{max} (Table II).

TABLE I: Results of statistical analysis showing all features with significant differences between (f-)TSST

Feature	W	p	Hedges' g
$\sigma(I)$	21	0.041*	–0.893
$\mu(F_1)$	14	0.010*	–0.740
$\sigma(F_1)$	2	<0.001***	–0.783
$\mu(F_2)$	0	<0.001***	–1.277
$\mu(F_3)$	10	0.004**	–0.708
$\mu(F_4)$	9	0.003**	–0.552
$\mu(MFCC_1)$	21	0.041*	0.985
$\mu(Shimmer)$	20	0.034*	0.927
$\hat{\mu}(Shimmer)$	19	0.028*	0.877

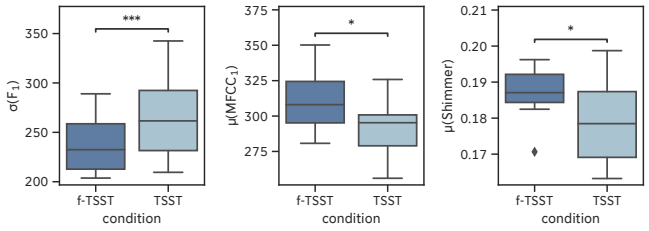

Fig. 2: Selection of features showing significant differences between (f-)tsst

TABLE II: Results of linear regression predicting Δc_{max}; β: standardized regression coefficient; σ: standard error; adj: adjusted

Feature	β	σ	t	p	R^2	adj. R^2
$\mu(F_1)$	8.228	1.564	5.263	<0.001	0.697	0.588
$\sigma(F_1)$	–1.842	1.094	–1.683	0.114	0.697	0.588
$\mu(F_4)$	–5.717	1.533	–3.729	0.002	0.697	0.588
$\mu(MFCC_1)$	6.521	1.363	4.785	<0.001	0.697	0.588
$\mu(Shimmer)$	–3.277	1.026	–3.195	0.006	0.697	0.588

C. Classification Experiments

The highest accuracy of 80.0 ± 12.7 % across the 5-fold model evaluation CV was achieved by the classification pipeline comprising Standard Scaler, SkB, and DT while the pipeline comprising Standard Scaler, RFE, and NB achieved the highest F1-score (Table III). Splitting the predictions by the condition order, i.e., whether the TSST or the f-TSST was performed on the first day, revealed that most misclassifications were performed when TSST was the first condition, with the TSST wrongly being classified as f-TSST (Figure 3).

IV. Discussion

The main objective of our work was to investigate the feasibility of identifying situations of acute psychosocial stress from speech information recorded during the gold standard protocol for laboratory stress induction (TSST) and a stress-free control condition (f-TSST).

In line with the findings by Baird et al. [8] who predicted sequential cortisol levels from acoustic information, we showed that stress-induced changes in verbal acoustics can predict the differences in cortisol increase between f-TSST and TSST. Our best-fitted SBMLR model was able to significantly explain the

TABLE III: Classification performance metrics (M ± SD) over the 5-fold model evaluation CV. For each classifier, the pipeline combination with the highest mean accuracy is shown. The pipelines with the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]
Standard	SkB	DT	**80.0 ± 12.7**	74.1 ± 20.5
Standard	RFE	NB	80.0 ± 17.0	**79.9 ± 17.0**
Standard	RFE	MLP	77.5 ± 9.4	74.7 ± 11.8
Standard	RFE	RF	77.5 ± 12.2	74.9 ± 14.6
Min-Max	RFE	Ada	75.0 ± 7.9	73.1 ± 16.7
Standard	RFE	kNN	75.0 ± 7.9	72.9 ± 10.1
Min-Max	SkB	SVM	72.5 ± 16.6	71.8 ± 15.4

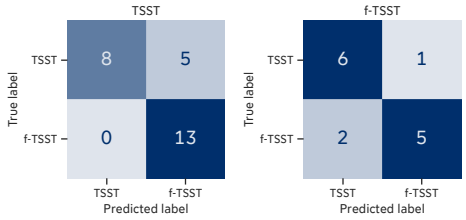

Fig. 3: Confusion matrices for the classification pipeline with the highest mean accuracy over the 5-fold model evaluation CV, split by condition order.

variance of the cortisol response. Similarly, the results of our classification experiments support the feasibility of leveraging vocal acoustics for detecting situations of acute stress. Despite a classification accuracy of 80.0 ± 12.7 % using a traditional machine learning approach with a rather simple decision tree as classification algorithm, the large number of TSST misclassifications in the group of participants performing the TSST on the Day 1 needs further investigation. A potential explanation could be that individuals performing the f-TSST on Day 2 were ruminating about the stressful situation on the previous day, resulting in vocal changes that made a classification between TSST and f-TSST more difficult.

Although both the results from our regression and machine learning approach are promising, the limited sample size of our study requires future work to assess the generalizability of our findings on a larger and more balanced dataset. Additionally, we plan to extend our analysis by integrating other digital biomarkers, such as facial activity [20], or speech semantics. Since acute stress detection from acoustic features alone yields promising results, we are convinced that also our approach could benefit from an extension by digital biomarkers of other modalities.

V. Conclusion and Outlook

In our work, we investigated the feasibility of acute stress detection from verbal acoustics. By comparing acoustic features extracted from a stress and non-stress condition with statistical measures, we demonstrated the potential for predicting the biological response to acute stress and to differentiate stress from non-stress situations. While our results confirmed the findings of previous work that acute stress alters the acoustics of speech and revealed the possibility to – at least partly – predict the cortisol response, they also highlight the potential of verbal acoustics being a reliable indicator for detecting acute psychosocial stress. Given its widespread accessibility, it holds great promise as a valuable component in automated, unobtrusive stress detection systems.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – Project-ID 442419336, EmpkinS.

References

[1] L. Kogler *et al.*, "Psychosocial versus physiological stress — Meta-analyses on deactivations and activations of the neural correlates of stress reactions," *NeuroImage*, vol. 119, pp. 235–251, Oct. 2015.

[2] S. S. Dickerson and M. E. Kemeny, "Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research," *Psychological Bulletin*, vol. 130, no. 3, pp. 355–391, 2004.

[3] C. Tsigos and G. P. Chrousos, "Hypothalamic–pituitary–adrenal axis, neuroendocrine factors and stress," *Journal of Psychosomatic Research*, vol. 53, no. 4, pp. 865–871, Oct. 2002.

[4] Jennifer Couzin-Frankel, "Inflammation bares a dark side," *Science*, vol. 330, no. 6011, pp. 1621–1621, 2010.

[5] C. Kirschbaum *et al.*, "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting," in *Neuropsychobiology*, vol. 28, 1993, pp. 76–81.

[6] C. Kirchhübel *et al.*, "Acoustic Correlates of Speech when Under Stress: Research, Methods and Future Directions," *IJSLL*, vol. 18, no. 1, pp. 75–98, Sep. 2011.

[7] A. Baird *et al.*, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 534–538.

[8] A. Baird *et al.*, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Front. Comput. Sci.*, vol. 3, p. 750284, Dec. 2021.

[9] M. Norden *et al.*, "Automatic Detection of Subjective, Annotated and Physiological Stress Responses from Video Data," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Nara, Japan: IEEE, Oct. 2022, pp. 1–8.

[10] U. S. Wiemers *et al.*, "A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women," *Stress*, vol. 16, no. 2, pp. 254–260, Mar. 2013.

[11] R. Richer *et al.*, "StressPose – Detection of Acute Psychosocial Stress from Body Posture and Movements."

[12] I. Labuschagne *et al.*, "An introductory guide to conducting the Trier Social Stress Test," *Neuroscience & Biobehavioral Reviews*, vol. 107, pp. 686–695, Dec. 2019.

[13] J. M. Smyth *et al.*, "Individual differences in the diurnal cycle of cortisol," *Psychoneuroendocrinology*, vol. 22, no. 2, pp. 89–105, 1997.

[14] J. Janson and N. Rohleder, "Distraction coping predicts better cortisol recovery after acute psychosocial stress," *Biological Psychology*, vol. 128, pp. 117–124, 2017.

[15] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, 2021.

[16] Y. Jadoul *et al.*, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[17] R. Richer *et al.*, "BioPsyKit: A Python package for the analysis of biopsychological data," *JOSS*, vol. 6, no. 66, p. 3702, Oct. 2021.

[18] A. Küderle *et al.*, "Tpcp: Tiny Pipelines for Complex Problems - A set offramework independent helpers for algorithms development andevaluation," *JOSS*, vol. 8, no. 82, p. 4953, Feb. 2023.

[19] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.

[20] J. Almeida and F. Rodrigues, "Facial Expression Recognition System for Stress Detection with Deep Learning:," in *Proceedings of the 23rd International Conference on Enterprise Information Systems*. SCITEPRESS - Science and Technology Publications, 2021, pp. 256–263.