Decoding the Hidden Language of Stress: Analyzing Speech Patterns as a Tool for Stress Detection

Bachelor's Thesis in Medical Engineering

submitted
by

Charlotte von Roznowski

born 28.04.2002 in Melbourne

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

in Cooperation with

Chair of Health Psychology
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Luca Abel M.Sc., Robert Richer M.Sc., Prof. Dr. Bjoern Eskofier,
Miriam Kurz M.Sc., Prof. Dr. Nicolas Rohleder (Chair of Health Psychology)

Started: 01.04.2024

Finished: 02.09.2024

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.
Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 2. September 2024

Übersicht

Stress ist ein Reaktionsmechanismus, der unserem Körper hilft, mit unmittelbaren Bedrohungen umzugehen, aber über längere Zeit schädliche Auswirkungen auf unsere Gesundheit haben kann. Daher ist die Stresserkennung ein stark erforschtes Themengebiet. Biomarker wie Cortisol, Alpha-Amylase und der Anstieg von Entzündungswerten stellen eine zuverlässige Methode der Stresserkennung dar und können in Laboren gemessen werden, was jedoch kostenintensiv und potenziell das natürliche Verhalten in Stresssituationen beeinflusst.

Aus diesem Grund gewinnen nicht-invasive, kontaktlose und schnelle Methoden zur Stresserkennung mit digitalen Biomarkern zunehmend an Bedeutung. Die Forschung hat gezeigt, dass Stressreaktionen durch Kameras und Mikrofone erkennbar sind, etwa durch Körper- und Gesichtsbewegungen sowie Veränderungen in Stimme und Sprachmuster.

Um diese Ansätze zu erweitern, wurde eine Studie mit 61 Teilnehmern durchgeführt, bei der Veränderungen in der Sprache während des ”Trier Social Stress Test” (TSST) und einer Kontrollbedingung, dem ”friendly-Trier Social Stress Test” (f-TSST) untersucht wurden. Die Teilnehmer wurden während der Tests anhand eines Mikrofons aufgezeichnet, um die Sprachveränderungen zu analysieren. Die Audioaufnahmen wurden so zugeschnitten, dass sie nur die Sprachanteile der Probanden beinhalteten, und dann transkribiert, um eine Sprach- und Stimmungs-Analyse zu ermöglichen. Übereinstimmend mit existierenden Studien sprachen die Teilnehmer während des TSST weniger, was sich in der reduzierten Anzahl von Wörtern und Sätzen zeigte. Interessanterweise ergab die Stimmungs-Analyse entgegen den Ergebnissen bestehender Studien während des TSST positivere Werte. Eine geschlechtsspezifische Analyse ergab keine signifikanten Unterschiede.

Um die Vorhersagefähigkeit von Sprachmerkmalen hinsichtlich stressbedingter Cortisolerhöhungen zu untersuchen, wurde ein maschinelles Lernmodell trainiert. Dabei erzielte das Modell eine Klassifikationsgenauigkeit von 95.0 % ± 7.3 % bei der Unterscheidung der Testbedingungen. Eine anschließende Regressionsanalyse zeigte jedoch, dass Cortisolerhöhungen anhand von Sprachmerkmalen nicht zuverlässig vorhergesagt werden konnten, was darauf hindeutet, dass das Modell eher die Testbedingungen als tatsächliche Stressreaktionen reflektiert und dass Cortisolanstiege mit allein der Sprachanalyse nicht vorhergesagt werden können. Dennoch erreichte die Klassifikation eine Genauigkeit von 62.7 %, wenn die Teilnehmer basierend auf ihrem Cortisolanstieg als Responder oder Non-Responder eingestuft wurden, was darauf schließen lässt, dass Sprachmerkmale potenziell wertvolle Indikatoren für Stress sein könnten.

Abstract

Stress is a response mechanism that helps our body cope with immediate threats but can have detrimental effects on our health if sustained over long periods. Consequently, stress detection is a vastly researched subject. Biomarkers such as cortisol, alpha amylase (α-amylase), and inflammatory increase can be measured through bodily fluids, and are typically evaluated in a laboratory setting. While this method provides reliable results, it requires significant funding due to the need for trained personnel to conduct the complex and time-consuming procedures. Furthermore, such measurements can impact natural behaviour, potentially altering the stress response.

Therefore, the detection of stress using non-invasive, contactless, and fast methods utilizing digital biomarkers is of growing interest. Research has shown that stress reactions can be detected through cameras and microphones, by observing a person's bodily and facial movements, as well as changes in voice or speech patterns.

To further explore stress detection through digital biomarkers, this study investigated changes in lexical speech during an acute psychosocial stress test. A total of 61 participants underwent the Trier Social Stress Test (TSST) and a control condition, the friendly Trier Social Stress Test (f-TSST). Participants were recorded using a microphone during the tests to analyze speech changes. The audio recordings were trimmed to include only the participants' talk segments, and then transcribed to enable lexical, grammatical, and sentiment analysis. Consistent with prior research, participants spoke less during the TSST, as evidenced by the reduced number of words and sentences. Interestingly, contrary to existing studies, sentiment was more positive during the TSST. Additionally, an analysis of gender differences in speech revealed no significant variations.

To evaluate the ability of speech features to predict stress-related cortisol increases, a machine learning (ML) model was trained to differentiate between the two conditions, achieving a notable classification accuracy of 95.0 % ± 7.3 %. However, when this was further validated through ML-based regression to predict maximum relative cortisol increases, the results indicated that cortisol increases could not be reliably predicted by speech features alone, suggesting that the classification model was more reflective of the test conditions than actual stress responses. Despite this, when participants were categorized as ”responders” or ”non-responders” based on their cortisol increase, classification accuracies improved to 62.7 %, indicating that speech features may still contain valuable indicators of stress.

Contents

Chapter 1

Introduction

Voice and speech are integral components of daily communication, but they may also serve as valuable indicators of physical and mental well-being. Stress, a common experience in various situations such as exams, job interviews, or physical and emotional challenges, triggers both physical and psychological reactions. While the body's stress response is designed to protect us by releasing adrenaline and other hormones that enhance brain and bodily functions [OCo21], prolonged stress can lead to severe health consequences. These include physical illnesses such as immune system dysfunction, gastrointestinal and cardiovascular diseases, and even cancer, as well as psychological issues like anxiety and depression [Sal08]. Consequently, stress has been recognized as a significant factor contributing to long-term physical and mental illnesses, making it a subject of extensive research [Rob18].

Stress triggers a cascade of neuro-endocrine responses in the body, primarily regulated by the sympathetic nervous system (SNS) and the hypothalamic-pituitary-adrenal (HPA) axis. The SNS initiates the "fight-or-flight" response, releasing the stress biomarker α-amylase [Roh06]. Furthermore, the activity of the SNS can be detected through heart rate and heart rate variability using electrocardiographic signals [Daw16]. On the other hand, the hormone cortisol marks the activation of the HPA axis [Kir94]. These responses are designed to help the body cope with immediate threats but can have detrimental effects if sustained over a prolonged period of time [Coh97].

To study these responses to stress, reliably inducing stress is crucial. Currently, the TSST is the gold standard for inducing acute psychosocial stress in a controlled laboratory setting [Kir93]. This test involves tasks designed to be discomforting, such as public speaking and a challenging arithmetic task, both performed in a socially evalulative context. To provide a comparison to a stress-free environment while maintaining similar cognitive demands, the f-TSST is used as a control condition [Wie13]. During these tests, participants are equipped with electrodes connected to an electrocardiogram to measure heart rate and heart rate variability. Additionally, blood and saliva samples are collected before and after the tests to measure stress on a biological level. However, these invasive methods can interfere with participants' natural behavior, potentially triggering cortisol spikes or restricting movement due to the ECG cables. Moreover, these procedures are costly, time-consuming, and require significant effort.

Given these challenges, there is growing interest in using digital biomarkers to detect stress, as they offer a faster, simpler, and less invasive approach. Research has shown promising results in recognizing stress through body movement and facial expression analysis [Ric24; Bla23], as well as through changes in voice and speech [Oes23; Buc14]. Digital biomarkers can facilitate early and personalized stress detection by identifying specific behavioral patterns, which could help mitigate stress. However, research on the lexical aspects of speech under stress is still limited, making it an interesting area for further investigation.

The aim of this bachelor's thesis is to analyze speech changes extracted from audio data to better understand stress responses, particularly in relation to the lexical aspects of speech. The data was collected as part of a study within the EmpkinS Collaborative Research Center [Emp21], involving 61 participants who performed the TSST and the control condition, the f-TSST, on two consecutive days. The focus was on identifying changes in lexical diversity, grammar, and sentiment measures. Speech features were extracted using *Whisper* for transcription [Rad23], *Natural Language Toolkit (nltk)* [Bir09] and *spaCy* [Hon17] for grammatical analysis, and *TextBlob* for sentiment analysis [Lor18]. Additionally, gender differences were examined, and ML techniques were employed to classify conditions and predict cortisol spikes.

Chapter 2

Related Work

Stress is a multifaceted response mechanism to a number of stressors, which influences changes in overall behaviour, body language, facial expressions, and speech. Hence, understanding these changes is crucial for the development of stress recognition systems. Stress recognition has been an active research area in general, with increasing contribution from ML communities [Pan19]. Thus, various studies examining methods of stress detection using digital biomarkers, focusing on body language, facial expressions and vocal or speech changes, will be outlined in the following Section.

Bodily Movements

The relationship between body movements and stress levels has been researched, with results indicating that stress often causes distinct postural and movement patterns.

For instance, Shahidi et al. discovered that mentally challenging tasks in a seated workplace environment can cause changes in cervical posture. Specifically, increased mental concentration resulted in a more forward head posture, while psychosocial stress alone led to increased muscle activity in the upper trapezius, indentifying this muscle as a potential stress marker [Sha13]. Similarly, Qiu et al. corroborated these results in a similar study, where an elevated mental workload resulted in a decrease in distance between the head and the display screen [Qiu12].

Besides posture, body movements - or the absence thereof - can also indicate states of stress. Focusing on gait, Lasselin et al. found that inflammatory processes, which can be triggered by stress [Roh19], manifest in a generally slower and more rigid walking pattern. In their study, the injection of lipopolysaccharide, which induces inflammation, led to participants exhibiting shorter, slower, and wider strides, less knee flexion, less arm extension, and a more downward-tilting head posture [Las20]. Conversely, a lack of body movement, or ”freezing”, can also pose as an indicator of stress. This response is a common defense mechanism among animals, helping them avoid detection by predators [Mis03]. This phenomenon was observed in a study by Roelofs et al., which exposed 50 female participants placed on a stabilometric force platform to social threats. The study found that participants displayed reduced body sway when presented with angry faces, compared to their reaction to neutral or happy faces. Along with reduced body sway, bradycardia and subjective anxiety was observed when participants were exposed to angry faces, indicating freeze-like behaviour to social threat cues in humans [Roe10].

Supporting these findings, Richer et al. documented similar freezing behaviour during the TSST compared to a control condition, the f-TSST, in a study involving 59 participants equipped with inertial measurement unit (IMU)-based motion capture suits. The psychosocial stress-inducing test resulted in reductions in overall body motion and prolonged static periods. Applying ML techniques to the motion data collected from the suits achieved a classification accuracy of 73.4 % in stress detection. Notably, features characterizing movements of the upper extremities and head had the largest effect size, suggesting their potential as stress biomarkers, thereby reinforcing Shahidi's et al. findings [Sha13]. This result implies that body posture and movements, especially freezing, can be utilized to detect psychosocial stress [Ric24]. Another study, which induced stress through an arithmetic task, achieved a mean accuracy of 77 % in stress detection, by observing both body and face movement using support vector machines (SVM). This study further validated the findings of Richer et al., demonstrating that movement quantity and postural changes were the most accurate classifiers when analyzing the body [Aig15].

Facial Expressions

Research has demonstrated a correlation between stress and negative emotions such as anger, anxiety, and disgust [Laz06; Oza21], and consequently, the associated facial expressions. For example, Almeida et al. utilized convolutional neural networks (CNN) to classify seven different emotions, including anger, sadness, fear, disgust, happiness, neutrality, and surprise, from video data, achieving an average F1-score of 89 %. Building on this, they employed ML techniques to detect stress by categorizing anger, disgust and fear as stress-related emotions, resulting in a binary evaluation with an average F1-score of 92 % [Alm21].

Moreover, Zhang et al. achieved an F1-score of 85.3 % by analyzing stress indicators such as increased blink rate, pupil dilation and mouth activity [Zha20]. Giannakis et al. confirmed these indicators as reliable signs of stress. Instead of associating stress solely with negative emotions, their study induced stress through video stimuli and mental tasks. The highest classification accuracy, 91.68 %, was obtained during the social exposure phase of the study using Adaptive

Boosting (AdaBoost) classifier. The contributing facial features included eye-related and mouth activity, head movement amplitude, and heart rate estimation elicited from alterations in facial skin colour [Gia17]. Another study employed the TSST to induce stress in 62 participants, measuring salivary cortisol, subjective experience, heart rate, heart rate variability, as well as facial activity using cameras. The analysis revealed changes in the aforementioned facial regions, when stress was induced. Especially, more frequent eyelid tightening, upper eyelid raising and upper lip raising was correlated with elevated cortisol levels [Bla23]. Furthermore, Aigrain et al. achieved classification accuracies exceeding 70 % by observing the correlation between brow lowering and nose wrinkling as raw features in their attempt to detect stress through body movements and facial activity [Aig15].

To address gender differences, female participants showed an increased cheek raising and lip corner pulling intensity, commonly associated with smiling. This finding is likely linked to the ”tend and befriend” stress response, which is particularly prevalent among women, as they often respond to psychosocial stress with positive affiliative social behavior [Bla23].

Vocal and Speech Changes

Recent research has explored the relationship between stress and vocal changes, with findings indicating that stress related cortisol spikes can be predicted through audio recordings. Baird et al. conducted a study involving 134 participants who underwent the TSST. Their results showed that stress-related physiological responses such as emotion, heart rate, and respiration could be detected through vocal analysis, despite significant individual variations in stress manifestation [Bai21].

Further supporting these findings, Laukka et al. investigated how anxiety influences non-verbal vocal parameters. Their research revealed that stress, induced by public speaking, typically leads to an increase in pitch, greater variability in pitch, higher speech intensity, and faster speech rate. These vocal changes are reflective of the physiological arousal and tension associated with stress [Lau08].

Oesten et al. developed *VoStress*, a system that analyzes the acoustic aspects of speech to predict cortisol spikes during stress. Their study involved 21 participants which underwent both the TSST and f-TSST. The process involved speaker diarization, trimming audio segments, and utilizing the *OpenDBM* library to compute digital biomarkers. The study identified key acoustic changes under stress, such as increased audio intensity variation, higher formant frequency, and decreased shimmer. By employing ML techniques, such as stepwise backward multiple linear regression, they achieved an 80.0 % classification accuracy in stress detection [Oes23].

While the correlation between stress and lexical changes in speech has not been studied extensively, several adjacent areas, such as sentiment analysis and fluency, have been explored.

The former was investigated by Chyan et al. who applied CNN to a dataset of audio recordings where actors displayed specific emotions. Their primary goal was to detect negative emotions associated with stress. In addition to sentiment detection, their study delved into vocal pattern analysis. They examined various acoustic features such as pitch, intensity, speech rate, and formant frequencies, which are known to be influenced by emotional states [Oes23; Lau08]. Sentiment analysis was performed using a Word-Embedding model that captured semantic relationships between words. By combining acoustic and sentiment features in a multi-stage model, they achieved a robust stress detection framework with an F1-score of 91 % [Chy23].

Similarly, a study by Belouali et al. explored the relationship between speech patterns and mental health, focusing specifically on suicidal ideation. This research utilized both acoustic and linguistic features to detect markers of mental distress. Key acoustic indicators, such as reduced speech rate and flatter pitch contours, were linked to suicidal tendencies. Though lexical and sentiment analysis did not show significant changes, certain linguistic trends, including the use of more superlative adverbs, possessive pronouns, and proper nouns, were noted. Using *XGBoost*, the study achieved a classification accuracy of 78 %, demonstrating the potential of combining acoustic and linguistic analyses for detecting mental health conditions, which aligns with the broader aim of identifying stress and emotional distress through speech analysis [Bel21].

Buchanan et al. explored the effects of stress on speech fluency by using the TSST and a control condition, the placebo Trier Social Stress Test (p-TSST) [Het09b]. The study observed that participants initially spoke fluently under stress, but as cortisol levels stabilized, their fluency decreased. This reduction may be linked to either the delayed effects of cortisol or the participants' limited preparation for the speech content. Notably, the TSST condition was associated with a higher frequency and longer duration of unfilled pauses, which correlated with elevated cortisol levels and heart rates. In contrast, participants in the p-TSST condition used more filler words, such as ”hmm” or ”uh”, resulting in higher speech fluency. Additionally, the study found that both communication rate and word productivity were higher in the TSST condition, particularly at the beginning of the tests [Buc14].

Chapter 3

Methods

In the following Sections, the study's methodology will be detailed, including the processes of data collection, feature selection, and ML approaches. These methods build upon the research discussed in the previous Chapter 2 and are designed to provide a robust framework for addressing the research objectives.

3.1 Data Acquisition

To assess the effects of acute psychosocial stress on speech and language, a study was conducted at the *Machine Learning and Data Analytics Lab (MaD Lab)* from December 2022 to December 2023. Participants were asked to undergo the TSST and the f-TSST on two consecutive days, with a pseudo-randomized condition order.

3.1.1 Study Population

In total, 61 participants (35 females, 25 males, and one identifying as non-binary) were recruited for the study. An overview of condition order and gender can be found in Table 3.1.

Table 3.1: Gender and condition order overview

	Gender			
Condition order	**Female**	**Male**	**Non-Binary**	Total
f-TSST first	21	11	0	32
TSST first	14	14	1	29
Total	35	25	1	61

Due to incomplete data, the final dataset was reduced to 40 participants (13 males, 26 females, and one identifying as non-binary). An overview over the condition order and gender can be found in Table 3.2. The demographic as well as anthropometric data of both the entire and final participant list can be found in Table 3.3.

Table 3.2: Gender and condition order overview - final

	Gender			
Condition order	**Female**	**Male**	**Non-Binary**	Total
f-TSST first	16	11	0	27
TSST first	5	7	1	13
Total	21	18	1	40

Participants were recruited using printed flyers distributed in university buildings, canteens, and libraries, as well as electronic flyers on social media platforms. Exclusion criteria included being under 18 or over 50 years old, not-native or insufficient German language skills, suffering from physical or mental health conditions, use of medication or drugs, smoking, obesity (Body Mass Index (BMI) over 30), or previous participation in a similar stress test. In addition, masters' students of psychology were excluded, as the probability of being familiar with the (friendly) Trier Social Stress Test ((f-)TSST) or a similar stress test was high. Applicants had to complete a digital screening beforehand and were only granted participation if the mentioned criteria were met.

Table 3.3: Demographic and anthropometric data of all and final participants; mean ± standard deviation (SD)

	Age (years)	**Height** [cm]	**Weight** [kg]	**BMI** [kg/m^2]
Female	21.69 ± 4.06	166.88 ± 6.42	60.69 ± 7.09	20.81 ± 4.63
Male	22.00 ± 2.86	184.0 ± 7.19	75.31 ± 8.70	22.27 ± 2.49
Non-Binary	24.00 ± 0.00	158.0 ± 0.00	45.00 ± 0.00	18.03 ± 0.00
Final	21.85 ± 3.63	168.00 ± 28.79	63.23 ± 14.73	21.22 ± 4.05
All	22.20 ± 3.95	174.39 ± 11.23	68.10 ± 11.89	22.29 ± 2.47

As compensation for participation in the study, participants had the option to either receive 50 € or five *Versuchspersonenstunden* (for psychology students).

3.1.2 Acute Stress Induction

In order to induce psychosocial stress, the TSST, the gold standard for assessing acute psychosocial stress in a laboratory setting, was used [Kir93]. As a counterpart, the f-TSST was used as a control condition, designed not to trigger a stress response while posing similar tasks and mental demands. Figure 3.1 provides an overview of the tasks involved in the conditions.

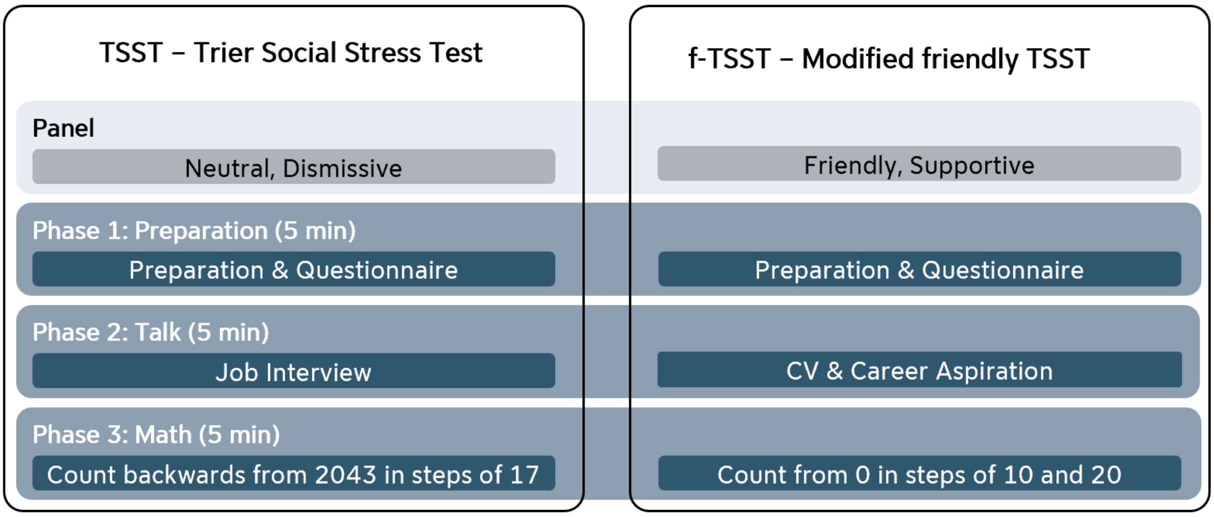
TSST – Trier Social Stress Test
f-TSST – Modified friendly TSST
Panel
Neutral, Dismissive
Friendly, Supportive
Phase 1: Preparation (5 min)
Preparation & Questionnaire
Preparation & Questionnaire
Phase 2: Talk (5 min)
Job Interview
CV & Career Aspiration
Phase 3: Math (5 min)
Count backwards from 2043 in steps of 17
Count from 0 in steps of 10 and 20

Figure 3.1: Protocol comparison of the TSST and f-TSST.

TSST

The TSST was held in front of a two person panel consisting of two female interviewers wearing white lab coats to create a more formal and laboratory-like environment. Due to a lack of male personnel, a slight alteration from Kirschbaums et al. version of the TSST was made, which enquires one female and one male interviewer [Kir93]. Consistent with Kirschbaums et al. TSST, the panel was instructed to remain entirely neutral towards the participants and show minimal reaction to their actions. The active member of the panel, placed to the left side from the participants perspective, handled the interactive part of the interview and was asked to only use a certain set of responses to ensure a neutral demeanor and uniform test.

The protocol consisted of three five minute phases: Preparation, talk and math. The preparation phase began after the study leader had explained the task and left the room. In the task explanation the participants were instructed to imagine being in a job interview for their dream job, which they disclosed to the study leader beforehand, with the panel deciding whether they get the job. The study leader emphasized that professional qualifications were already known and the participants should focus on their personality. Then, participants were given a three minute silent preparation period to take notes for their talk and a two-minute period to answer the questions of the primary appraisal secondary appraisal (PASA) questionnaire [Gaa09], which is designed to asses cognitive appraisal processes in stressful situations [Cam04]. The notes taken during this period were not allowed to be used during the talk.

After the preparation phase, two cameras including one microphone (for details see Section 3.1.3) were turned on, and participants were instructed to start their five-minute talk. The active panel member only interrupted the test subject if they sidetracked to topics other than their personality traits, did not speak for more than 20 seconds, or avoided eye contact. Following the talk, participants were challenged with an arithmetic task, in which they were instructed to count backward from 2043 in steps of 17. The active interviewer stopped the participants if a mistake was made and ordered them to start over, or interrupted them if they did not hold eye contact.

f-TSST

To create a comparable control condition, the f-TSST was set up similarly to the TSST, but without the stressful components. This included identical timelines, less stressing yet relatable tasks, and a two-person panel, which however, abstained from wearing lab coats to ensure a more comfortable environment. Furthermore, both interviewers were permitted to interact with participants during the test and were instructed to behave in a friendly and encouraging manner throughout the protocol. To further relax the participants, the active panel member left the room during the preparation phase.

The talk section of the f-TSST differed from the original TSST in that participants were asked to go into detail about their Curriculum Vitae and career aspirations, as opposed to the more challenging topic in the TSST. Additionally, the panel asked targeted questions to encourage participants to speak more about themselves, creating a more relaxing and conversational environment.

In contrast to the original (f-)TSST protocol, which does not include an arithmetic task [Wie13], this f-TSST incorporated a numerical task. Participants were asked to add the values 10 and 20 to the starting value 0 in alternating steps. If a mistake was made, the panel intervened in a friendly manner and asked to continue from the last correctly calculated value. This approach alines with the protocol of the placebo-TSST [Het09a] and was integrated to establish a low-stress environment while maintaining consistency.

3.1.3 Test Procedure

Pre-Test Phase

Upon arrival at the laboratory, participants were guided to the preparation room, seated, and asked to sign a declaration of consent. The study leader then provided a brief summary of the study timeline and proceeded in taking the first of eight saliva samples (S_0). An overview of the saliva sampling times throughout the study is provided in Table 3.4.

Table 3.4: Saliva sampling times relative to the (f-)TSST start.

Relative time [min]	-40	-1	0-15	$+16$	$+25$	$+35$	$+45$	$+60$	$+75$
Saliva samples	S_0	S_1	*(f-)TSST*	S_2	S_3	S_4	S_5	S_6	S_7

After collecting S_0, participants were given 200 ml of grape juice, or glucose water for those with fructose-intolerance, to minimize differences in results due to variations in energy availability [Zän20]. To account for cycle-related differences in cortisol responses to acute stress, female participants additionally provided a passive drool sample to measure progesterone levels [Ham20]. Furthermore, the body weight, body fat, and muscle percentage was documented using a scale, along with overall height. Directly before proceeding to the next step, participants gave their second saliva sample (S_1).

(f-)TSST

Approximately 40 minutes after arrival, participants were guided to a second room where the (f-)TSST, as described in Section 3.1.2, was conducted. All participants were filmed by two separate cameras: an RGB camera capturing the head and face (Sony SRG-300H, Minato, Japan), and a smartphone capturing the body (Google Pixel 7A, Foxconn, Tucheng, Taiwan). Additionally, the smartphone was connected to a microphone (Type C TL350, TONOR, Hong Kong) via Bluetooth to obtain audio footage. The microphone was attached to participants' clothes, mostly around the neckline, to ensure optimal audio quality. Moreover, the phases of the (f-)TSST (see Figure 3.1) were logged by the panel using a smartphone application.

Post-Test Phase

Participants were guided back to the preparation room after completing the (f-)TSST and provided the third saliva sample (S_2: approximately 15min after the start of the (f-)TSST). The remaining saliva samples were taken according to the relative times listed in Table 3.4. Upon providing the final saliva sample, participants were either reminded of their next session for the following day (Day 1) or debriefed and asked to sign a non-disclosure agreement regarding the study setup (Day 2).

3.1.4 Endocrinological Measures

The previously mentioned eight saliva samples, which were taken according to the timeline shown in Table 3.4, were collected with Salivettes (Sarstedt AG & Co. KG, Numbrecht, Germany). The Salivettes containing saliva were kept at room temperature until the end of the session and were then stored in a freezer at a temperature of −18 °C until the end of the study. Later, the cortisol concentrations were extracted from the saliva samples in a laboratory, as described in previous work [Ric21a].

To assess the activity of the HPA axis and confirm the anticipated stress reaction induced by the TSST, three saliva features were calculated from the raw cortisol values obtained by the saliva samples (S_1 - S_7). This included the maximum cortisol increase (Δc_{max}), the "Area under the curve with respect to ground" (AUC_g) as well as the "Area under the curve with respect to increase" (AUC_i) [Pru03]. These features were calculated as follows, where S_i represents the corresponding cortisol level and t_i the time of the measurement. The maximum increase in cortisol (Δc_{max}), which depicts the difference between the highest cortisol level after the (f-)TSST and the cortisol level S_1, measured right before the (f-)TSST:

$$\Delta c_{max} = \text{max}\{S_2, ..., S_7\} - S_1 \tag{3.1}$$

The area under the curve was calculated to determine the quantity of cortisol released as a result to the (f-)TSST. The AUC_g was computed as:

$$AUC_g = \sum_{i=1}^{6} \frac{(S_{i+1} + S_i) \cdot \Delta t_i}{2} \tag{3.2}$$

Furthermore, the AUC_i was calculated using following equation:

$$AUC_i = (\sum_{i=1}^{6} \frac{(S_{i+1} + S_i) \cdot \Delta t_i}{2}) - (6 \cdot S_1) \tag{3.3}$$

3.2 Speech Feature Calculation

3.2.1 Pre-Processing

To process the speech data from the (f-)TSST, the audio channels from the body video recordings were first extracted using *ffmpeg*[1] and stored in .wav files. The audio recordings were then trimmed to focus on the talk segment of the (f-)TSST, excluding the first minute and the last eight minutes of the recording, as these portions approximately cover the segment in which participants talked about their job (f-TSST) or personality (TSST).

Since this trimmed segment contained conversations from both the participant and panel, a speaker diarization algorithm implemented in the *pyannote.audio* Python package was utilized to identify different speakers within the audio [Bre20; Bre21]. The obtained Pandas dataframe included the start and stop times, length, and speaker-ID for each individual segment. Assuming the participant spoke for the majority of the time, the speaker-ID with the most segments was identified as the participant, and therefore all other speaker-IDs were excluded from the dataframe.

Next, the participant's speech segments were isolated and concatenated using *ffmpeg*, based on the start and stop times of the participants speech segments. The resulting audio file contained only the participant's speech. To analyze the lexical aspects of the speech, the audio data was transcribed using *Whipser*, an OpenAI speech recognition model designed for various speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection [Rad23]. The largest model size (large-v3) was used to transcribe the audio data into German text.

In some *Whipser* transcripts, repeated words or sentences were observed, particularly before transitioning to the next speaking segment. These repetitions are a known issue [Rad23] that can occur when the model struggles to accurately interpret certain sections of audio, possibly due to unclear speech or background noise. Since these repetitions could alter the speech features (refer to Table 3.5), any repeated words or sentences that occurred more than three times were consolidated into a single instance to ensure consistency and reliability in the processed transcripts. Furthermore, some transcripts contained errors, however there were no major errors that changed the meaning of the answers. These errors were not corrected manually, as the goal was to assess the feasibility of an automated approach of the linguistic analysis of speech.

Due to some limitations in the speaker diarization process using the *pyannote.audio* tool, certain sentences or parts of sentences were unintentionally excluded from the transcript. This issue likely arose from the algorithm's difficulty in distinguishing between speakers with similar

[1] https://ffmpeg.org/

vocal characteristics or due to overlapping speech. Since this inconsistency was observed across both the TSST and f-TSST, it was deemed acceptable for the purposes of this study.

3.2.2 Feature Extraction

This section outlines the feature extraction process applied to the generated text, focusing on numerical metrics related to lexical, grammatical, and sentiment aspects. A summary of the extracted features is provided in Table 3.5.

To calculate these features, the text was tokenized using *nltk* [Bir09] to determine the number of words and sentences. The tokens were then analyzed and categorized into different word types using *spaCy's* German language model "de_core_news_sm" [Hon17]. Each word type was normalized to assess its frequency relative to the total number of words, ensuring that text length did not influence the results.

Sentiment analysis was conducted using *TextBlob-de* [Lor18], which measures both polarity and subjectivity. Polarity reflects how positively or negatively a text is phrased (e.g., TextBlob("TextBlob is amazingly simple to use. What great fun!") yields a positive polarity score of 0.4). Subjectivity, on the other hand, indicates the degree to which the text expresses personal opinions rather than objective facts, with the same example yielding a subjectivity score of approximately 0.4.

Lexical Diversity

Research has shown that not all measures of lexical diversity (also referred to as lexical richness) are representative of the actual diversity of the text or speech, as most tend to vary on the length of said text and become more accurate with increasing length [Bes23; Tor13]. Therefore, a selected amount of lexical richness features were chosen for the analysis of participants' speech, mostly including features insensitive to text length, as the average text length of participants is approximately 564 words. The features chosen are ttr, mattr, Maas's ttr, mtld, and HD-D and were calculated using the codebase from Shen [She22]. To provide more detail on these features, a brief summary of each will be presented.

ttr is the simplest and in this context likely most variable feature, as it measures the relationship between the number of individual terms (t) and the entire number of words spoken/written (n) 3.4. According to Torruella et al., this feature is rather sensitive to the length of the text and therefore might vary most in the analysis of the participants speeches.

$$ttr = \frac{t}{n} \tag{3.4}$$

A feature which is less sensitive to text length is mattr, which is obtained by moving a window of length (w) tokens along the tokens of the text. The ttr is calculated for each window and the final score is the average of each calculated ttr (refer to equation 3.6). The window length chosen

Table 3.5: Feature overview

Name	Short Form
Number of words	num_words
Number of sentences	num_sentences
Average sentence length	mean_sentence_length
SD of sentence length	std_sentence_length
Number of nouns	num_nouns
Nouns in comparison to all words	nouns/words
Number of verbs	num_verbs
Verbs in comparison to all words	verbs/words
Number of adjectives	num_adj
Adjectives in comparison to all words	adj/words
Number of adverbs	num_adv
Adverbs in comparison to all words	adv/words
Number of articles	num_art
Articles in comparison to all words	art/words
Number of pronouns	num_pron
Pronouns in comparison to all words	pron/words
Number of conjunctions	num_conj
Conjunctions in comparison to all words	conj/words
Number of prepositions	num_prep
Prepositions in comparison to all words	prep/words
Number of numbers	num_num
Numbers in comparison to all words	num/words
Average sentiment (polarity)	mean_sent_polarity
SD of sentiment (polarity)	std_sent_polarity
Average sentiment (subjectivity)	mean_sent_subjectivity
SD of sentiment (subjectivity)	std_sent_subjectivity
type-token-ratio (ttr)	ttr
Maas' ttr	maas
Moving average type-token-ratio (mattr)	mattr
Measure of textual lexical diversity (mtld)	mtld
Hypergeometric Distribution Diversity (HD-D)	hdd

for this thesis is w = 70, as the default window size in Shen's code of w = 100 [She21; She22] is too large with an average text length of 564 words per participant and the recommended length of w = 50 by Bestgen [Bes23] is for research on language learning.

$$mattr = \frac{1}{n-w+1} \sum_{i=1}^{n-w+1} ttr \tag{3.5}$$

Another ttr was developed by Maas which also is insensitive to text length and is calculated as follows:

$$maas = \frac{\log n - \log t}{\log^2 n} \tag{3.6}$$

The mtld is computed as the mean length of sequential words in a text while maintaining a minimum threshold ttr score. The recommended threshold score is between 0.66 and 0.72, for the analysis of the participants text a threshold of 0.68 was used, as the mean ttr was approximately 0.43 .

HD-D provides an estimate of how varied the vocabulary in a text is, without being too heavily influenced by how long the text is. It computes the probability of a term appearing at least once in a random draw size (default value draws = 42) and does this for each term of a text. This lexical richness measure is the most recommended, as it is least sensitive to text length, while providing an accurate and uniform diversity value [Bes23].

3.3 Evaluation

3.3.1 Statistics

For the statistical evaluation a repeated measure analysis of variance (ANOVA) was employed, with paired *t*-test as *post-hoc* comparison. Since all participants were exposed to both TSST and f-TSST, *condition* was used as within-variable, for which the control and intervention samples are interdependent. Due to the exploratory nature of the study the only multi comparison correction applied was the Benjamini-Hochberg method [Ben95], designed to control the False Discovery Rate, reducing the likelihood of false positives in multiple comparisons. All statistical analyses were performed using the Python package *biopsykit* [Ric21b] based on *pingouin* [Val18]. Effect sizes for *t*-tests were reported using Hedge's g.

3.3.2 Classification

To differentiate stress from non-stress conditions, ML models were trained using the speech features derived from the transcripts of the (f-)TSST video recordings . Classification was performed using a standard ML pipeline, as depicted in Figure 3.2.

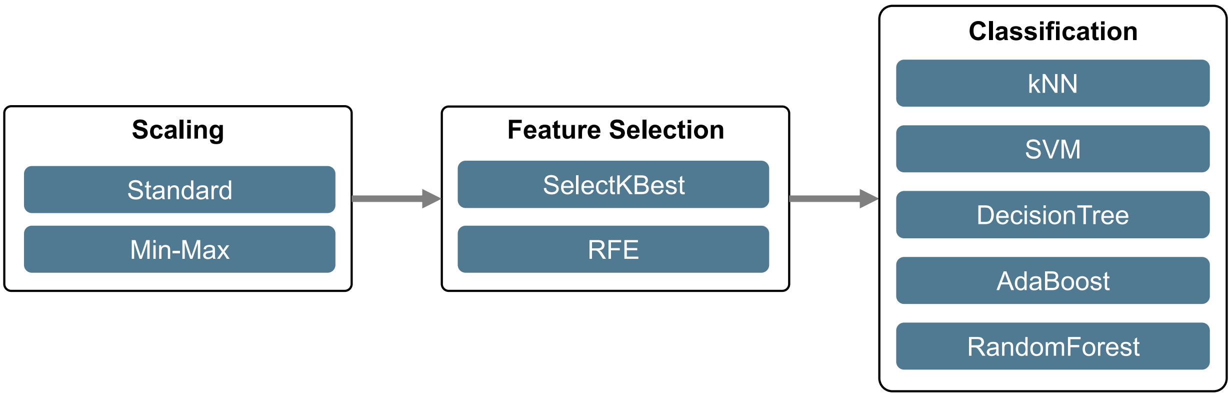

Figure 3.2: Trained classification pipeline.

For the classification task, five models were trained: *k-nearest neighbors (kNN)*, *SVM*, *DecisionTree*, *AdaBoost*, and *RandomForest*. Each model's performance was assessed using a five-fold cross validation (CV) based on mean test accuracy. During each iteration of CV, hyperparameters for both feature selection and the classifiers were optimized using a five-fold CV with grid search (random search with 100 iterations for *RandomForest*).

First, features were scaled in each CV fold using either *StandardScaler* or *MinMaxScaler*. *StandardScaler* applied z-score normalization to standardize feature distributions to a mean of 0 and a standard deviation of 1, while *MinMaxScaler* mapped features to a 0-1 range.

Next, dimensionality reduction was performed with *SelectKBest* or *recursive feature elimination (RFE)*. *SelectKBest* selected the top k features with the highest ANOVA F-score, while *RFE* recursively eliminated the least important features, as determined by SVM, until the desired feature count was reached. Table 3.6 summarizes the hyperparameter optimization space for each model. This CV approach has been validated in prior studies [Abe22].

3.3.3 ML-Based Regression

Following a similar approach to the classification task, machine learning-based regression was employed, with the objective of predicting the continuous endocrine measure, Δc_{max}, rather than categorical condition labels. A five-fold CV was used to assess model performance, with an additional five-fold CV nested within each fold to optimize hyperparameters for both feature selection and the regressors, aiming for the highest R^2 value. The models were implemented in their respective regression forms: kNN Regressor (kNNRegressor), Support Vector Regressor (SVR), Decision Tree Regressor, AdaBoost Regressor, and Random Forest Regressor. Table 3.6 presents the hyperparameter grid, along with the scaling and feature selection methods, consistent with those used in the classification task.

Table 3.6: Hyperparameter grid; [1] only for RBF kernel; [2] RandomizedSearch was used for RandomForest

Feature Selection	**Hyperparameter**	**Values**
SelectKBest	k	2 to 4; steps of 2; all
RFE	n	2 to 4
Classifier/Regressor	**Hyperparameter**	**Values**
kNN	k	2 to 10; steps of 2
	weights	uniform, distance
SVM	kernel	linear, RBF
	C	$10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$
	gamma [1]	$10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$
DecisionTree	criterion classification	gini, entropy
	criterion regression	squared_error
	depth	2 to 10; steps of 2
	min_samples_split	2 to 5; steps of 1
	min_samples_leaf	1 to 4; steps of 1
AdaBoost	base_estimator	DecisionTree
	n_estimators	20 to 110; steps of 10
	learning_rate	0.6 to 1.1; steps of 0.1
RandomForest [2]	bootstrap	True
	n_estimators	50, 100, 150, 200
	max_depth	None, 2, 4, 6, 8, 10
	min_samples_split	2 to 5; steps of 1
	min_samples_leaf	1 to 4; steps of 1
	max_features	sqrt, log2

Chapter 4

Results & Discussion

This Chapter outlines the outcomes of the statistical analysis of endocrinological measures and speech features, which demonstrated promising results. In addition, the Chapter includes a comprehensive discussion of these outcomes. Furthermore, the performance of the classification and regression models is presented and interpreted, along with an analysis of their limitations.

4.1 Stress Response Assessment

Figure 4.1 illustrates the cortisol response across conditions, indicating a nearly twofold mean increase in cortisol levels following the TSST. Specifically, mean cortisol concentrations surged by 99 % from the baseline (*S1*) to the expected peak approximately 25 minutes post-TSST start (*S3*). In contrast, the f-TSST elicited a smaller, though still notable, increase of 37 %.

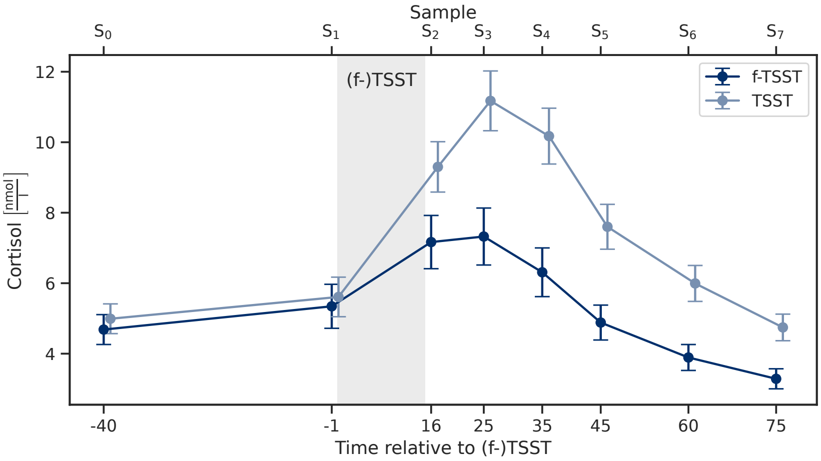

Figure 4.1: Cortisol response

The derived cortisol features, presented in Figure 4.2, show a marked increase in cortisol levels after the TSST compared to the f-TSST. Notably, the feature Δc_{max} exhibited the largest effect size, with an average increase of 4.12 nmol/l following the TSST compared to the f-TSST. Additionally, AUC_g showed an increase of nearly 50 % following the TSST.

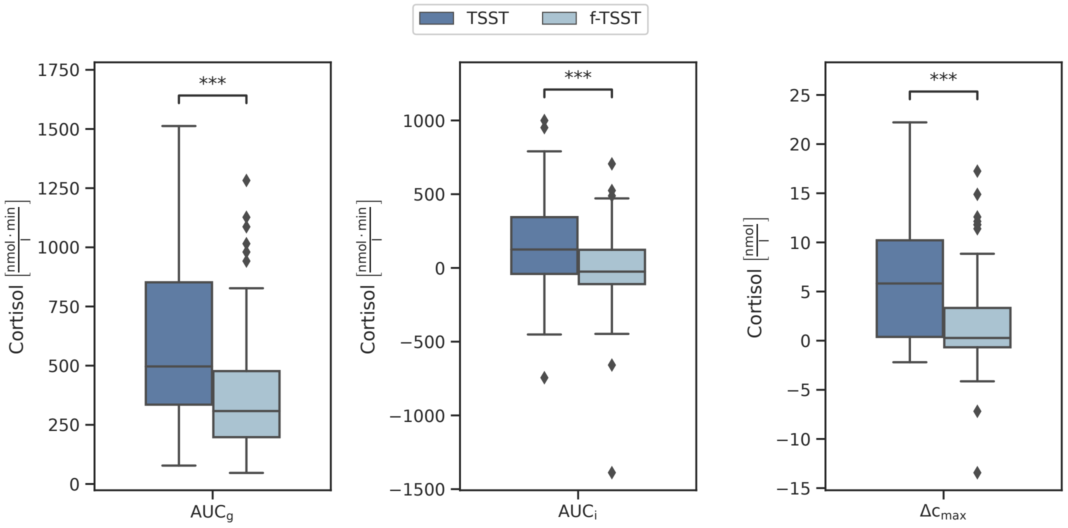

Figure 4.2: Cortisol features

The statistical analysis of these cortisol features, detailed in Table 4.1, reveals medium to high effect sizes, aligning broadly with previous research findings [Wie13]. However, a considerable number of outliers, particularly in Δc_{max} and AUC_g, were observed in the f-TSST group (see Figure 4.2), indicating that several f-TSST sessions still elicited notable stress responses.

Table 4.1: t-test results of cortisol features; $^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Feature	$t(53)$	p	Hedges' g
AUC_g	6.016	<0.001***	0.629
AUC_i	3.748	<0.001***	0.549
$\Delta\ c_{max}$	4.742	<0.001***	0.705

To further explore this, participants were categorized as "responders" or "non-responders" based on two different thresholds of maximum cortisol increase, in order to assess how the distribution of stress responses corresponded to the conditions. Wiemers et al. proposed a threshold of 2.5 nmol/l to classify a participant as a "responder," indicating a stress response. Using this threshold, they reported an 83 % stress response rate in the TSST, while the f-TSST induced a stress response in only 9 % of participants [Wie13]. On the other hand, Miller et al. suggested a lower threshold of 1.5 nmol/l as a more accurate indicator, categorizing over 90 % of participants in the TSST group as ”responders”, and less than 1 % in the f-TSST group [Mil13]. These thresholds were applied to the cortisol data from the participants in this study, with the results illustrated in Figure 4.3.

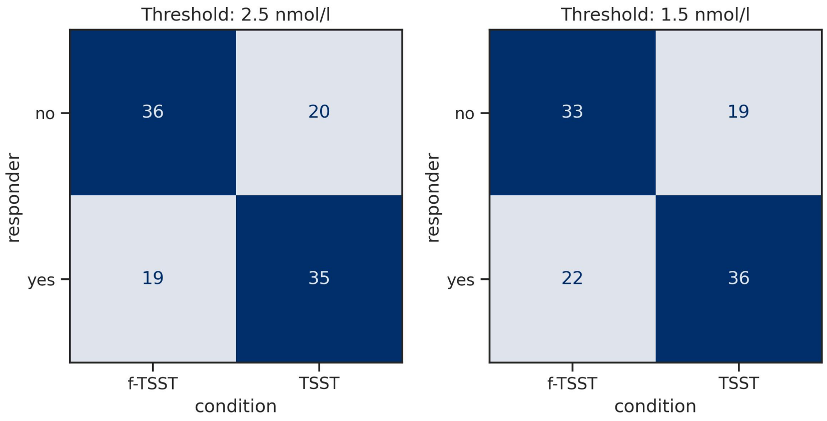

Figure 4.3: Confusion matrices showing correlation between condition and cortisol responders using different thresholds

The 2.5 nmol/l threshold resulted in a slightly more accurate alignment between condition and stress response, with 63.6 % of participants exhibiting a stress response after the TSST and 34.5 % after the f-TSST. This threshold produced a nearly balanced distribution of 56 ”responders” and 54 ”non-responders” among the 55 participants who underwent both conditions. The lower threshold of 1.5 nmol/l, as anticipated, identified a higher percentage of ”responders”, with 65.4 % for the TSST and 40 % for the f-TSST. However, this threshold resulted in a less balanced distribution of 58 ”responders” and 52 ”non-responders”, indicating that the 2.5 nmol/l threshold provided a more precise alignment for this study.

Nevertheless, the overall correlation between experimental conditions and stress responses was less robust than in the referenced studies. This variance may be due to subtle differences in the study design, such as having two female panel members or incorporating the arithmetic task in the f-TSST, which could have influenced the stress response outcomes.

4.2 Speech Feature Evaluation

Given the extensive number of features extracted from the transcripts, only the most relevant ones are discussed in this section, with detailed statistical results available in Table B.1 of Appendix B.

Among the numerous features analyzed, the ten most prominent, identified by the highest t-values, are illustrated in Figure 4.4. Notably, the most significant differences between conditions were observed in average sentiment polarity. The mean polarity score for the f-TSST was 0.14, contrasting with the TSST at 0.34, indicating that participants used more positively inclined words in the TSST. A closer examination of this feature revealed that participants often began their speech in the TSST with a positive tone, for instance, by expressing that they were ”happy to be here”. Moreover, words like ”very” and ”good” were frequently used in the TSST, often in the context of describing their personality or suitability for a job. Specifically, the word ”very” ranked among the 30 most common words for 32 out of 40 participants in the TSST, compared to only 17 participants in the f-TSST. Similarly, the word ”good” appeared in the top 30 words for 21 participants in the TSST, but for only 2 participants in the f-TSST, likely contributing to the higher positive sentiment score during the stress test. Further analysis of sentiment polarity showed changes in the timeline of the tests: the average polarity score in the first half of the TSST transcript was approximately 0.38, which decreased to 0.31 in the second half, demonstrating a decline in positive sentiment as the speech progressed. A similar pattern was observed in the f-TSST, where the polarity score dropped from an average of 0.2 in the first half to 0.1 in the second half.

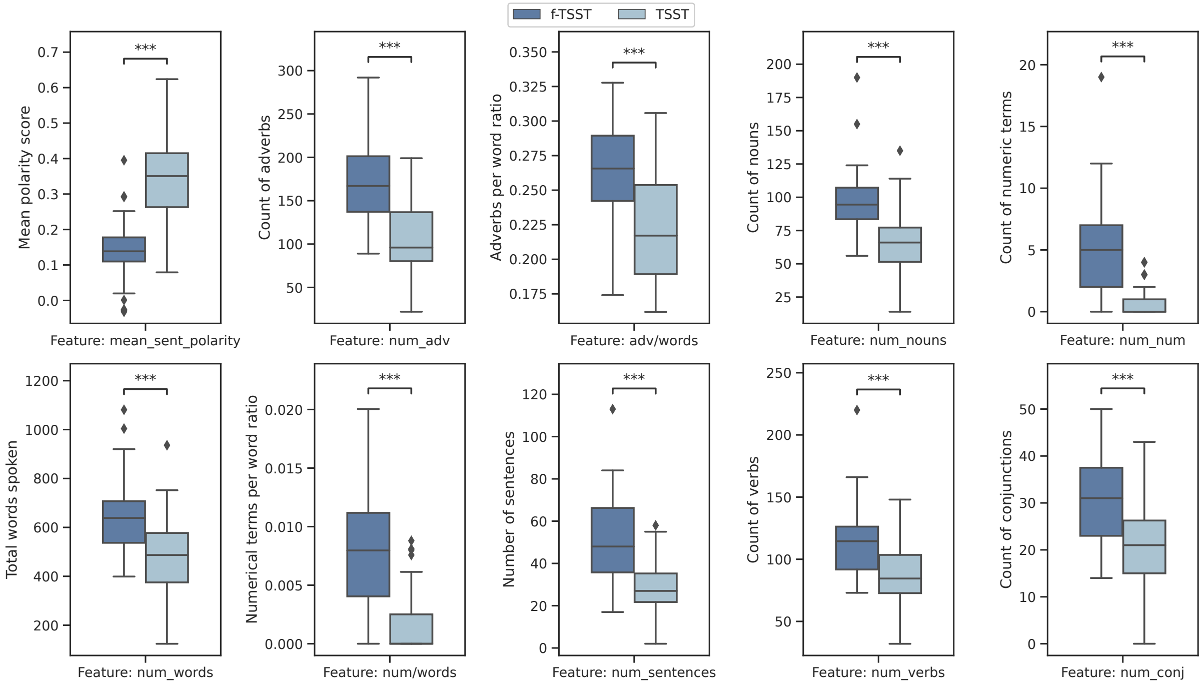

Figure 4.4: Best performing features (highest absolute t-values)

Additionally, participants used more numerical terms during the f-TSST, as reflected in both the absolute count and the ratio of numerical terms relative to the total word count. The number of words, and consequently the number of sentences, also decreased in the TSST compared to the f-TSST, with the average word count in the f-TSST being 652 ± 147, while in the TSST, it was only 478 ± 164. Furthermore, a notable difference was observed in the use of adverbs between the two conditions, with both the absolute number of adverbs and the ratio of adverbs to words decreasing from the control condition to the stress test. The average adverb to word ratio was 26 % ± 4 % during the f-TSST and 22 % ± 4 % during the TSST. Figure A.1 in Appendix A illustrates the correlation of the top 24 speech features between the TSST and f-TSST.

Gender Differences

In the analysis of gender differences, only minor variations were observed. On average, women spoke less during the TSST compared to men, with men's mean word count decreasing by 100 from the f-TSST to the TSST, while women's count dropped by 200. However, women's average sentence length increased by over 50 % in the TSST, whereas men's length increased by only 15 %.

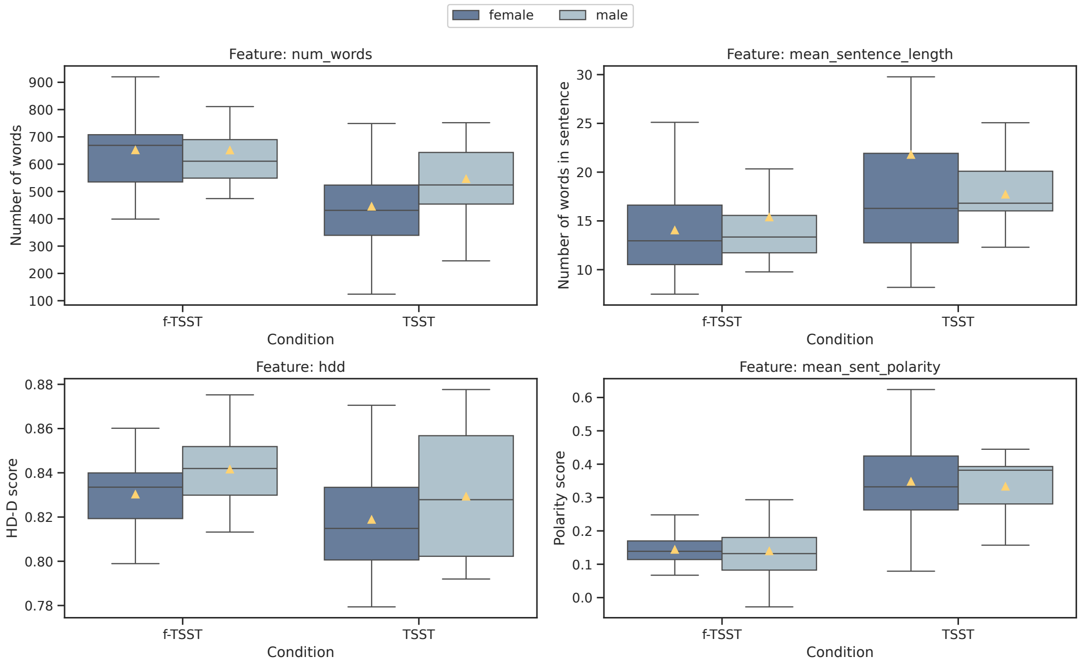

Figure 4.5: Features showing gender differences; yellow triangle represents mean value; plots without outliers

Additionally, women exhibited a slightly lower mean mtld and HD-D in both conditions and demonstrated a marginally more positive sentiment, aligning with Blasberg et al.'s findings on "tend and befriend" behaviors in stressful situations [Bla23]. These results are represented in Figure 4.5. As a final note, the single non-binary participant was excluded from this analysis, as a statistical evaluation based on a single individual would not provide meaningful or reliable insights.

Discussion

The speech characteristics observed during the TSST align with previous studies regarding word and sentence count, suggesting a reduction in fluency under stress [Buc14; Bel21]. While fluency was not directly measured, the reduction in word count may indicate decreased fluency, potentially due to a "freezing" response similar to other stress-induced behaviors [Roe10; Ric24]. Alternatively, participants may struggle to form coherent speech due to the restrictive and less interactive nature of the TSST, which limits topics to personality traits, excluding more concrete areas like work history or achievements. The lack of conversational dynamics further constrains participants, possibly making it more difficult for them to articulate coherent speech. It is also important to note that fluency is not solely determined by word count. Buchanan et al. included an analysis of filler words, which this study did not explore, as *Whisper* only transcribed them in a few instances. The omission of filler words might have affected the total word count and should be considered in future research. Buchanan et al. suggested that filler words were used more frequently during the p-TSST, potentially further impacting word count and fluency. Therefore, future research should incorporate a more direct measurement of fluency, such as tracking the number of words spoken per minute or within a specified time span, to verify whether the observed reduction in word count under stress correlates with decreased fluency.

Contrary to previous research [Chy23], sentiment increased significantly during the TSST, resulting in a more positive tone compared to the almost neutral sentiment observed in the f-TSST. This could be attributed to the nature of the TSST, where participants were in a job interview setting, prompting them to emphasize their positive personality traits, despite also being asked to discuss their negative traits. The relatively high SD in sentiment polarity during the TSST might suggest an initial positivity that diminished as stress levels increased. This suggestion is further supported by the specific language choices made by participants during the TSST. For instance, words such as "very" and "good" were frequently used at the beginning of the interview, indicating an effort to create a favorable first impression. This is corroborated by the more positive sentiment observed in the first half of the transcripts compared to the latter half. The decline in sentiment observed in both the TSST and f-TSST could have multiple explanations. One possibility is that participants naturally begin with a friendly tone, which then neutralizes over time. Another plausible explanation is that as cortisol levels rise, sentiment becomes increasingly negative, potentially indicating that the f-TSST also induced a certain level of stress. This notion is supported by the relatively high number of responders — 19 out of 35 — (see Figure 4.3) and the 37 % increase in mean cortisol levels in the f-TSST (refer to Figure 4.1).

Other noteworthy features, such as the frequency of numbers and adverbs, likely reflect the content constraints of the tests. For instance, the f-TSST, which involves discussing one's resume, naturally leads to the mentioning of dates and time periods, resulting in a higher number count compared to the TSST, where the focus on personality traits does not typically involve numerical references. The increased use of adverbs during the f-TSST might also be attributed to the broader range of topics covered, while the TSST restricts discussion to personality traits. However, this difference could suggest that participants have fewer ways to describe themselves under stress.

Regarding the results of lexical diversity, the ttr was found to be higher during the TSST, suggesting greater diversity, however, as discussed in Section 3.2.2, ttr is sensitive to text length, and given that the text produced during the TSST is shorter than that of the f-TSST, this metric may be misleading. In contrast, the HD-D, which is less affected by text length and a generally more recommended measure of lexical richness, indicates a slightly higher diversity during the f-TSST, implying that participants may have experienced difficulty finding words under stress. Notably, this observation contrasts with the findings of Buchanan et al., who reported an increase in communication rate and word productivity during the TSST [Buc14]. However, since the measures used in both studies are not directly comparable, the apparent discrepancy may not be directly related.

4.3 Classification

The best-performing pipeline achieved a mean accuracy of 95.0 % ± 7.3% in classifying between conditions using speech features. The parameters for this pipeline were as follows:

- Scaling: StandardScaler
- Feature Selection: RFE
- Classification: RandomForestClassifier

Figure 4.6 presents the confusion matrix of the best-performing pipeline, underlining its high accuracy. Table 4.2 summarizes the performance of the other pipelines, which show consistent

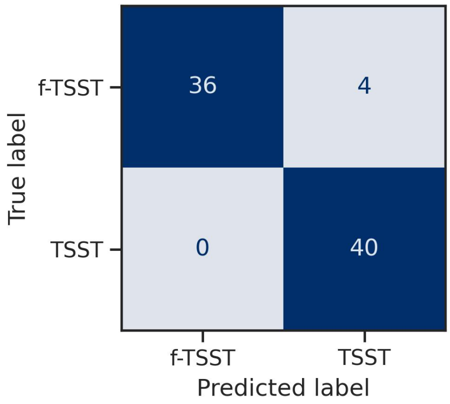

Figure 4.6: Confusion matrix condition classification

results, reflected in the small standard deviations, suggesting a reliable prediction and reducing concerns about overfitting despite the relatively small sample size.

Table 4.2: Comparison of different ML pipelines; accuracy ± SD; best performing pipeline per classifier highlighted

Feature Selection	RFE		SelectKBest	
Scaling	MinMax	Standard	MinMax	Standard
Classification				
KNeighborsClassifier	**92.5** ± **7.3**	91.2 ± 6.4	88.8 ± 7.3	91.2 ± 6.4
SVC	**93.8** ± **6.8**	92.5 ± 4.7	91.2 ± 7.5	91.2 ± 6.4
DecisionTreeClassifier	**86.2** ± **9.2**	85.0 ± 8.5	83.8 ± 8.5	**86.2** ± **9.2**
AdaBoostClassifier	**91.2** ± **6.4**	87.5 ± 6.8	88.8 ± 4.7	87.5 ± 5.6
RandomForestClassifier	92.5 ± 10.0	**95.0** ± **7.3**	90.0 ± 7.5	91.2 ± 6.4

These promising results align with the distinctive features identified in the statistical analysis, such as *mean_sent_polarity*, *num_words*, and *num_num*. However, it is crucial to consider the study's limitations, including the small sample size and the specific nature of the stress tests. To address these limitations, additional classifications were performed using speech features that were not correlated with the number of words, focusing solely on relative features. The confusion matrices for the two best-performing models, namely the *KNeighborsClassifier* (utilizing either *StandardScaler* or *MinMaxScaler* with *SelectKBest*) and the *RandomForestClassifier* (using *StandardScaler* with *SelectKBest*), are presented in Figure 4.7. These models achieved a mean accuracy of 93.75 % ± 5.59%. The specific features utilized and their respective relevance to the classification are illustrated in Figure 4.8.

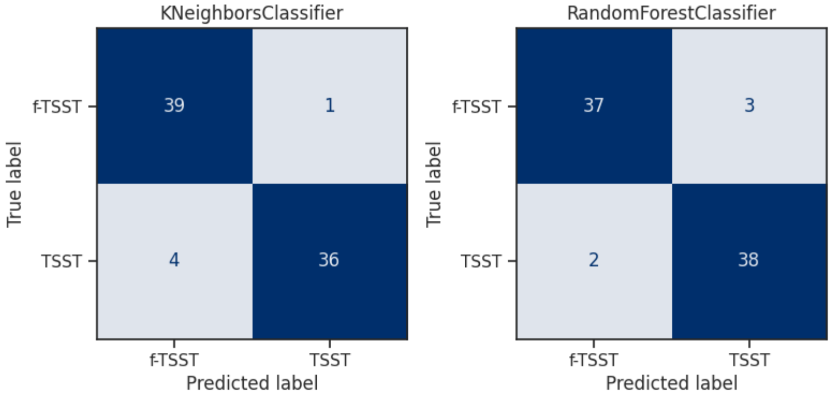

Figure 4.7: Confusion matrices of best condition classification pipelines using relative features

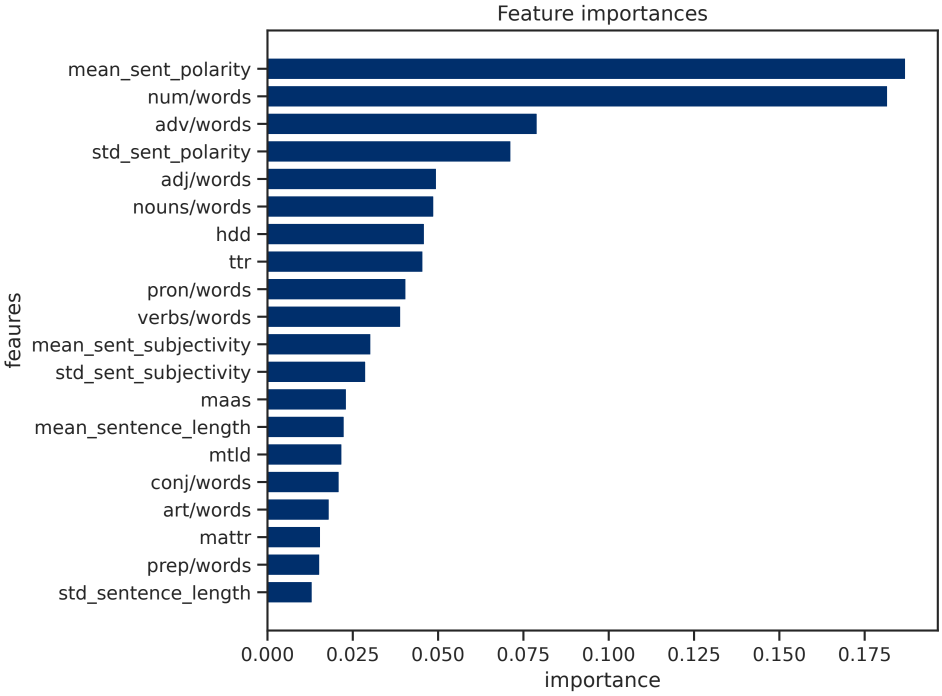

Figure 4.8: Feature importances of *RandomForestClassifier* based on mean decrease in impurity

Given the persistently high accuracy, which raised concerns that the model might still be relying on the testing protocols rather than actual stress indicators, further classification was conducted using cortisol levels as an indicator of stress. For this analysis, the condition labels were adjusted from ”TSST” and ”f-TSST” to ”responders” or ”non-responders”, based on the participant’s cortisol response, specifically the peak increase (Δc_{max}). While Millers et al. study suggest that a salivary cortisol increase of 1.5 nmol/l is a reliable marker of stress, Wiemers et al. threshold of 2.5 nmol/l proved to be better suited for this study, as stated in Section 4.1. Using this threshold, 32 ”responders” and 44 ”non-responders” were labeled, with two participant’s data excluded due to missing cortisol levels. This reclassification yielded a mean accuracy of 62.68 % ± 11.09 % using the best-performing pipeline, which comprised *MinMaxScaler*, *SelectKBest*, and *KNeighborsClassifier*. The corresponding confusion matrix is depicted in Figure 4.9. Due to the relatively uneven distribution of this labeling, a classification using the lower threshold of 1.5 nmol/l for responder categrization was used, as the distribution was slightly more even with 34 ”responders” and 42 ”non-responders”. However, this classifications highest accuracy was worse at only 57.5 %.

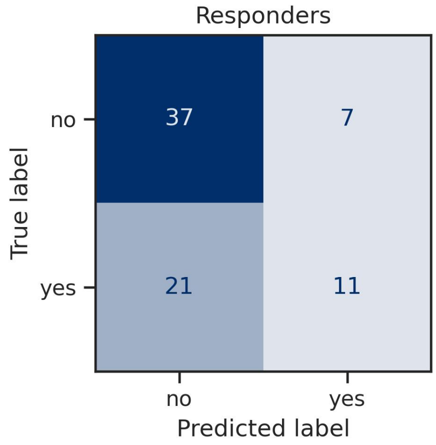

Figure 4.9: Confusion matrix depicting responder classification

To assess whether the 62.68 % accuracy was achieved by chance, a randomized condition categorization was performed, resulting in a maximum accuracy of 58.93 %. Given the proximity of this value to the observed 62.68 % accuracy for ”responders”, a comprehensive analysis of all classification results was conducted and compared. Figure 4.10 illustrates the comparison across different classification tasks, revealing that condition categorization was the most accurate. Participants labeled as ”responders” — categorized by a cortisol increase threshold of 2.5 nmol/l — were classified more accurately than randomized labels. Therefore, although the highest accuracies for both ”responder” and randomized classifications were similar, the average accuracy for ”responder” classification was notably higher. This finding suggests that, despite the less-thanideal alignment between ”responders” and test conditions, the model was still more effective at detecting cortisol increases above the threshold than it was at classifying random labels.

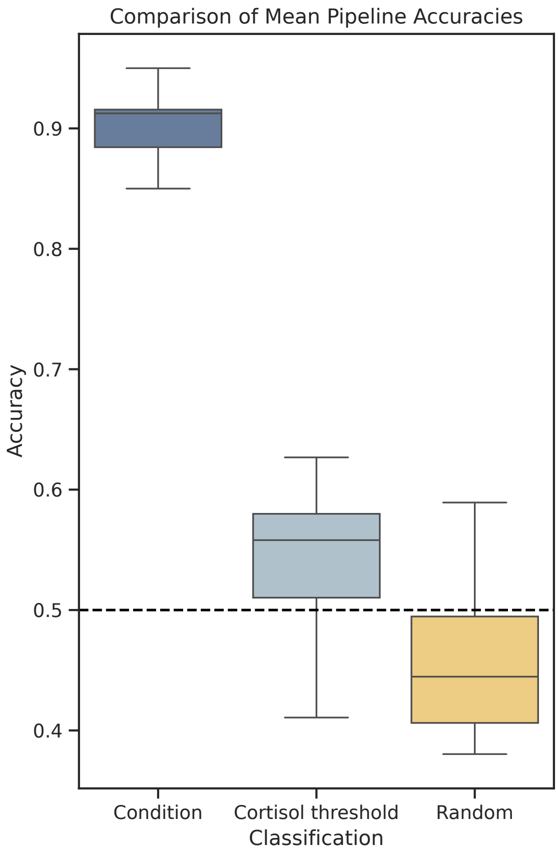

Figure 4.10: Comparison of mean classification accuracies of condition, cortisol threshold and random assignment

To further evaluate whether lexical speech truly reflects stress, these findings will be cross-referenced with cortisol levels in Section 4.4 through regression analysis.

4.4 ML-based Regression

The regression model developed to predict the maximum cortisol increase (Δc_{max}) based on the selected features yielded negative R^2 values, with the highest R^2 being -0.108. Negative R^2 values suggest that the model is performing worse than a simple horizontal line representing the mean of the target variable. This indicates that the chosen features and model configuration were not effective in capturing the underlying patterns or relationships necessary to accurately predict cortisol increases.

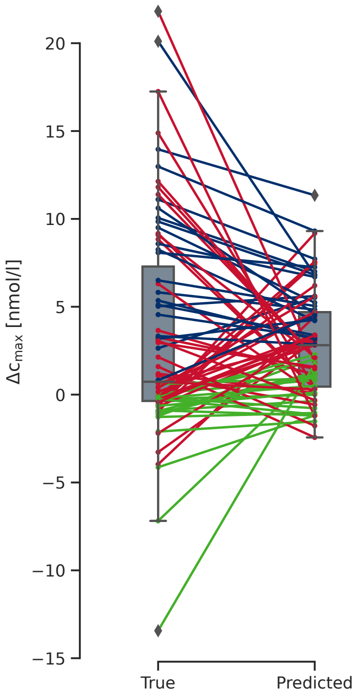

Figure 4.11: Regression results Δc_{max} ; blue: true and predicted value over the respective median; green: true and predicted value under the respective median; red: true and predicted values not on the same side of the respective median

To assess whether the model's predictions were at least somewhat aligned with the expected range, a further analysis was conducted, as shown in Figure 4.11 The objective was to determine whether samples with true Δc_{max} values above the median had corresponding predicted values also above the median, and similarly for values below the median. However, as illustrated in Figure 4.11, this analysis yielded similarly insufficient results. This suggests that speech features alone may not be able to reliably indicate changes in cortisol levels.

4.5 General Discussion & Limitations

The findings of this study suggest that digital biomarkers derived from speech can effectively predict the condition, using both absolute and relative speech features. The statistical analysis revealed clear distinctions in speech patterns between the TSST and f-TSST, which results in the high classification accuracy achieved. However, when the focus shifts to classifying actual stress or predicting cortisol increase, the model's accuracy diminishes. This decline could be attributed to the test protocol, which restricts and guides participants' speech, or to the inherent variability in cortisol levels, likely a combination of both factors.

The statistical analysis highlights three significant features for classification: average sentiment polarity, total word count, and the frequency of numerical terms relative to total words spoken. Notably, the finding that sentiment was more positive during the TSST contradicts previous research, which typically associates stress with negative sentiment [Chy23]. This divergence may be attributed to the specific test protocol. The polarity score tends to decrease over the course of the TSST, possibly due to the delayed onset of cortisol release as stress levels build up progressively. The reduction in word count observed during the TSST, which could correlate with decreased fluency as seen in earlier studies [Buc14], may result from the restrictive nature of the TSST or a stress-induced ”freezing” response [Roe10; Ric24]. The prominence of numerical references is clearly tied to the protocol differences, as the f-TSST allows a broader range of topics, including dates and numbers, whereas the TSST focuses narrowly on personality traits. This suggests that these features may be more reflective of the protocol itself rather than genuine stress responses.

Moreover, while Figures 4.1 and 4.2 demonstrate differences in cortisol levels between the test conditions, these differences represent averages and vary significantly across individuals. This variability is evident in the confusion matrices shown in Figure 4.3, where 34.5 % of cases were classified as false positives (i.e., participants undergoing the f-TSST were categorized as ”responders”) and 36.4 % as false negatives (i.e., participants undergoing the TSST were categorized as ”non-responders”) using Wiemers et al. threshold of 2.5 nmol/L cortisol increase. For comparison, Wiemers et al. reported 9 % false positives and 17 % false negatives using the same threshold [Wie13]. These results suggest that modifications to the study setup may be necessary to ensure more accurate stress induction.

A broader limitation is the relative novelty of the f-TSST, which has not been extensively utilized in previous research, resulting in limited expertise on standardizing this protocol. The f-TSST was designed to modify the p-TSST and create a more comparable control condition for the TSST by incorporating a simpler arithmetic task. As a result, the timeline and cognitive demands of both conditions are similar, though the f-TSST was not intended to induce stress. Nevertheless, as shown in Figure 4.1, a smaller but measurable cortisol spike was observed on average, which may stem from the design of the control condition. While the alternating addition of 10 and 20 is relatively simple, performing this task in front of a panel could still be stressful for participants less confident in mathematics. Moreover, public speaking, inherently stressful for many, was required in the f-TSST, unlike the p-TSST, which may have contributed to the cortisol increase. Hence, further exploration of different control conditions for the TSST is warranted to expose participants to similar tasks without triggering a significant HPA axis response. Additionally, for research into lexical changes in speech due to stress, alternative study designs or stress tests may be required. The current protocols, with their distinct conversation topics, lead to speech features that reflect the protocol rather than stress itself. A study design that employs identical conversation topics across varying environments may better isolate speech changes associated with stress.

Another limitation involves the reliability of *Whisper's* transcription and *pyannote's* speaker diarization, which were not entirely accurate. Some speech segments were cut off or incorrectly transcribed, indicating a need for improvement in these tools. Moreover, even if stress could be identified exclusively from speech, real-time stress recognition remains unfeasible with current technology. For instance, processing a six-minute audio segment took approximately 30 minutes on the available hardware.

In summary, while the current study highlights both the potential and limitations of using speech features for stress recognition, it sets the stage for future research aimed at improving accuracy and understanding the patterns of stress responses. By addressing these limitations and refining methodologies, subsequent studies can build on these findings to develop more effective tools for digital stress assessment.

Chapter 5

Conclusion & Outlook

This bachelor thesis represents an initial exploration into detecting stress through lexical and grammatical aspects of speech. The significance of this research lies in the potential for real-time, contactless stress detection, which could have wide-reaching implications for public health and well-being. By incorporating features from previous studies, such as lexical diversity and sentiment analysis, the study achieved promising statistical and classification results, with mean accuracies exceeding 93 %. These findings suggest that speech features could potentially serve as digital biomarkers for stress detection.

The TSST, a well-established standard for inducing psychosocial stress, was utilized for this research. Despite minor protocol modifications, such as the use of an all-female panel, the TSST proved to be a generally reliable tool for stress induction. The f-TSST served as a viable control condition, maintaining relatively low activation of the HPA axis, as indicated by endocrine level analysis. However, when categorizing participants as “responders” and “non-responders” based on a cortisol increase threshold, which is recognized for its sufficient accuracy, the alignment of conditions and cortisol response was approximately 65 %, indicating that refinements to the stress induction methodology may be necessary.

The study identified sentiment polarity, overall word count, and the ratio of numbers and adverbs to total words as the most significant features for classifying conditions. While these findings align with some aspects of previous research, such as the reduction in fluency under stress, as evidenced by the decreased word count, they also present contradictions, particularly regarding the increase in positive sentiment during stress. This suggests that while the model was effective at classifying test conditions, it may have been influenced by the experimental protocol rather than actual stress responses. This hypothesis is supported by the regression analysis, which indicated that speech features alone were insufficient for accurately predicting cortisol levels.

No significant gender differences were observed in the study, though minor trends included slightly more positive sentiment and reduced word count under stress for women. These observations suggest that while gender may influence certain aspects of speech under stress, these differences were not pronounced enough to impact the overall classification accuracy.

In summary, while the classification model achieved a high accuracy of 95.0 % in categorizing the test conditions, its performance in identifying ”responders” was notably lower, with an accuracy of 62.7 %. The regression analysis was even less promising, as it demonstrated that the model was unable to predict cortisol levels effectively. These findings suggest that the TSST and f-TSST may not correlate strongly with the discrete categories of ”responders” and ”non-responders,” and that the model primarily learned to differentiate between conditions rather than detect physiological stress responses. This indicates that speech features alone may not be sufficient to capture changes in cortisol levels.

For near future research in this area, several enhancements are recommended. First, the study design should be refined to better isolate and differentiate stress responses in speech. One promising approach could involve standardizing conversational topics across conditions, with one condition inducing stress and the other serving as a neutral baseline. This would reduce the likelihood of the model classifying conditions based on specific word types, such as numbers, that are tied to particular topics rather than stress itself. Second, it is essential to improve the accuracy and processing speed of speech transcription and speaker diarization tools, such as *Whisper* and *pyannote*, to advance toward real-time stress detection capabilities. Lastly, exploring a multimodal approach could significantly enrich the analysis. Integrating prosodic and phonetic features, including general vocal characteristics, has shown promising results in previous research. Extending the range of speech features to include measures of filler words and fluency may also be beneficial. Moreover, combining speech features with other stress biomarkers, such as heart rate and heart rate variability via smartwatches or facial expression analysis through cameras, could provide a more comprehensive assessment by capturing multiple dimensions of the stress response.

The long-term future objective is to achieve real-time, contactless stress detection using digital biomarkers, ideally integrating voice and speech analysis. Building on the promising results from previous studies, this approach could facilitate the identification of stress in everyday settings, enabling timely interventions and mitigating the risks associated with prolonged or chronic stress.

Appendix A

Additional Figures

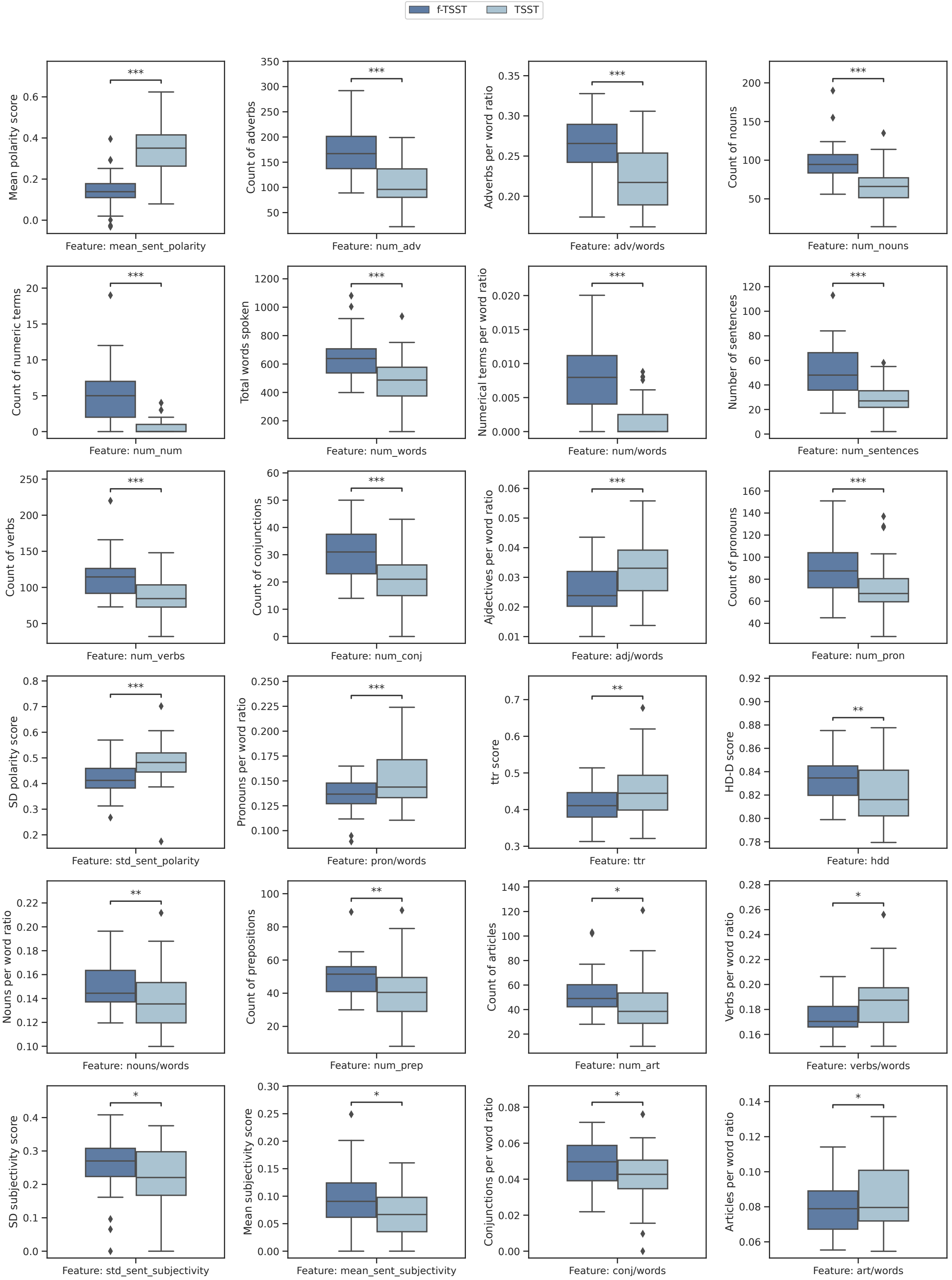

Figure A.1: Statistical analysis feature box-plots

Appendix B

Additional Tables

Table B.1: t-test results of all speech features

Feature	$t(39)$	p	Hedges' g
adj/words	-5.108	$<0.001^{***}$	-0.815
adv/words	7.754	$<0.001^{***}$	1.027
art/words	-2.240	0.040^{*}	-0.415
conj/words	2.334	0.034^{*}	0.463
hdd	3.406	0.003^{**}	0.524
maas	-1.667	0.119	-0.221
mattr	0.455	0.674	0.065
mean_sent_polarity	-9.738	$<0.001^{***}$	-1.883
mean_sent_subjectivity	2.361	0.033^{*}	0.497
mean_sentence_length	-1.900	0.077	-0.410
mtld	-0.189	0.851	-0.033
nouns/words	3.081	0.007^{**}	0.506
num/words	6.911	$<0.001^{***}$	1.515
num_adj	0.744	0.493	0.099
num_adv	9.095	$<0.001^{***}$	1.403
num_art	2.727	0.015^{*}	0.460
num_conj	6.342	$<0.001^{***}$	1.074
num_nouns	7.419	$<0.001^{***}$	1.227
num_num	7.383	$<0.001^{***}$	1.643
num_prep	2.965	0.009^{**}	0.587
num_pron	4.999	$<0.001^{***}$	0.741
num_sentences	6.880	$<0.001^{***}$	1.329
num_verbs	6.709	$<0.001^{***}$	0.981
num_words	7.366	$<0.001^{***}$	1.104
prep/words	-1.960	0.071	-0.393
pron/words	-3.893	$<0.001^{***}$	-0.777
std_sent_polarity	-4.028	$<0.001^{***}$	-0.921
std_sent_subjectivity	2.444	0.028^{*}	0.544
std_sentence_length	-1.095	0.310	-0.228
ttr	-3.584	0.002^{**}	-0.527
verbs/words	-2.709	0.015^{*}	-0.637

List of Figures

List of Tables

Bibliography

[Abe22] Luca Abel. “Machine Learning-Based Detection of Acute Psychosocial Stress from Dynamic Movements”. PhD thesis. Master Thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg, 2022. url …, 2022.

[Aig15] Jonathan Aigrain, Séverine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani. “Person-specific behavioural features for automatic stress detection”. In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 3. IEEE. 2015, pp. 1–6.

[Alm21] José Almeida and Fátima Rodrigues. “Facial Expression Recognition System for Stress Detection with Deep Learning”. In: *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*. Ed. by Joaquim Filipe, Michal Smialek, Alexander Brodsky, and Slimane Hammoudi. SCITEPRESS, 2021, pp. 256–263. isbn: 978-989-758-509-8. doi: 10.5220/0010474202560263. url: https://doi.org/10.5220/0010474202560263.

[Bai21] Alice Baird, Andreas Triantafyllopoulos, Sandra Zänkert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, Brigitte M. Kudielka, Nicolas Rohleder, Harald Baumeister, and Björn W. Schuller. “An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress”. In: *Frontiers in Computer Science* 3 (Dec. 6, 2021), p. 750284. issn: 2624-9898. doi: 10.3389/fcomp.2021.750284. url: https://www.frontiersin.org/articles/10.3389/fcomp.2021.750284/full (visited on 04/10/2023).

[Bel21] Anas Belouali, Samir Gupta, Vaibhav Sourirajan, Jiawei Yu, Nathaniel Allen, Adil Alaoui, Mary Ann Dutton, and Matthew J Reinhard. “Acoustic and language analysis of speech for suicidal ideation among US veterans”. In: *BioData mining* 14 (2021), pp. 1–17.

[Ben95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[Bes23] Yves Bestgen. "Measuring lexical diversity in texts: The twofold length problem". In: *Language Learning* (2023).

[Bir09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[Bla23] Jost U. Blasberg, Mathilde Gallistl, Magdalena Degering, Felicitas Baierlein, and Veronika Engert. "You look stressed: A pilot study on facial action unit activity in the context of psychosocial stress". In: *Comprehensive Psychoneuroendocrinology* 15 (2023), p. 100187. ISSN: 2666-4976. DOI: https://doi.org/10.1016/j.cpnec.2023.100187. URL: https://www.sciencedirect.com/science/article/pii/S2666497623000218.

[Bre20] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. "pyannote.audio: neural building blocks for speaker diarization". In: *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain, May 2020.

[Bre21] Hervé Bredin and Antoine Laurent. "End-to-end speaker segmentation for overlap-aware resegmentation". In: *Proc. Interspeech 2021*. Brno, Czech Republic, Aug. 2021.

[Buc14] Tony W. Buchanan, Jacqueline S. Laures-Gore, and Melissa C. Duff. "Acute stress reduces speech fluency". In: *Biological Psychology* 97 (Mar. 2014), pp. 60–66. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2014.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0301051114000441 (visited on 11/08/2022).

[Cam04] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe. "Multimodal analysis of expressive gesture in music and dance performances". In: *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers 5*. Springer. 2004, pp. 20–39.

[Chy23] Phie Chyan, Andani Achmad, Ingrid Nurtanio, and Intan Sari Areni. "Multi-Stage Approach for Stress Detection Using Speech Lexical Analysis". In: *2023 IEEE 7th International Conference on Information Technology, Information Systems and Elec-*

trical Engineering (ICITISEE). 2023, pp. 157–162. DOI: 10.1109/ICITISEE58992.2023.10404529.

[Coh97] Sheldon Cohen, Ronald C Kessler, and Lynn Underwood Gordon. *Measuring stress: A guide for health and social scientists*. Oxford University Press, USA, 1997.

[Daw16] Michael E. Dawson, Anne M. Schell, and Diane L. Filion. “The electrodermal system”. In: *Handbook of psychophysiology, 4th ed.* Cambridge handbooks in psychology. New York, NY, US: Cambridge University Press, 2016, pp. 217–243. ISBN: 978-1-107-05852-1 978-1-316-72858-1. URL: https://doi.org/10.1017/9781107415782.010.

[Emp21] EmpkinS. *Empatho-Kinaesthetic Sensor Technology – Sensor Techniques and Data Analysis Methods for Empatho-Kinaesthetic Modeling and Condition monitoring*. https://www.empkins.de/. Accessed: (21.08.2024). 2021.

[Gaa09] Jens Gaab. “PASA–primary appraisal secondary appraisal”. In: *Verhaltenstherapie* 19.2 (2009), pp. 114–115.

[Gia17] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P.G. Simos, K. Marias, and M. Tsiknakis. “Stress and anxiety detection using facial cues from videos”. In: *Biomedical Signal Processing and Control* 31 (2017), pp. 89–101. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2016.06.020. URL: https://www.sciencedirect.com/science/article/pii/S1746809416300805.

[Ham20] Ajna Hamidovic, Kristina Karapetyan, Fadila Serdarevic, So Hee Choi, Tory Eisenlohr-Moul, and Graziano Pinna. “Higher circulating cortisol in the follicular vs. luteal phase of the menstrual cycle: a meta-analysis”. In: *Frontiers in endocrinology* 11 (2020), p. 532846.

[Het09a] S Het, N Rohleder, D Schoofs, C Kirschbaum, and OT19307062 Wolf. “Neuroendocrine and psychometric evaluation of a placebo version of the ‘Trier Social Stress Test’”. In: *Psychoneuroendocrinology* 34.7 (2009), pp. 1075–1086.

[Het09b] 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’”. In: *Psychoneuroendocrinology* 34.7 (2009), pp. 1075–1086. ISSN: 0306-4530. DOI: https://doi.org/10.1016/j.psyneuen.2009.02.008. URL: https://www.sciencedirect.com/science/article/pii/S0306453009000614.

[Hon17] Matthew Honnibal and Ines Montani. “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. To appear. 2017.

[Kir93] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. “The ’Trier Social Stress Test’ – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting”. In: *Neuropsychobiology*. Vol. 28. Issue: 1-2ISSN: 0302282X. 1993, pp. 76–81. ISBN: 978-0-87421-656-1. DOI: 10.1159/000119004.

[Kir94] Clemens Kirschbaum and Dirk H Hellhammer. “Salivary cortisol in psychoneuroendocrine research: recent developments and applications”. In: *Psychoneuroendocrinology* 19.4 (1994), pp. 313–333.

[Las20] Julie Lasselin, Tina Sundelin, PM Wayne, MJ Olsson, S Paues Göranson, John Axelsson, and Mats Lekander. “Biological motion during inflammation in humans”. In: *Brain, behavior, and immunity* 84 (2020), pp. 147–153.

[Lau08] Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissiota, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. “In a nervous voice: Acoustic analysis and perception of anxiety in social phobics’ speech”. In: *Journal of Nonverbal Behavior* 32 (2008), pp. 195–214.

[Laz06] Richard S Lazarus. *Stress and emotion: A new synthesis*. Springer publishing company, 2006.

[Lor18] Steven Loria. “textblob Documentation”. In: *Release 0.15* 2 (2018).

[Mil13] Robert Miller, Franziska Plessow, Clemens Kirschbaum, and Tobias Stalder. “Classification criteria for distinguishing cortisol responders from nonresponders to psychosocial stress: evaluation of salivary cortisol pulse detection in panel designs”. In: *Psychosomatic medicine* 75.9 (2013), pp. 832–840.

[Mis03] René Misslin. “The defense system of fear: behavior and neurocircuitry”. In: *Neurophysiologie Clinique/Clinical Neurophysiology* 33.2 (2003), pp. 55–66.

[OCo21] Daryl B. O’Connor, Julian F. Thayer, and Kavita Vedhara. “Stress and Health: A Review of Psychobiological Processes”. In: *Annual Review of Psychology* 72 (Jan. 4, 2021), pp. 663–688. ISSN: 1545-2085. DOI: 10.1146/annurev-psych-062520-122331.

[Oes23] Marie Oesten, Robert Richer, Luca Abel, Nicolas Rohleder, and Björn Eskofier. "VoStress – Voice-based Detection of Acute Psychosocial Stress". In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* (Pittsburgh). IEEE, Oct. 15–18, 2023. DOI: 10.1109/BHI58575.2023.10313458. URL: https://www.mad.tf.fau.de/files/2023/12/oesten23%7B%5C_%7Dstress%7B%5C_%7Ddetection%7B%5C_%7Dvoice.pdf.

[Oza21] Sachiyo Ozawa. "Emotions induced by recalling memories about interpersonal stress". In: *Frontiers in Psychology* 12 (2021), p. 618676.

[Pan19] Suja Sreeith Panicker and Prakasam Gayathri. "A survey of machine learning techniques in physiology based mental stress detection systems". In: *Biocybernetics and Biomedical Engineering* 39.2 (2019), pp. 444–469.

[Pru03] Jens C Pruessner, Clemens Kirschbaum, Gunther Meinlschmid, and Dirk H Hellhammer. "Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change". In: *Psychoneuroendocrinology* 28.7 (2003), pp. 916–931.

[Qiu12] Jing Qiu and Rolf Helbig. "Body posture as an indicator of workload in mental work". In: *Human factors* 54.4 (2012), pp. 626–635.

[Rad23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR. 2023, pp. 28492–28518.

[Ric21a] Robert Richer, Arne Küderle, Jana Dörr, Nicolas Rohleder, and Bjoern M Eskofier. "Assessing the influence of the inner clock on the cortisol awakening response and pre-awakening movement". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[Ric21b] Robert Richer, Arne Küderle, Martin Ullrich, Nicolas Rohleder, and Bjoern M Eskofier. "BioPsyKit: A Python package for the analysis of biopsychological data". In: *Journal of Open Source Software* 6.66 (2021), p. 3702.

[Ric24] Robert Richer, Veronika Koch, Luca Abel, Felicitas Hauck, Miriam Kurz, Veronika Ringgold, Victoria Müller, Arne Küderle, Lena Schindler-Gmelch, Bjoern M Eskofier, et al. "Machine learning-based detection of acute psychosocial stress from body posture and movements". In: *Scientific Reports* 14.1 (2024), p. 8251.

[Rob18] Alexandra Robinson. “Let’s Talk About Stress: History of Stress Research”. In: *Review of General Psychology* 22 (Feb. 2018). DOI: 10.1037/gpr0000137.

[Roe10] Karin Roelofs, Muriel A Hagenaars, and John Stins. “Facing freeze: social threat induces bodily freeze in humans”. In: *Psychological science* 21.11 (2010), pp. 1575–1581.

[Roh06] Nicolas Rohleder, Jutta M Wolf, Enrique F Maldonado, and Clemens Kirschbaum. “The psychosocial stress-induced increase in salivary alpha-amylase is independent of saliva flow rate”. In: *Psychophysiology* 43.6 (2006), pp. 645–652.

[Roh19] Nicolas Rohleder. “Stress and inflammation–The need to address the gap in the transition between acute and chronic stress effects”. In: *Psychoneuroendocrinology* 105 (2019), pp. 164–171.

[Sal08] Mohd Razali Salleh. “Life event, stress and illness”. In: *The Malaysian journal of medical sciences: MJMS* 15.4 (2008), p. 9.

[Sha13] Bahar Shahidi, Ashley Haight, and Katrina Maluf. “Differential effects of mental concentration and acute psychosocial stress on cervical muscle activity and posture”. In: *Journal of electromyography and kinesiology* 23.5 (2013), pp. 1082–1089.

[She21] Lucas Shen. *Measuring Political Media Slant Using Text Data*. 2021. URL: https://www.lucasshen.com/research/media.pdf.

[She22] Lucas Shen. *LexicalRichness: A small module to compute textual lexical richness*. 2022. DOI: 10.5281/zenodo.6607007. URL: https://github.com/LSYS/lexicalrichness.

[Tor13] Joan Torruella and Ramón Capsada. “Lexical statistics and tipological structures: a measure of lexical richness”. In: *Procedia-Social and Behavioral Sciences* 95 (2013), pp. 447–454.

[Val18] Raphael Vallat. “Pingouin: statistics in Python.” In: *J. Open Source Softw.* 3.31 (2018), p. 1026.

[Wie13] Uta S Wiemers, Daniela Schoofs, and Oliver T Wolf. “A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women”. In: *Stress* 16.2 (2013), pp. 254–260.

[Zän20] Sandra Zänkert, Brigitte M Kudielka, and Stefan Wüst. “Effect of sugar administration on cortisol responses to acute psychosocial stress”. In: *Psychoneuroendocrinology* 115 (2020), p. 104607.

[Zha20] Huijun Zhang, Ling Feng, Ningyun Li, Zhanyu Jin, and Lei Cao. “Video-Based Stress Detection through Deep Learning”. In: *Sensors* 20 (Sept. 2020), p. 5552. DOI: 10.3390/s20195552.

Appendix C

Acronyms

HPA hypothalamic-pituitary-adrenal

TSST Trier Social Stress Test

f-TSST friendly Trier Social Stress Test

p-TSST placebo Trier Social Stress Test

(f-)TSST (friendly) Trier Social Stress Test

ML machine learning

MaD Lab Machine Learning and Data Analytics Lab

SD standard deviation

BMI Body Mass Index

CNN convolutional neural networks

PASA primary appraisal secondary appraisal

IMU inertial measurement unit

SNS sympathetic nervous system

α-amylase alpha amylase

ttr type-token-ratio

mattr Moving average type-token-ratio

mtld Measure of textual lexical diversity

HD-D Hypergeometric Distribution Diversity

nltk Natural Language Toolkit

ANOVA analysis of variance

kNN k-nearest neighbors

CV cross validation

SVM support vector machines

AdaBoost Adaptive Boosting

RFE recursive feature elimination