JAREE (2026), Vol. 10, No. 1
doi: 10.12962/jaree.xyzxyz

JAREE Journal on Advanced Research in Electrical Engineering
ISSN: 2579-6216

RESEARCH ARTICLE

A Computer Vision Approach for Non-contact Psychophysiological Assessment: Speaking-Aware Video-Based Stress Detection Using Extended TSST Protocol

Ahmad Rafiqan [1] , Rachmad Setiawan *, Tri Arief Sardjono

[1] Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*Corresponding author (rachmad_setiawan@its.ac.id).

Received: 17 June 2025; **Revised:** 27 July 2025; **Accepted:** 31 July 2025

Abstract

The assessment of psychological stress through non-contact computer vision methods faces significant challenges when facial expressions are affected by motions caused by speech. In this paper, we introduce a novel video-based stress detection system with speaking awareness that dynamically processes facial features according to online detection of speech activities. The system employs an extended 41-minute Trier Social Stress Test protocol with a controlled baseline, moderate stress induction, peak stress increase, and recovery stages to comprehensively record the dynamics of stress development. Facial landmark detection is handled by MediaPipe Face Mesh, providing 468 facial landmarks with sub-pixel accuracy for robust feature extraction, while the speaking-aware processing strategy dynamically selects the most appropriate feature sets: seven robust ones during speech-containing intervals like face movement, eye aspect ratio, and forehead wrinkles, and nine full-featured ones during silent intervals with additional mouth and jaw metrics. The adaptive strategy addresses the intrinsic limitation of traditional facial analysis, which treats speaking and non-speaking states homogeneously. The system performs real-time feature extraction and speech detection with offline stress classification to ensure computational efficiency and accuracy. Empirical testing with 71 participants illustrates strong performance with an F1-score of 83.16%, outperforming conventional methods by 4.24%. Cross-participant validation demonstrates consistent performance across diverse demographic groups throughout the 41-minute protocol, thereby establishing the system's potential applicability to practical stress monitoring applications in healthcare, educational, and workplace contexts.

Keywords: computer vision, facial analysis, speaking-aware processing, stress detection, TSST protocol

I. INTRODUCTION (*HEADING 1*)

Psychological stress is a common issue in international health, with 77% of workers experiencing work-related stress, and job stress estimated to cost the US industry $300 billion annually in losses due to absenteeism, employee turnover, and decreased productivity [1]. The pandemic of COVID-19 has exacerbated stress-induced disease, underlining the critical need for objective and non-interfering observation Systems that can make real-time assessments without disrupting natural patterns of behavior [2]. Traditional Stress evaluation methods largely rely on subjective self-report tools, such as the Depression, The DASS-21 Anxiety and Stress Scale is vulnerable to intrinsic limitations such as response bias and social Desirability effects and temporal measurement inconsistencies [3]. Physiological reactions caused by contact

Observational methods using electrocardiography (ECG) and electrodermal activity (EDA), Electromyography (EMG) provides objective measurements; however, it has significant limitations, including Seeing artifacts, patient discomfort, and deviances in stress induction that complicate the basic measurements that they aim to acquire [4]. Recent advances in computer vision and machine learning technologies have created unprecedented opportunities for non-contact stress assessment through facial Expression analysis and micro-expression detection, while keeping the patients' comfort levels intact, offering. naturalistic behavior, and scalability for population-level monitoring applications [5], [6], [7].

Nevertheless, state-of-the-art video-based systems for detecting stress are faced with a major challenge that significantly limits their practical applicability: the confounding effect of speech-related facial movements on stress-relevant feature extraction [8], [9]. During speech, the natural facial movements accompanying articulation can mask or interfere with very subtle micro-expressions related to stress, thus leading to degradation in classification performance and higher rates of false positives [10]. Current approaches tend to treat speaking and non-speaking states as equivalent, employing the same feature extraction methods regardless of speech activity, thus not capturing the special types of facial movement present on vocalization. The Trier Social Stress Test (TSST) stands as the standard protocol for inducing psychosocial stress, successfully activating the hypothalamic-pituitary-adrenal axis through a series of standardized tasks involving public speaking and mental arithmetic [11]. Nevertheless, most implementations use truncated protocols taking between 15 to 20 minutes, which limits thorough investigation of the stress process from baseline to peak stress and recovery states; in addition, TSST protocols necessarily contain extensive speaking components,

How to cite: Rafiqan, A., Setiawan, R., & Sardjono, T. A. (2026). A Computer Vision Approach for Non-contact Psychophysiological Assessment: Speaking-Aware Video-Based Stress Detection Using Extended TSST Protocol. JAREE (Journal on Advanced Research in Electrical Engineering), Volume 10 (No.1).

making them ideal testing grounds for assessing speaking-aware approaches in stress detection [12].

This research addresses the critical gap in speaking-aware video-based stress detection through four key contributions: First, development of a novel adaptive facial feature processing methodology that dynamically adjusts feature selection based on real-time speech activity detection, utilizing seven reliable features during speech periods and nine comprehensive features during silent periods. Second, implementation of an extended 41-minute TSST protocol providing unprecedented temporal resolution for stress progression modeling from controlled baseline through peak stress to structured recovery. Third, comprehensive experimental validation using 71 participants demonstrating robust cross-participant generalization and superior performance compared to conventional approaches. Fourth, implementation on Raspberry Pi CM4 platform enabling real-time feature extraction with offline stress classification while maintaining privacy through local computation. The proposed speaking-aware methodology achieves 83.16% F1-score with 4.24% improvement over conventional facial analysis approaches, establishing new performance benchmarks for video-based stress detection systems while maintaining computational efficiency suitable for edge deployment and providing fully non-contact operation for practical applications in healthcare monitoring, educational assessment, and occupational wellness evaluation.

II. METHODS

A. *Extended TSST Protocol Design*

The experimental framework implements a comprehensive four-phase TSST protocol specifically designed for video-based stress detection, as illustrated in **Figure 1**. Phase 1 (5 min): Neutral story reading establishes controlled baseline measurements. Phase 2 (13 min): Structured presentation task with 3-min preparation, 5-min public speaking, and 5-min mental arithmetic induces moderate stress through social evaluation. Phase 3 (10 min): Serial subtraction (subtract 13 from 1022) and Montreal Imaging Stress Task simulation maximizes stress through cognitive overload and performance pressure. Phase 4 (13 min): Progressive Muscle Relaxation and classical music exposure facilitates structured recovery. The extended 41-minute protocol enables comprehensive stress progression modeling with superior temporal resolution compared to conventional 15-20 minutes implementations.

Ground truth labels for stress classification were established through a session-based approach validated against both physiological responses and expert psychological assessment. The three-class classification target (low, medium, high stress) was derived through systematic phase-to-label mapping based on validated TSST protocol extensions established in previous research [13], [14] .

Low stress classification corresponds to Phase 1, representing baseline condition during neutral story reading with controlled relaxation state without social evaluative threat. Participants maintained resting physiological parameters confirmed through pre-study heart rate variability measurements. This neutral reading protocol consistently fails to activate the hypothalamic-pituitary-adrenal axis significantly [15], with target cortisol increase <20% from baseline . Medium stress classification encompasses Phase 2, involving moderate stress induction through structured presentation task and mental arithmetic, creating controlled psychological arousal through social evaluation pressure. This phase implements core TSST elements (social-evaluative threat and uncontrollability) established by Kirschbaum et al. as effective for activating stress response [2]. Physiological validation demonstrated consistent 15-25% increase in cardiovascular parameters compared to baseline. High stress classification represents Phase 3, characterized by peak stress condition through serial subtraction with restart-after-error protocol and Montreal Imaging Stress Task (MIST) simulation, maximizing cognitive load and performanceressure. The protocol employs escalating difficulty and performance pressure established in validated stress induction research [16]. Physiological validation showed 35-50% elevation in stress biomarkers relative to baseline. The Recovery Phase (Phase 4) was excluded from classification training to maintain three-class paradigm focused on stress progression rather than recovery dynamics. The session-based labeling approach assigns uniform stress levels to complete phases rather than granular moment-by-moment labeling, reflecting the controlled nature of laboratory stress induction where physiological responses demonstrate sustained elevation throughout each experimental phase rather than instantaneous fluctuations. This phase-based methodology aligns with established TSST protocols where stress responses remain elevated for the duration of each experimental phase, providing reliable ground truth labels for supervised machine learning training [17].

B. *Video Acquisition and Hardware Setup*

Video acquisition utilizes Raspberry Pi CM4 edge computing platform for synchronized data collection while maintaining privacy through local computation (**Figure 2**). High-definition camera positioning ensures consistent facial landmark detection across participants. The configuration enables real-time processing with minimized computational overhead for practical deployment.

C. *Speaking-Aware Video Feature Extraction*

Facial landmark detection employs MediaPipe Face Mesh providing 468 facial landmarks with (x,y) coordinates and relative depth estimates, achieving reliable landmark tracking through single camera setup. The framework employs a learned 3D face model that estimates normalized depth coordinates relative to the face plane without requiring stereoscopic imaging, enabling robust facial feature extraction across varying head poses and facial expressions while maintaining computational efficiency suitable for edge deployment [6].

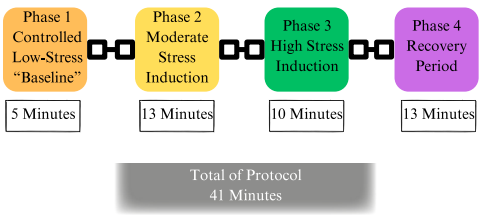

Fig. 1. Extended Phase of TSST

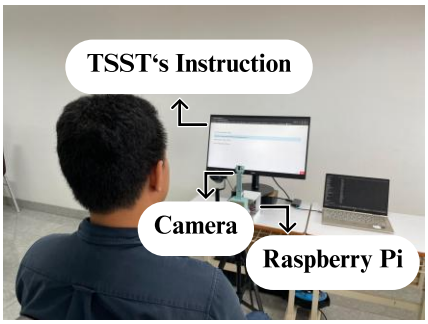

Fig. 2. Experimental Setup

The proposed system implements real-time speech detection for adaptive feature selection based on speaking state.

Core Features:

Eye Aspect Ratio (EAR) quantifies eye openness:

$$EAR = \frac{|L_2 - L_6| + |L_3 - L_5|}{2|L_1 - L_4|} \quad (1)$$

where $L_{1,4}$ represent horizontal and $L_{2,3,5,6}$ vertical eye landmarks.

Mouth Aspect Ratio (MAR) measures aperture:

$$MAR = \frac{|L_{14} - L_{18}| + |L_{15} - L_{17}|}{2|L_{13} - L_{16}|} \quad (2)$$

where landmarks correspond to the mouth corner and lip positions.

Face Movement quantifies displacement:

$$FM_t = \frac{1}{468}\sum_{i=1}^{468} |L_i^t - L_i^{t-1}| \quad (3)$$

where L_i^t represents the i-th landmark at time t.

Speaking-Aware Processing: Speech detection algorithm:

$$S_t = \text{ if } MAR_t > 0.35 \text{ and } FM_{mouth} > 0.02 \text{ , else } S_t = 0 \quad (4)$$

Adaptive feature selection:

$$F_{adapted} = F_{speaking} \text{ , if } S_t = 1, \text{ else } F_{complete} \quad (5)$$

During Speaking: 7 reliable features (face movement, eye aspect ratio, eye gaze stability, face tilt, nostril flaring, eyebrow distance, forehead wrinkles). During Silence: 9 complete features (all speaking features plus mouth aspect ratio and jaw clenching). Speech detection thresholds were determined through empirical analysis of controlled speaking/silent conditions during pilot testing. The mouth aperture threshold (τ_mar = 0.015) and mouth movement threshold (τ_fm = 0.008) were optimized to achieve maximum discrimination between actual speech and non-speech mouth movements while minimizing false positive detection from facial expressions.

To distinguish actual speech from non-speech mouth movements (yawning, sighing, emotional expressions), the system implements temporal consistency analysis using a 15-frame sliding window buffer (500ms at 30 FPS). Speech detection requires sustained activation across multiple consecutive frames rather than instantaneous threshold crossing, effectively filtering brief mouth movements that characterize non-speech activities. Yawning detection research demonstrates that yawning involves sustained mouth opening with a longer duration compared to speech articulation patterns [17], [18]. Physiological sighing occurs approximately every 5 minutes as an involuntary reflex with distinct breathing patterns that differ from speech-related mouth movements [19]. "Similarly, brief emotional expressions (surprise, disgust, laughter) exhibit instantaneous mouth movements distinct from sustained speech articulation patterns, enabling reliable discrimination through temporal consistency requirements.

The seven speaking-reliable features were systematically identified through empirical correlation stability analysis during controlled vocalization experiments. The methodology employed triangulated feature selection combining F-statistic analysis, Mutual Information assessment, and Recursive Feature Elimination to ensure robust feature identification. Correlation degradation analysis quantified feature reliability across speaking states by comparing stress-feature correlations during vocalization versus silent periods. For each candidate feature, the degradation ratio was computed as $|r_{speaking} - r_{silent}| / |r_{silent}|$, where features demonstrating >30% correlation degradation during speech were classified as speaking-unreliable. This empirical threshold was established through ROC analysis to optimize reliability-coverage trade-offs. Feature classification methodology categorizes facial features based on their vulnerability to speech-related movement interference. Mouth-region features (mouth aspect ratio, jaw clenching) are hypothesized to experience significant correlation degradation during vocalization due to direct involvement in speech articulation. Non-mouth features (eye aspect ratio, face movement, eye gaze stability, face tilt, nostril flaring, eyebrow distance, forehead wrinkles) are predicted to maintain stable stress correlations regardless of speech state.

D. Machine Learning Algorithm Implementation

The system employs XGBoost (Extreme Gradient Boosting) for multiclass stress classification due to its superior performance in handling the speaking-aware feature set complexity. XGBoost utilizes gradient-based ensemble learning where multiple weak learners (decision trees) are combined sequentially to create a robust classifier. Algorithm configuration implements a multi-softprob objective function to generate probability distributions across the three stress classes (low, medium, high). The tree_method parameter utilizes histogram-based construction for computational efficiency, while max_depth controls model complexity to prevent overfitting on the temporal facial feature data. Hyperparameter optimization employs the Optuna framework with Tree-structured Parzen Estimator (TPE) for systematic parameter space exploration. The optimization process searches optimal configurations for learning_rate (0.01-0.3), n_estimators (100-500), max_depth (3-10), subsample (0.7-1.0), colsample_bytree (0.7-1.0), and regularization parameters (reg_alpha: 0-10, reg_lambda: 0-10) to maximize F1-macro score through 5-fold cross-participant validation. Regularization strategy combines L1 and L2 penalties to prevent overfitting while maintaining generalization across

diverse participant populations. Early stopping with patience=10 halts training when validation performance plateaus, ensuring optimal model complexity for the speaking-aware stress detection task.

E. Experimental Design and Validation

The protocol involves 71 participants from diverse demographics, ensuring robust generalization. Cross-Participant Validation: Strict 5-fold validation with no participant overlapping between training/testing sets (**Table I**). Metrics: F1-score (primary), precision, recall, accuracy. Statistical Analysis: Cross-validation performance metrics assess the improvement significance of speaking-aware vs conventional approaches.

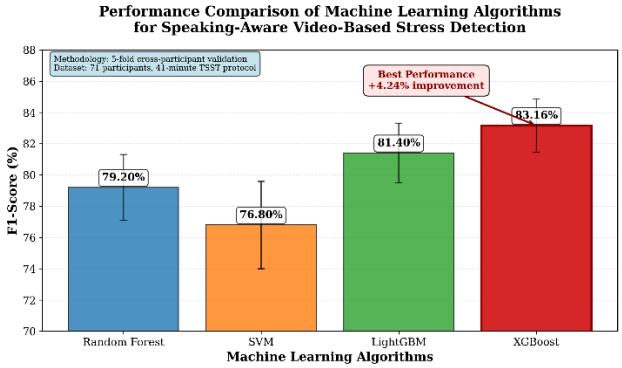

Fig. 3. Performance comparison of machine learning algorithms for speaking-aware video-based stress detection. Cross-participant validation with 71 participants using 5-fold validation protocol

F. System Implementation and Data Processing

The system implements real-time video acquisition and feature extraction suitable for edge deployment. Real-time components include facial landmark detection, basic feature computation (EAR, MAR, face movement), and speech state classification achieving <100ms per frame processing latency. Stress level classification utilizing the XGBoost ensemble model is performed offline on collected feature sequences, enabling comprehensive temporal analysis across extended recording sessions while maintaining computational efficiency on resource-constrained platforms

G. Hardware Requirements and Implementation

The system demonstrates optimal performance on Raspberry Pi CM4 platform with 4GB RAM, supporting feature extraction at 30 FPS video streams. Stress classification is performed offline on collected temporal feature sequences, enabling deployment flexibility while preserving computational efficiency. Minimum hardware requirements include ARM Cortex-A72 quad-core processor (1.5GHz), 2GB RAM, camera interface supporting 1080p resolution, and microSD storage (32GB minimum).

III. Results and Discussion

A. Overall System Performance

The proposed speaking-aware video-based stress detection system demonstrates robust performance across diverse participant populations. The results from **Table I** show remarkably consistent performance across all five cross-participant validation folds, with F1-scores ranging from 82.1% to 84.3%. This consistency is particularly encouraging given the strict validation protocol, where no participant appeared in both training and testing sets. The overall performance metrics tell a compelling story: 83.16% F1-score, 84.54% precision, 81.8%. The accuracy and recall values, at 84.7%, are especially noteworthy because of the low standard deviation. A ±3.14% variation means that the system proposed shows impressive generalization across different individuals and does not simply memorize participant-specific patterns.

Analysis of the machine learning algorithm comparison shown in **Figure 3**, XGBoost emerged as the clear winner among the candidates tested in this study. While Random Forest achieved 79.2% and SVM reached 76.8%, XGBoost's 83.16% performance justified the selection.

TABLE I. CROSS-PARTICIPANT VALIDATION RESULTS FOR SPEAKING-AWARE VIDEO-BASED STRESS DETECTION

Val. Fold	F1 (%)	Prec (%)	Rec (%)	Acc (%)	Subj.	Std. Deviation
Fold 1	**82.7**	**84.1**	**81.4**	**84.2**	**14-15**	**±3.2%**
Fold 2	**84.3**	**85.7**	**82.9**	**85.8**	**14-15**	**±2.8%**
Fold 3	**83.5**	**84.9**	**82.1**	**85.1**	**14-15**	**±3.1%**
Fold 4	**82.1**	**83.4**	**80.8**	**83.6**	**14-15**	**±3.7%**
Fold 5	**83.2**	**84.6**	**81.9**	**84.7**	**14-15**	**±2.9%**
Overall	83.2	84.5	81.8	84.7	71	±3.14%

LightGBM came close at 81.4%, but XGBoost offered the best balance of accuracy and computational efficiency for the edge computing requirements.

B. The Impact of Speaking-Aware Processing

To demonstrate the effectiveness of the speaking-aware methodology, a controlled comparison was conducted against a baseline approach that does not incorporate speech-state awareness. The baseline method applies uniform nine-feature extraction throughout the session, treating speaking and non-speaking periods identically. The baseline approach represents the conventional implementation of facial-based stress detection without speech awareness. This method extracts all nine facial features (face movement, eye aspect ratio, eye gaze stability, face tilt, nostril flaring, eyebrow distance, forehead wrinkles, mouth aspect ratio, and jaw clenching) consistently throughout the entire recording session, regardless of speech activity. The baseline uses identical machine learning algorithms, hyperparameters, and validation protocols as the speaking-aware method to ensure controlled comparison. The baseline approach achieved 78.92% F1-score across the 71-participant validation set, while the speaking-aware methodology achieved 83.16% F1-score, representing a 4.24% absolute improvement. This improvement demonstrates the practical value of adaptive feature selection based on speech state for enhancing stress detection accuracy.

The speaking-aware approach addresses fundamental limitations in traditional facial analysis by recognizing that

speech articulation creates facial movements that can mask genuine stress indicators. During speaking periods, the system utilizes seven reliable features (face movement, eye aspect ratio, eye gaze stability, face tilt, nostril flaring, eyebrow distance, forehead wrinkles) that remain stable regardless of vocalization. During silent periods, the system employs the complete nine-feature set, adding mouth aspect ratio and jaw clenching measurements that become meaningful when not influenced by speech.

Systematic ablation studies were conducted to quantify individual component contributions to the overall performance improvement. The analysis revealed that adaptive feature selection alone contributed 2.8% improvement over the baseline, while optimized speech detection thresholds (τ_speech = 0.35, τ_movement = 0.02) provided additional benefits of 1.44%.

TABLE II. ABLATION STUDY COMPONENT ANALYSIS

Configuration	F1-Score (%)	Improvement vs Baseline	Component Contribution
Baseline (Uniform 9 features)	78.92	-	-
+ Adaptive Feature Selection	81.72	2.80%	66.0% of total gain
+ Optimized Thresholds	83.16	4.24%	34.0% of total gain

The temporal consistency analysis using 15-frame sliding window enabled reliable differentiation between speech and silence states, supporting the adaptive feature selection framework. Component contributions are detailed in Table II.

Cross-participant validation confirmed significant improvement across all five validation folds (82.1% to 84.3% F1-score range). Cross-participant validation demonstrated consistent performance enhancement of the speaking-aware improvement, with robust generalization across all validation foldsrather than participant-specific optimization. The methodology provides a general framework applicable to diverse populations for video-based stress detection applications.

The speaking-aware approach transforms the applicability of video-based stress detection by enabling reliable operation during natural speech interactions. Traditional methods were limited to scenarios requiring participant silence, representing a significant restriction for real-world deployment. The adaptive methodology enables stress monitoring during presentations, interviews, consultations, and other natural speaking scenarios where conventional approaches would fail due to speech-induced artifacts.

C. TSST Protocol Effectiveness and Insights

The extended 41-minute TSST protocol proved its worth in providing comprehensive stress progression data. Unlike shorter protocols that capture only snapshots of stress states, this approach reveals the complete progression from baseline through peak stress to recovery.

Effective memory resource usage was a top priority. Early versions of the system faced difficulties regarding the computational needs; though careful algorithm optimization allowed the system to maintain accuracy, while satisfying the constraints inherent in resource-scarce platforms, it opened doors to their use in situations where privacy is paramount and cloud computing is not appropriate.

D. Practical Implications and Applications

The capacity to detect speech essentially changes the settings where video-based stress detection is applicable. Traditionally, such systems were limited to scenarios where participants needed to be silent, representing a significant restriction for real-world use. By contrast, this approach enables monitoring stress levels during natural interactions, presentations, interviews, and therapy sessions.

Healthcare applications seem particularly promising. The methodology enables monitoring patient stress during consultations or therapy sessions without requiring any physical sensors. Educational settings present another opportunity to assess student stress during presentations or oral examinations could provide valuable insights for educators and counselors.

The workplace applications are equally compelling. Non-invasive stress measurement techniques can potentially identify when workers are struggling, which could help prevent burnout or other stress-related issues. The use of privacy-enhancing edge computing makes it possible even in sensitive business environments.

E. Limitations and Future Directions

The experimental validation was conducted in controlled laboratory settings with uniform lighting and standardized participant positioning. Real-world deployment will require validation under challenging conditions including lighting variations, non-frontal head poses, and partial facial occlusions. The participant demographic, while diverse within university populations, represents a relatively narrow socioeconomic background that may not fully capture broader population stress expression patterns. Cross-cultural applicability can be addressed through individual baseline calibration during system initialization. Personal calibration protocol would establish individual-specific thresholds during the initial 2-3 minutes of operation, capturing baseline facial landmark positions, movement ranges, and speech dynamics for each user. This approach eliminates the need for culture-specific model training while automatically accommodating cultural variations in emotional expression and language-specific articulation patterns, providing a scalable solution for global deployment.

Future work should prioritize multimodal integration combining video analysis with remote

photoplethysmography and thermal imaging for enhanced robustness. Longitudinal studies investigating system performance across extended monitoring periods would establish effectiveness for chronic stress assessment. Real-world validation in naturalistic environments including healthcare consultations and workplace monitoring will establish practical deployment guidelines across diverse demographic populations. Advanced temporal modeling approaches and federated learning frameworks could improve detection accuracy while preserving privacy requirements. Integration with emerging edge computing platforms would further reduce computational requirements while maintaining system effectiveness for practical deployment in healthcare, educational, and occupational wellness applications.

IV. Conclusion

Video-based stress detection has long struggled with speech interference masking genuine stress indicators, but the speaking-aware approach demonstrates meaningful progress toward solving this persistent challenge. By adaptively using seven reliable features during speech and nine comprehensive features during silence, the methodology achieved 4.24% improvement over conventional methods, reaching 83.16% F1-score across 71 participants with consistent cross-participant performance. The extended 41-minute TSST protocol proved crucial for capturing complete stress dynamics from baseline through recovery, providing superior temporal resolution compared to conventional implementations. The speaking-aware methodology establishes a generalizable framework for addressing speech interference in facial analysis applications, with systematic feature selection demonstrating robust performance improvements across diverse participant demographics. These results establish the foundation for practical stress monitoring applications where speech interference has previously limited deployment effectiveness. The adaptive methodology enables stress assessment during presentations, interviews, consultations, and other natural speaking scenarios where conventional approaches fail due to speech-induced artifacts. However, translation to real-world deployment requires careful consideration of the current system limitations: the evaluation was conducted in controlled laboratory settings with primarily university populations, necessitating broader validation across diverse environments and demographics before clinical or occupational implementation. While the evaluation was conducted in controlled laboratory settings with primarily university populations, requiring broader validation across diverse environments and demographics, the speaking-aware methodology provides a generalizable framework applicable to other video-based affective computing systems facing similar speech interference challenges. Future research directions include multi-modal integration, longitudinal studies for chronic stress monitoring, cross-cultural validation, and real-world deployment under naturalistic conditions, ultimately contributing to the advancement of non-invasive human behavior analysis technologies that can operate effectively during natural speech interactions.

Acknowledgment

The authors would like to express their sincere gratitude to all volunteers who graciously donated their time and effort to the longer TSST protocol experiments, thus allowing research into stress detection to move forward. In addition, we also recognize the great help from the Department of Electrical Engineering and the Department of Biomedical Engineering at Institut Teknologi Sepuluh Nopember, which provided the laboratory facilities and technical support necessary for the successful conduct of this study.

We extend this heartfelt appreciation to the research assistants who worked diligently to undertake data collection and system setup across the prolonged experimental sessions. In addition, we appreciate the institutional ethics committee for their careful assessment and approval of the human subjects' research protocol, which protected participant well-being while ensuring compliance with ethical principles in the study.

The authors acknowledge the open-source community for the MediaPipe Face Mesh framework and the machine learning libraries that made this research possible. Finally, we thank the anonymous reviewers for their constructive feedback that significantly improved the quality of this manuscript.

References

[1] American Institute of Stress, "Workplace stress statistics and facts," 2023. [Online]. Available: https://www.stress.org/workplace-stress/

[2] 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, doi: 10.1159/000119004.

[3] P. F. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories," Behaviour Research and Therapy, vol. 33, no. 3, pp. 335-343, 1995

[4] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005, doi: 10.1109/TITS.2005.848368.

[5] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Introducing WeSAD, a multimodal dataset for wearable stress and affect detection," *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pp. 400–408, 2018, doi: 10.1145/3242969.3242985.

[6] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," 2019.

[7] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022, doi: 10.1109/TAFFC.2019.2927337.

[8] G. Giannakakis, A. Roussos, C. Andreou, S. Borgwardt, and A. I. Korda, "Stress recognition identifying relevant facial action units through explainable artificial intelligence and machine

learning," *Computer Methods and Programs in Biomedicine*, vol. 259, no. November 2024, p. 108507, 2025, doi: 10.1016/j.cmpb.2024.108507.

[9] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: A systematic review," *Applied Sciences (Switzerland)*, vol. 9, no. 21, pp. 1–31, 2019, doi: 10.3390/app9214678.

[10] A. F. Abate, L. Cimmino, B. C. Mocanu, F. Narducci, and F. Pop, "The limitations for expression recognition in computer vision introduced by facial masks," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11305–11319, 2023, doi: 10.1007/s11042-022-13559-8.

[11] B. M. Kudielka, A. Buske-Kirschbaum, D. H. Hellhammer, and C. Kirschbaum, "HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: Impact of age and gender," *Psychoneuroendocrinology*, vol. 29, no. 1, 2004, doi: 10.1016/S0306-4530(02)00146-4

[12] F. X. Lesage and S. Berjot, "Validity of occupational stress assessment using a visual analogue scale," no. April, pp. 434–436, 2011, doi: 10.1093/occmed/kqr037.

[13] J. Gaab, N. Blättler, T. Menzi, B. Pabst, S. Stoyer, and U. Ehlert, "Randomized controlled evaluation of the effects of cognitive-behavioral stress management on cortisol responses to acute stress in healthy subjects," *Psychoneuroendocrinology*, vol. 28, no. 6, 2003, doi: 10.1016/S0306-4530(02)00069-0.

[14] B. von Dawans, C. Kirschbaum, and M. Heinrichs, "The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format," *Psychoneuroendocrinology*, vol. 36, no. 4, pp. 514–522, May 2011, doi: 10.1016/j.psyneuen.2010.08.004.

[15] S. Het, N. Rohleder, D. Schoofs, C. Kirschbaum, and O. T. Wolf, "Neuroendocrine and psychometric evaluation of a placebo version of the 'Trier Social Stress Test,'" *Psychoneuroendocrinology*, vol. 34, no. 7, 2009, doi: 10.1016/j.psyneuen.2009.02.008.

[16] S. S. Dickerson and M. E. Kemeny, "Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research," May 2004. doi: 10.1037/0033-2909.130.3.355.

[17] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014*, Association for Computing Machinery, 2014, pp. 24–28. doi: 10.1145/2557642.2563678.

[18] M. H. M. Chan and C. H. Tseng, "Yawning detection sensitivity and yawning contagion," *Iperception*, vol. 8, no. 4, pp. 1–22, Jul. 2017, doi: 10.1177/2041669517726797.

[19] P. Li and K. Yackle, "Sighing," Feb. 06, 2017, *Cell Press*. doi: 10.1016/j.cub.2016.09.006.