Interspeech 2026
27 September - 1 October 2026, Sydney, Australia

Automatic Detection of Stress from Speech in the Trier Social Stress Test

Hanna Drimalla [1,**], *Wieland R. Cremer* [1,*], *Christine Kraus* [1,*], *Oliver T. Wolf* [2]

[1] Human-Centered Artificial Intelligence Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany.
[2] Department of Cognitive Psychology, Faculty of Psychology, Ruhr University Bochum, Bochum, Germany.

drimalla@uni-bielefeld.de

Abstract

Automatically detecting stress in speech provides an unobtrusive way to gain insights relevant to behavioral research or clinical assessment. This study investigates the automatic differentiation between a stressful and non-stressful situation, and the prediction of physiological and affective stress responses. Speech data was collected from 50 participants who either completed the Trier Social Stress Test (TSST) or a non-stressful control condition. With a processing pipeline that included speaker diarization and machine learning models, we achieved stress detection performance significantly above a mean baseline. Moreover, relevant physiological and affective stress responses were partially predictable from acoustic-prosodic features. Feature-importance analyses identified the most informative predictors contributing to model performance. The findings demonstrate that speech can serve as a meaningful and unobtrusive indicator of multiple dimensions of the human stress response.

Index Terms: stress, machine learning, speech, voice, cortisol

1. Introduction

In research and clinical practice, stress is commonly measured using self-report instruments and physiological biomarkers such as salivary cortisol and salivary alpha-amylase (sAA). While these measures provide important reference points, they are sensitive to procedural and contextual factors and can be difficult to collect unobtrusively, frequently and at scale in a standardized manner [1, 2]. These constraints highlight the need for alternative measures that still relate to well-established physiological indices of stress. Speech-based markers represent a particularly promising candidate in this regard. Stress has been linked to systematic modulation of speech production and shown to leave measurable traces in both prosody and voice quality [3, 4, 5]. Across various stress-inducing settings, speakers show subtle yet reliable acoustic and temporal changes, including shifts in fundamental frequency (F0) [6, 7], intensity [8], as well as differences in speaking rate or pausing behavior [7, 6, 9]. Moreover, speech can be recorded unobtrusively and repeatedly in both remote and real-world contexts.

Whether speech can serve as a potential biomarker of stress has been investigated via machine learning (ML) on datasets encompassing a broad range of stressors, from speaking in a foreign language [10] and acted stress [11], to experimentally induced stress in laboratory settings [12, 13]. The variety of stressors and the often unclear presence and intensity of stress may have contributed to inconsistent results [3]. Standardized laboratory protocols offer a controlled and well-validated way to experimentally induce stress and link resulting vocal changes to established physiological stress markers. Prior work [14, 15] has investigated such speech-based detection of stress responses using the Trier Social Stress Test (TSST) [16], which is widely considered the "gold-standard" among standardized psychosocial stress-inducing paradigms [17]: it reliably elicits acute psychosocial stress through socio-evaluative threat by a speech task and a mental arithmetic task in front of a fake committee. Baird et al. [14] reported moderate associations between acoustic speech representations and time-varying physiological stress responses measured via salivary cortisol following participation in a TSST. However, to answer the question whether vocal acoustics can be used to distinguish between stressed and non-stressed speech, a control condition is needed. The friendly-TSST (f-TSST) [18] was introduced as such an analogue, preserving the overall structure of the TSST while removing key stress-inducing elements of social-evaluative threat. Recent work incorporating the f-TSST has provided first evidence that acoustic cues can distinguish stress and control conditions [19]. However, the authors noted an important limitation of their within-subject design: the f-TSST may have elicited mild stress responses when the TSST was conducted prior, as reflected in the difficulty of classifying f-TSST samples recorded after the TSST.

In this paper, we therefore aim to evaluate speech-based automated detection of acute psychosocial stress in a fully between-subject setting using a newly collected dataset with both the TSST and the f-TSST as a non-stressed control condition. We investigate whether automatically extracted acoustic features can (i) discriminate TSST from f-TSST speech under cross-participant evaluation, and (ii) predict stress responses indexed by changes in salivary cortisol, sAA and self-reported affect. Further, we explore the most important features contributing to automatic speech-based detection of stress. By combining a matched control condition with affective and physiological outcomes, our work provides further assessment of whether stress-induced speech cues can link to acute stress responses.

2. Methods

2.1. Data Collection

Data was collected from healthy German-speaking university students participating in a laboratory stress-induction experiment. The study was approved by the local ethics committee of the Faculty of Psychology of Ruhr University Bochum and the Declaration of Helsinki was followed. Exclusion criteria comprised prior TSST experience; night-shift work; relevant illness, medication use, medical or psychotherapeutic treatment; smoking; substance abuse; and exercising, eating or drinking

[*] These authors contributed equally.
[**] indicates the corresponding author.

10.21437/Interspeech.2026-671

before testing (see superordinate study [20] for details). Inclusion criteria included a BMI between 19 and 28 and, for female participants, the use of monophasic oral contraceptives during pill intake to reduce variability in endocrine stress responses related to menstrual-cycle phase. Participants provided informed consent and were randomly assigned to either the stress (TSST) or control (f-TSST) condition. In the *stress condition*, a modified version [21] of the TSST [16] was used. Following a 5 min preparation phase, participants delivered 10 min of free speech framed as a simulated job interview in front of a reserved two-person committee (one male, one female) while being videotaped. The committee did not provide verbal feedback and prolonged pauses were tolerated. In the *control condition* (f-TSST [18]), the same structure was applied without the stress-inducing socio-evaluative threat elements: participants were informed about being in the control condition and could choose from a set of topics; the committee was introduced as laboratory employees to have a friendly conversation with; and sessions were explicitly not videotaped. Unlike in the TSST, the committee engaged by asking follow-up questions. At baseline (−1 min), participants completed the Positive and Negative Affect Schedule (PANAS) [22] and provided a saliva sample. During the preparation phase and the speech task, audio was recorded via eye-tracking glasses. After the task, the second saliva sample and PANAS were collected (1 min). A third saliva sample was obtained at 20 min post-manipulation.

2.2. Stress measures

2.2.1. Salivary biomarkers

Saliva samples were analyzed for salivary cortisol, reflecting hypothalamic–pituitary–adrenal (HPA) axis activity [23], and sAA, indicating sympathetic nervous system arousal [24]. Saliva was collected using Salivettes and stored at −18 °C until analysis. Salivary cortisol concentrations (nmol/L) were quantified using a Dissociation-Enhanced Lanthanide Fluorescent Immunoassay (DELFIA [25]; detection limit 0.5 nmol/L). sAA activity (U/L) was assessed via a colorimetric test using the CNP-G3 substrate reagent [26, 27]. Biomarker reactivity indices were computed as the maximum post-task value minus the baseline value [28, 29], resulting in values for cortisol reactivity and sAA reactivity as ground truth variables.

2.2.2. Self-reported affect

PANAS [22] is a validated 20-item measure of positive and negative affect. Participants indicate the intensity of 10 positive and 10 negative emotions on a 5-point Likert scale (1 = 'very slightly or not at all', 5 = 'extremely'). Positive (PA) and negative affect (NA) scores were obtained for pre- and post-manipulation. Change scores (ΔPA and ΔNA; post − pre) served as ground truth variables.

2.3. Audio data

Audio was recorded via the built-in microphone of SMI Eye Tracking Glasses 2.0 (SensoMotoric Instruments GmbH, Teltow, Germany) as uncompressed 16 kHz *.wav* files. Recordings started at the onset of the preparation phase and ended shortly after task completion. To reduce irrelevant noise (e.g., experimenter interaction), all raw audio files (TSST: 16.62 ± 0.32 min; f-TSST: 16.72 ± 0.34 min) were trimmed to 9 min segments starting at minute 7. For isolating participant speech, speech diarization was performed automatically using *Sortformer* [30], a pretrained transformer encoder-based end-to-end speaker diarization model by NVIDIA NeMo Speech AI. The participant was identified as the speaker with the longest total speaking time. Participant-only audio was generated by retaining their diarized speech segments, removing overlaps (50 ms collar) and concatenating the segments into a single waveform per recording without pauses longer than those occurring in natural speech (mean lengths of recordings after processing: TSST: 4.13 ± 1.85 min; f-TSST: 6.51 ± 0.98 min). A random subset ($n = 12$) of the pre-processed recordings was manually inspected to assess diarization quality, ensure the absence of non-participant speech and compared to diarization using *pyannote* [31, 32]. Noise accounted for less than 5% of the total recording time for each participant. Acoustic features were then extracted from the participant speech using three complementary toolchains: *librosa* (v0.11.0) [33] was used to extract 40 Mel-Frequency Cepstral Coefficients (MFCCs). Audio was resampled to 22.05 kHz and MFCCs were computed frame-wise and then averaged. Furthermore, 15 classical voice parameters (e.g., mean/SD F0, HNR, median pitch, jitter, shimmer) were extracted using *Praat* (v6.1.38) [34] via Parselmouth (v0.4.7) [35]. Additionally, the eGeMAPSv02 feature set [36] was extracted using *openSMILE* (v2.6.0) [37], containing 88 statistical functionals summarizing pitch, energy, spectral and other voice-quality measures. For each participant, feature sets were concatenated into a single participant-level vector with sex added as a covariate to account for related differences in the voice, resulting in a 144-dimensional feature vector per participant. Within each cross-validation split, each feature was z-standardized across participants in the training fold. The same z-standardization was subsequently applied to the corresponding hold-out participant-level feature vectors.

2.4. Machine learning models and evaluation

All models were trained on the participant-level feature vectors obtained from the audio recordings. For comparison, models were additionally trained on reduced-dimensional representations obtained via principal component analysis (PCA). The code for preprocessing, ML analysis and evaluation as well as additional figures are publicly available on GitHub (`https://github.com/mbp-lab/tsst-speech-stress`).

2.4.1. Classification

The objective of the binary classification task was to distinguish between participants who underwent the TSST and those who underwent the f-TSST. Four classification algorithms, covering a range of model complexities from linear to nonlinear ensemble methods, were trained: logistic regression (LR) [38], support vector machine (SVM) [39], random forest (RF) classifier [40] and XGBoost (XGB) classifier [41]. The LR model was tuned for the regularization coefficient $\lambda \in \{0.1, 1, 2, 10, 100\}$ and penalty type (L1 or L2). The SVM was optimized with respect to the kernel function (linear or RBF), the regularization coefficient $\lambda \in \{0.01, 0.1, 1, 10\}$ and for the RBF kernel, the kernel coefficient $\gamma \in \{\texttt{scale}, 0.001, 0.01, 0.1, 1, 10\}$. For RF, the number of trees was fixed at 1000, while maximum tree depth $\{1, 2, 4, 8\}$ and minimum number of samples per split $\{1, 2, 4\}$ were tuned. For XGB, the number of trees $\{50, 100, 150\}$, maximum tree depth $\{1, 2, 4, 8\}$ and learning rate $\{0.03, 0.1, 0.2\}$ were optimized.

2.4.2. Regression

Regression analyses were conducted to predict different stress responses from speech-derived acoustic features. Ground-truth variables included cortisol and sAA reactivity and changes in positive and negative affect: for each, separate models were calculated. Three regression algorithms were trained: support vector regression (SVR) [42], a random forest regression (RFR) [40] and an XGB regressor. The SVR was tuned over the kernel function (linear or RBF), the regularization coefficient $\lambda \in \{0.01, 0.1, 1, 10\}$ and for the RBF kernel, the kernel coefficient $\gamma \in \{\texttt{scale}, 0.001, 0.01, 0.1, 1, 10\}$. For RFR, the number of trees was fixed at 1000, while the maximum tree depth $\{2, 4, 5, 10\}$ was optimized. For the XGB regressor, number of trees $\{50, 100, 150\}$, maximum tree depth $\{1, 2, 4, 8\}$ and learning rate $\{0.03, 0.1, 0.2\}$ were optimized. Separate models were trained on the full sample and the TSST subsample.

2.4.3. Evaluation

Cross-validation was performed across participants, with each participant represented by a single feature vector. All preprocessing steps, including feature standardization and PCA, were carried out exclusively within the training folds. For classification, a nested cross-validation scheme was used, with an outer 10-fold and an inner 3-fold cross-validation for hyperparameter tuning. Performance was evaluated using classification accuracy and the area under the ROC curve (AUC). We included a majority-class baseline and conducted corrected paired t-test proposed by Nadeau and Bengio [43] that accounts for the dependency due to the cross-validation scheme. For regression, a nested Leave-One-Out (LOO) cross-validation scheme was used, with an outer LOO loop and an inner 5-fold cross-validation for hyperparameter tuning. Performance is reported as mean absolute error (MAE) and Spearman's correlation between true and predicted values. A mean-value baseline MAE was calculated on each fold, averaged across folds and tested for statistical significance using corrected paired t-test by Nadeau and Bengio [43]. Because overlapping training folds violate the independence assumption, significance tests are not reported for correlations. Feature importance was computed through Shapley Additive Explanations (SHAP) [44] and averaged across folds.

3. Results

3.1. Participants

After exclusions (five dropouts, five instances of data loss, one corrupted data recording, one baseline cortisol outlier > 3 standard deviation (SD) and one cortisol non-responder), the final sample comprised 50 participants (25 per condition). Overall, mean age was 23.24 years ($\pm$3.92) and mean BMI was 23.28 ($\pm$2.35); 23 (46.0%) participants were female. Descriptives by condition were comparable (TSST: mean age 22.48 $\pm$3.45, mean BMI 23.30 $\pm$2.77, 12 female (48.0%); f-TSST: mean age 24 ($\pm$4.27), mean BMI 23.27 ($\pm$1.90), 11 female (44.0%)).

3.2. Manipulation check

To verify successful stress induction, we compared physiological and affective responses between TSST and f-TSST. Log-transformed salivary cortisol and sAA reactivity indices, along with changes in self-reported affect (ΔNA, ΔPA), were analyzed. Cortisol responses were higher in the TSST group, with mixed-effects models showing significantly greater increases at 1 min ($b = 0.40$, $p = .001$) and 20 min ($b = 0.65$, $p < .001$), and no baseline difference ($b = 0.18$, $p = .27$). The sAA trajectories did not differ significantly between conditions (1 min: $b = -0.02$, $p = .85$; 20 min: $b = -0.03$, $p = .78$). Negative affect increased under stress and decreased in the control condition (ΔNA: TSST 3.04 ± 4.04; f-TSST -2.12 ± 4.58; $b = 3.96$, $p = .019$), while positive affect showed the opposite, nonsignificant trend (ΔPA: TSST -0.04 ± 4.68; f-TSST 4.28 ± 5.42; $b = -3.72$, $p = .065$).

Table 1: *Confusion matrix for the XGB classifier.*

	Predicted TSST	**Predicted f-TSST**
Actual TSST	20	5
Actual f-TSST	4	21

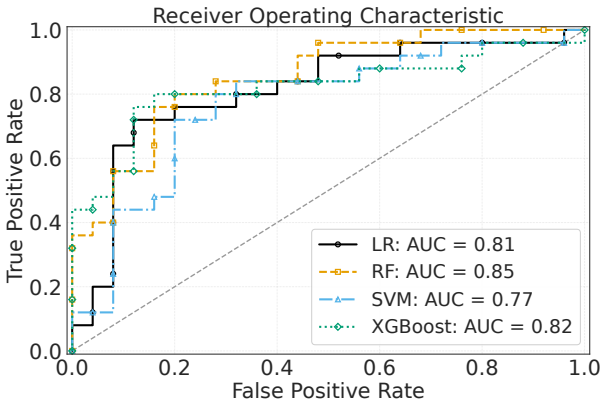

Figure 1: *ROC curves for the classification models.*

3.3. Classifications

The four ML models classified whether a participant was in the stress condition (TSST) or the control condition (f-TSST) . The highest accuracy was achieved by the XGB classifier (accuracy 0.82 ± 0.11), outperforming the majority-class baseline (corrected $t = 8.05$, $p < .001$). Misclassifications were relatively balanced, as illustrated by the confusion matrix in Table 1. Based on SHAP values averaged across all folds, the most relevant features were identified as the variability of voiced spectral flux, very-low- and low-frequency spectral energy, the rate of voiced speech segments and the variability of local shimmer (see Figure 2). The RF showed comparable performance (accuracy 0.80 ± 0.18; corrected $t = 4.62$, $p = .001$). LR reached an accuracy of 0.78 ± 0.23 (corrected $t = 3.45$, $p = .004$), while SVM achieved an accuracy of 0.74 ± 0.18 ; corrected $t = 3.90$, $p = .002$. The corresponding Receiver Operating Characteristic (ROC) curves are shown in Figure 1. Additional model variants using dimensionality reduction with PCA did not yield improved performance.

3.4. Regressions

ML regression models were trained to predict both physiological and affective stress reactivity. Among the stress responses, cortisol reactivity and negative affect could be predicted from speech-derived features, with performance varying across models (Table 2). Using the full dataset, including participants

Table 2: *Regression results comparison.*

(A) Performance of regression models for predicting stress responses (full sample)

Model	Cortisol						sAA			PANAS (Affect)					
	Reactivity			+20 min			Reactivity			ΔNA			ΔPA		
	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ
RFR	**3.73**	0.68 (.25)	0.20	5.04	-0.22 (.41)	0.10	**32.37**	1.04 (.15)	0.21	3.17	-0.09 (.47)	0.17	4.11	-0.50 (.31)	-0.10
SVR	**3.10**	2.01 (.02)	0.01	**4.78**	0.49 (.32)	-0.11	**35.87**	0.77 (.22)	-0.82	3.37	-0.57 (.29)	0.28	3.96	-0.17 (.43)	0.05
XGB	**3.41**	0.98 (.17)	0.42	5.63	-1.13 (.13)	-0.08	**35.55**	0.27 (.39)	0.18	**3.10**	0.07 (.47)	0.49	4.15	-0.58 (.28)	-0.13
Dummy	4.04	–	–	4.93	–	–	37.27	–	–	3.14	–	–	3.93	–	–

(B) Performance of regression models for predicting stress responses (TSST subsample)

Model	Cortisol						sAA			PANAS (Affect)					
	Reactivity			+20 min			Reactivity			ΔNA			ΔPA		
	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ	MAE	t (p)	ρ
RFR	**4.93**	0.73 (.24)	0.27	5.81	-0.35 (.37)	0.09	**40.14**	0.37 (.36)	0.18	**2.82**	0.94 (.18)	0.53	3.79	-0.16 (.44)	-0.06
SVR	**4.43**	1.48 (.08)	0.34	6.20	-1.30 (.10)	-0.54	43.43	-0.21 (.42)	0.07	3.45	-0.51 (.30)	0.10	3.96	-0.69 (.25)	-0.45
XGB	5.55	0.00 (.50)	0.06	5.78	-0.39 (.35)	-0.08	50.05	-0.79 (.22)	0.12	**2.08**	2.11 (.02)	0.67	3.74	-0.04 (.48)	-0.01
Dummy	5.55	–	–	5.53	–	–	42.34	–	–	3.22	–	–	3.72	–	–

Note. Dummy = mean baseline. $t(p)$ = corrected t-test: outperforms dummy. ρ = Spearman's correlation. Boldface = lower MAE than baseline.

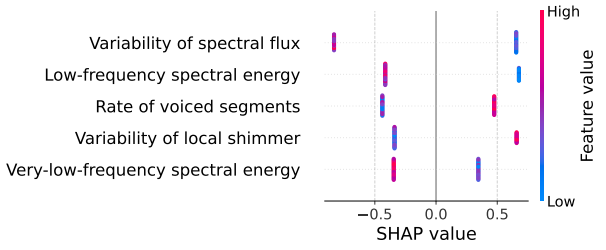

Figure 2: *Top 5 SHAP values for XGB Classifier.*

from both the stress and control conditions, the SVR model predicted cortisol reactivity more accurately than the baseline. The most relevant features for this prediction were the rate of voiced segments, low and mid-frequency spectral energy, variation in spectral tilt and the spread of low pitch values. For the TSST subsample, the SVR achieved a lower MAE than the mean baseline, although the improvement was only marginally significant. For negative affect, the XGB regressor outperformed the baseline, but the improvement reached statistical significance only for the TSST subsample. The most relevant acoustic features were the mean F0 rising slope, F1 bandwidth, Hammarberg index, alpha ratio, and the SD of F0.

4. Discussion

Our work demonstrates that subtle acoustic-prosodic characteristics of speech can be leveraged for automatic stress detection. The randomized between-participant design established a validated TSST/f-TSST contrast in stress, as confirmed by the manipulation check. The experimental conditions could be automatically distinguished from speech-derived features using different ML models; although influence of protocol-related cues cannot be fully excluded, the successful stress manipulation supports interpreting the learned acoustic differences as stress-related.This extends not only studies without a control condition [14, 15] but also a within-subject study [19], where order effects may have confounded results. Furthermore, we were able to predict key markers of distinct stress responses elicited by the TSST, namely cortisol reactivity and changes in negative affect, with performance exceeding that of a mean-baseline regressor. Statistical testing confirmed these effects, at least for the best-performing models, for relevant physiological and affective stress indices. Also, predicted values showed positive correlations with observed responses. In line with previous research, the feature-importance analysis identified features related to pitch [6, 7], speech productivity [9, 7, 6] and shimmer [45, 46]. In addition, some less frequently studied features emerged, which have recently shown promising associations with stress, namely the alpha ratio [12] and the Hammarberg index [47].

In contrast to related work [14, 15], it was not possible to consistently predict the 20 min cortisol value. This limitation may stem from not standardizing these measurements due to the small number of available time-points. However, since cortisol reactivity is considered a key marker of physiological stress [28, 29], we regard its successful prediction especially relevant. That sAA reactivity could not be predicted is unsurprising, given that, consistent with previous work [18], the f-TSST also induces a sAA response. Beyond the physiological response, we also predicted the negative affect change, which is important given that previous work [13] has shown the necessity of considering multiple facets of stress in automatic stress-response modeling. While differences in committee interaction between TSST and f-TSST limit direct comparability of pause structures, the same pattern for cortisol reactivity and ΔNA observed in TSST-only regressions supports the robustness of our results. With XGB, we chose a classifier well-suited for feature-level interpretation and that has been shown to outperform deep learning methods on tabular data [48]. Nevertheless, future research should compare such interpretable feature-based approaches with modern pretrained, end-to-end and multimodal deep learning models in similarly controlled designs, particularly as architectures incorporating temporal information, such as LSTMs, have been shown to improve performance [15]. Overall, our findings highlight speech as a promising digital biomarker of stress and demonstrate the feasibility of automated processing for stress detection and both physiological and affective stress response prediction. The analysis pipeline may serve as a useful basis for advancing objective stress assessment in both research and clinical practice.

5. Acknowledgments

The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/3 2026 – 438445824. The superordinate study [20] that provided the data was supported by the DFG project B4 of the Collaborative Research Centre (SFB) 874 "Integration and Representation of Sensory Processes" awarded to Oliver T. Wolf. We would like to thank Dennis Pomrehn, Nadja Herten and Sarah Weusthoff for their valuable contributions to this work during their time at the Department of Cognitive Psychology, Faculty of Psychology, Ruhr University Bochum.

6. Generative AI Use Disclosure

Generative AI was used only for editing and polishing the manuscript.

7. References

[1] A. Brewis, B. A. Piperata, I. Dengah, H. J. François, W. W. Dressler, M. A. Liebert, S. M. Mattison, R. Negrón, R. Nelson, K. S. Oths, J. G. Snodgrass, S. Tanner, Z. Thayer, K. Wander, and C. C. Gravlee, "Biocultural Strategies for Measuring Psychosocial Stress Outcomes in Field-based Research," *Field Methods*, vol. 33, no. 4, pp. 315–334, Nov. 2021.

[2] J. M. Bell, T. M. Mason, H. G. Buck, C. S. Tofthagen, A. R. Duffy, M. W. Groër, J. P. McHale, and K. E. Kip, "Challenges in Obtaining and Assessing Salivary Cortisol and α-Amylase in an over 60 Population Undergoing Psychotherapeutic Treatment for Complicated Grief: Lessons Learned," *Clinical Nursing Research*, vol. 30, no. 5, pp. 680–689, Jun. 2021.

[3] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, "Vocal Indices of Stress: A Review," *Journal of Voice*, vol. 27, no. 3, pp. 390.e21–390.e29, May 2013.

[4] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance," *Frontiers in Psychology*, vol. 9, Nov. 2018.

[5] L. Schewski, M. M. Doss, G. Beldi, and S. Keller, "Measuring Negative Emotions and Stress through Acoustic Correlates in Speech: A Systematic Review," *PLOS ONE*, vol. 20, no. 7, p. e0328833, Jul. 2025.

[6] M. Kappen, G. Vanhollebeke, J. Van Der Donckt, S. Van Hoecke, and M.-A. Vanderhasselt, "Acoustic and prosodic speech features reflect physiological stress but not isolated negative affect: a multi-paradigm study on psychosocial stressors," *Scientific Reports*, vol. 14, no. 1, p. 5515, Mar. 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-55550-3

[7] K. Pisanski and P. Sorokowski, "Human Stress Detection: Cortisol Levels in Stressed Speakers Predict Voice-Based Judgments of Stress," *Perception*, vol. 50, no. 1, pp. 80–87, 2021.

[8] R. Sabo and J. Rajčáni, "Designing the database of speech under stress," *Jazykovedny Casopis*, vol. 68, no. 2, pp. 326–335, 2017.

[9] T. W. Buchanan, J. S. Laures-Gore, and M. C. Duff, "Acute stress reduces speech fluency," *Biological Psychology*, vol. 97, pp. 60–66, Mar. 2014.

[10] H. Han, K. Byun, and H.-G. Kang, "A Deep Learning-based Stress Detection Algorithm with Speech Signal," in *proceedings of the 2018 workshop on audio-visual scene understanding for immersive multimedia*, 2018, pp. 11–15.

[11] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress Detection Through Speech Analysis," in *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: ICETE,*, INSTICC. SciTePress, 2018, pp. 394–398.

[12] F. Menne, H. Lindsay, J. Tröger, S. Paulmann, A. König, N. Steinbach, A. Reif, M. M. Plichta, and M. Schmidt-Kassow, "Voice as objective biomarker of stress: association of speech features and cortisol," *Acta Neuropsychiatrica*, vol. 37, p. e84, 2025. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0924270825100379/type/journal_article

[13] M. Norden, O. T. Wolf, L. Lehmann, K. Langer, C. Lippert, and H. Drimalla, "Automatic Detection of Subjective, Annotated and Physiological Stress Responses from Video Data," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2022, pp. 1–8.

[14] A. Baird, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. W. Schuller, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *Proc. Interspeech 2019*, 2019, pp. 534–538.

[15] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, N. Rohleder, H. Baumeister, and B. W. Schuller, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Frontiers in Computer Science*, vol. 3, 2021.

[16] 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," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[17] 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, May 2004.

[18] U. S. Wiemers, D. Schoofs, and O. T. Wolf, "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.

[19] M. Oesten, R. Richer, L. Abel, N. Rohleder, and B. M. Eskofier, "VoStress – Voice-based Detection of Acute Psychosocial Stress," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Oct. 2023, pp. 1–4.

[20] N. Herten, T. Otto, and O. T. Wolf, "The Role of Eye Fixation in Memory Enhancement under Stress – An Eye Tracking Study," *Neurobiology of Learning and Memory*, vol. 140, pp. 134–144, Apr. 2017.

[21] U. S. Wiemers, M. M. Sauvage, D. Schoofs, T. C. Hamacher-Dang, and O. T. Wolf, "What we remember from a stressful episode," *Psychoneuroendocrinology*, vol. 38, no. 10, pp. 2268–2277, Oct. 2013.

[22] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988.

[23] D. H. Hellhammer, S. Wüst, and B. M. Kudielka, "Salivary cortisol as a biomarker in stress research," *Psychoneuroendocrinology*, vol. 34, no. 2, pp. 163–171, Feb. 2009.

[24] U. M. Nater and N. Rohleder, "Salivary alpha-amylase as a non-invasive biomarker for the sympathetic nervous system: Current state of research," *Psychoneuroendocrinology*, vol. 34, no. 4, pp. 486–496, May 2009.

[25] R. A. Dressendörfer, C. Kirschbaum, W. Rohde, F. Stahl, and C. J. Strasburger, "Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement," *The Journal of Steroid Biochemistry and Molecular Biology*, vol. 43, no. 7, pp. 683–692, Dec. 1992.

[26] K. Lorentz, B. Gütschow, and F. Renner, "Evaluation of a direct α-amylase assay using 2-chloro-4-nitrophenyl-α-D-maltotrioside," *Clinical Chemistry and Laboratory Medicine*, vol. 37, no. 11–12, pp. 1053–1062, Nov.-Dec. 1999.

[27] E. S. Winn-Deen, H. David, G. Sigler, and R. Chavez, "Development of a direct assay for alpha-amylase," *Clinical Chemistry*, vol. 34, no. 10, pp. 2005–2008, Oct. 1988.

[28] G. E. Miller, E. Chen, and E. S. Zhou, "If it goes up, must it come down? Chronic stress and the hypothalamic-pituitary-adrenocortical axis in humans," *Psychological Bulletin*, vol. 133, no. 1, pp. 25–45, Jan. 2007.

[29] J. E. Khoury, A. Gonzalez, R. D. Levitan, J. C. Pruessner, K. Chopra, V. S. Basile, M. Masellis, A. Goodwill, and L. Atkinson, "Summary cortisol reactivity indicators: Interrelations and meaning," *Neurobiology of Stress*, vol. 2, pp. 34–43, 2015.

[30] T. Park, I. Medennikov, K. Dhawan, W. Wang, H. Huang, N. R. Koluguri, K. C. Puvvada, J. Balam, and B. Ginsburg, "Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 48 153–48 169. [Online]. Available: https://proceedings.mlr.press/v267/park25h.html

[31] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.

[32] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[33] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

[34] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Version 6.1.38, available at http://www.praat.org/, 2021, accessed: 2026-03-01.

[35] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[36] F. Eyben *et al.*, "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, extended versions (eGeMAPS) are implemented in the openSMILE toolkit; see https://github.com/nokiagiant/egemaps.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462.

[38] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, Jul. 1958.

[39] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[40] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[41] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[42] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems*, vol. 9. MIT Press, 1996.

[43] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, Sep. 2003.

[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[45] C.-K. Park, S. Lee, H.-J. Park, Y.-S. Baik, Y.-B. Park, and Y.-J. Park, "Autonomic function, voice, and mood states," *Clinical Autonomic Research: Official Journal of the Clinical Autonomic Research Society*, vol. 21, no. 2, pp. 103–110, Apr. 2011.

[46] M. Kappen, K. Hoorelbeke, N. Madhu, K. Demuynck, and M.-A. Vanderhasselt, "Speech as an indicator for psychosocial stress: A network analytic approach," *Behavior Research Methods*, vol. 54, no. 2, pp. 910–921, 2022.

[47] L. Tavi, "Acoustic correlates of female speech under stress based on /i/-vowel measurements," *The International Journal of Speech, Language and the Law*, vol. 24, no. 2, pp. 227–241, 2017. [Online]. Available: https://doi.org/10.1558/ijsll.32506

[48] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information fusion*, vol. 81, pp. 84–90, 2022.