INVESTIGATING THE EFFECTS OF PHYSIOLOGY-DRIVEN VIBRO-TACTILE BIOFEEDBACK FOR MITIGATING STATE ANXIETY DURING PUBLIC SPEAKING

A Thesis

by

JASON DAVID RAETHER

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Theodora Chaspari
Co-Chair of Committee,	Annmarie MacNamara
Committee Member,	Ricardo Gutierrez-Osuna
Head of Department,	Scott Schaefer

May 2022

Major Subject: Computer Science

ABSTRACT

For some, public speaking can cause heightened moments of stress while giving a speech or presentation. These moments are quantifiable through one's physiology and vocal characteristics, measurable through sensor-enabled smart technology. Through these measurements, we can assess the current state of the individual to determine opportune moments to deliver interventions that alleviate symptoms of stressful moments.

Recent work in wrist-worn vibrotactile biofeedback suggests that it is a promising intervention towards reducing state-based anxiety for public speaking. However, since the vibrotactile stimulus is delivered constantly, adaptation could risk diminishing relieving effects. Therefore, we administer vibrotactile biofeedback as a just-in-time adaptive intervention during in-the-moment heightened levels of stress. We evaluate two types of vibrotactile feedback delivery mechanisms in a between-subjects design – one that delivers stimulus randomly and one that delivers stimulus during moments of heightened physiological reactivity, as determined by changes in electrodermal activity. The results from these interventions indicate that vibrotactile biofeedback administered during high physiological arousal appears to improve stress-related measures early on, but these effects diminish over time. However, we also observe no significant differences in self-reported state anxiety scores between experiment groups.

In the latter half of this thesis, we will explore methods for personalizing machine learning models that detect the onset of heightened moments of stress in real-time. Results indicate that baseline-norming, fine-tuning on participant-specific data, and providing individual-specific trait information are all helpful techniques for improving stress detection performance.

DEDICATION

To my mother Mary Lou, my father Helmut, my sister Kirsten, and my dog Taz, without whom I would have never been able to accomplish this.

ACKNOWLEDGMENTS

First off, a huge thank you goes out to Dr. Chaspari for helping me complete this thesis every step of the way and helping to guide me throughout my academic career at Texas A&M. Dr. Gutierrez-Osuna and Dr. MacNamara were also extremely helpful in providing their insight and further guidance for this thesis. I am forever grateful for my committee and I'm very lucky to have chosen a group of talented, passionate individuals with extensive history in related fields.

I would also like to thank Ehsan for his extensive help and guidance in the data analysis for physiology and speech, as well as his support throughout the process, and Sakib for his help in setting up the experiment room and for generously allowing me to use his equipment for my experiment.

I would also like to thank my girlfriend, Lexi, my parents, Helmut and Mary Lou, and my sister, Kirsten, for their constant love and support while going through graduate school. I wouldn't have been able to complete this without them.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Theodora Chaspari (Chair) and Dr. Gutierrez-Osuna (Member) of the Computer Science and Engineering department and Dr. Annmarie MacNamara of the Department of Psychology and Brain Sciences (Co-Chair).

The user study for this thesis was conducted with the help of Dr. Theodora Chaspari, Md Nazmus Sakib, and Ehsanul Haque Nirjhar. Md Nazmus Sakib assisted with the setup of the virtual reality system and Ehsanul Haque Nirjhar and Dr. Theodora Chaspari helped in providing software tools and guidance for the data analysis and real-time signal processing.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

This work has been supported by the Engineering information Foundation (EiF) through grant number (18.02). The author gratefully acknowledges the support from EiF. Any opinions, findings, conclusions, and recommendations expressed in this thesis are those of the author and do not necessarily represent those of the EiF.

NOMENCLATURE

PSA	Public Speaking Anxiety
EDA	Electrodermal Activity
GSR	Galvonic Skin Response
HRV	Heart Rate Variability
HR	Heart Rate
BVP	Blood Volume Pulse
SCR	Skin Conductance Response
SCL	Skin Conductance Level
F0	Fundamental Frequency
mHealth	Mobile Health
EMA	Ecological Momentary Assessment
EMI	Ecological Momentary Intervention
JITAI	Just-In-Time Adaptive Intervention
IBI	Interbeat Interval
RMSSD	Root-mean-square of successive differences between heartbeats
SDNN	Standard deviation between NN intervals
HF	Absolute power of high frequency band in HRV
LF	Absolute power of low frequency band in HRV

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Public speaking is an important skill for both professionals and academics to communicate their ideas effectively to their audience [3, 4]. However, there are many people that face public speaking with apprehension and anxiety. Public speaking anxiety (PSA) is defined by the American Psychological Association as the “fear of giving a speech or presentation in public because of the expectation of being negatively evaluated or humiliated by others" [5]. It is categorized as one of the sub-phobias that affect individuals with social anxiety disorder (SAD) [6], but it also affects the general population. This has been approximated as being about 30% of the general population [7, 8], but some estimates are as high as 75% [9]. Individuals suffering from both SAD and PSA are shown to have more negative cognition with regards to their speaking skills, resulting in poor public speaking performance [10]. In order to help these individuals cope with their public speaking anxiety symptoms while giving or practicing giving a speech, we can design interventions that can help them, either through cognitive restructuring [11, 12] or subconscious relief [13, 14].

With how ubiquitous smart devices and smart technology are today due to the Internet of Things (IoT), the approach of using mobile technology to augment and assist users efforts in managing their health and wellness has rapidly expanded into the field named mHealth [15, 16]. mHealth has been applied to a variety of domains, including behavior monitoring, health diagnostics, vitals tracking, and other targeted applications like dieting, binge-eating, alcohol abuse, and sedentary activity prevention [17, 18, 19, 20]. mHealth has the advantage over traditional methods due to increased accessibility and scalability [21]. The technology is accelerating rapidly, but the growth is outpacing the research, with a call for further studies evaluating the feasibility and effectiveness of mHealth solutions [15, 16, 22, 23].

One intervention that has been investigated is vibrotactile biofeedback. The vibrotactile sensation is usually delivered through a wrist-worn device, constantly throughout the stress-inducing task [14, 13]. Since the delivery is constant, the next logical step is to evaluate its effectiveness delivered in shorter durations. The reasons for doing this are two-fold: one is to maintain calm-

ing effects should the user habituate to the vibrotactile stimulus, and the other is its relevance to interventions for in-the-moment interventions. Due to the tactile medium in which vibrotactile biofeedback is administered, it could be a good candidate for a real-time accessible intervention that operates on a 'subconscious' level, requiring no interaction from the user.

This study is grounded in mHealth design principles and draws from existing research on intervention frameworks and public speaking interventions, as well as suggestions and challenges for the future of mHealth. We seek to investigate the efficacy of vibrotactile biofeedback as a real-time in-the-moment intervention during a public speaking task. The public speaking task will be performed within a virtual reality (VR) environment of a virtual audience. Furthermore, we explore methods for personalizing statistical models that detect high stress moments within a speech. This contrasts with existing work which typically evaluates over a very large timescale. Related work which attempts to classify individuals with high anxiety only does so at the level of the overall session, not within the session itself. The latter strategy presents significantly more challenges due to limited analysis windows and real-time, fault-tolerant system requirements.

1.1 Related Work

1.1.1 Public Speaking Anxiety

Prior research has introduced two main theoretical models to describe PSA. The first is the Trait-State model, which separates PSA into *trait* and *state* distinctions. The trait aspect of PSA represents a general susceptibility of the individual to experience anxiety-related symptoms while giving a speech. The state aspect of PSA defines the act of public speaking itself as a non-deterministic task, and thus having stochastic or random components to it that could invoke anxiety depending on the environment [24].

The three-systems model is the second theoretical model used to explain PSA. According to [24], the three-systems model defines three components of PSA – these are physiology, cognition, and behavior. The *physiology* component represents the changes in physiological indicators with increased levels of PSA. These changes are mainly attributed to the sympathetic branch of the

autonomic nervous system (ANS), which controls the 'fight-or-flight' response in humans [25, 24]. Some examples of the physiological indicators that change as a result of increased levels of PSA are electrodermal activity (EDA), as well as blood volume pulse (BVP), of which we can derive heart rate (HR) and heart rate variability (HRV) from [26, 27, 28, 29]. Both of these can be measured directly and unobtrusively using sensors incorporated into wearable devices.

The individual's own internal cognition of themselves and their speaking performance covers the *cognitive* component of the three-systems model. These can be quantified by having the individual complete self-reports after the public speaking task [24]. Furthermore, we can estimate a more granular representation of their cognitive state by having the individual annotate the playback of their speech with a numerical representation of their stress levels at that point in time [30].

Finally, the behavioral component of the three-systems model describes the observable behaviors the individual expresses during the speech. Some of these behaviors include fidgeting, stuttering, voice trembling, or frequent use of stop words. These can also be measured by self-reports from the individual or from 3rd party observations, but can also be quantified through speech processing and analysis [24].

Virtual reality environments are a promising medium that can be used to elicit PSA from an individual. When giving a speech to a VR audience, humans still experience the same reactions and emotions associated with a real audience, but to a slightly lesser degree [31]. Regardless, using a VR environment is more feasible in practice, and we can emulate different conditions in the audience, like negative facial expression or background noise. It also reduces attrition rates in carrying out the actual experiment [32]. Related work has investigated using VR coaches to augment cognitive restructuring during or after the speech [33, 34]. Other studies have measured the improvement in real-life public speaking performance and reductions in PSA after subsequent exposure to VR public speaking experiences [35, 36, 30, 37].

1.1.2 mHealth and Intervention Frameworks

As stated in section 1.1.1, wearable devices can be used to measure a user's physiology, they can also be used to provide interventions in mHealth applications. There are many important fac-

tors to consider when designing an mHealth system that interfaces with the human individual. If there are devices the user must where, some of these include locality of the devices on the user, physical appearance of any devices, social acceptability, comfort, maintenance, physiological accuracy/relevancy, and required interactions from the user [38, 39, 40, 41, 42]. Wrist-worn devices mimicking the form factor of a watch are excellent candidates that fit this criteria and rank the highest on social acceptability and comfort on behalf of the wearer [43, 44].

While the location of any measuring devices is important, designing the intervention and determining when and how to deliver it carries more complexity. Previous work on interventions define Ecological Momentary Interventions (EMI) as a framework for mobile technology to deliver health and wellness interventions for real-world applications [45]. EMI are informed by Ecological Momentary Assessment (EMA), which defines how and how often to sample from the current state of the user [46]. This sampling and overall assessment after sampling is then used to guide the intervention timing and magnitude of delivery (whether that be strength, frequency, or duration depends on the intervention).

Building on these frameworks, Just-In-Time Adaptive Interventions (JITAI) were introduced to provide mHealth designers a computational framework for EMI and how to assess their effectiveness [47]. The JITAI framework covers the requirements for designing an effective intervention. These requirements include distal (long-term) and proximal (short-term) outcomes, tailoring variables (how the treatment is affected depending on variables), decision points and decision rules (what do we use as the criteria and when do we decide to intervene?), and intervention options. A well-designed JITAI should adapt to the changing requirements of the user, leading to the importance of personalization for the JITAI decisions and tailoring variables. Many studies have used this design framework to great success. Goldstein et al., 2017 created a JITAI app that helped prevent users from overeating, using machine learning models to assess the individual at any moment [48]. Choi et al., 2019 used JITAI to prevent alcohol binging episodes [20]. There has also been work done in designing JITAI models which prevent sedentary behavior [18].

Regarding interventions, one topic not covered in the JITAI framework is the framing of the

intervention back to the user. The framing is important since the outcome of the intervention can change based on what the user's interpretation of the intervention's purpose. Costa et al., 2016 delivered vibrotactile biofeedback to individuals at 60 bpm. However, for one group, the user's were deceived into believing the vibration rate was actually their heart beat, leading them to believe their heart rate was lower than it actually was [49]. This 'false-rate feedback' had a substantial impact on anxiety-related measures for the experiment when compared to the group who received the vibrations at a rate that was their true heart rate. Similarly, Hollis et al., 2018 ran a study that would notify the user when increased levels of EDA were recognized [50]. One group was informed the notifications were signaling that they were "alert" and "engaged" (positive-framing), whereas another group was informed the notifications detected they were "stressed" (negative-framing). The negative-framing group had higher levels of anxiety-related measures compared to the positive-framing group, further supporting the perception of the intervention on behalf of the user is an important design consideration.

1.1.3 Quantifying Anxiety

There are several modalities of which we can collect information about the user that can be used to estimate the user's affective state. Non-invasive methods include measuring electrodermal activity (sometimes known as galvanic skin response, or GSR), blood volume pulse, and speech. Features extracted from EDA are known to correlate with high state anxiety in some individuals [29]. EDA is frequently used in other studies as an indicator of increased stress or measuring when a stimulus has been introduced to the individual [51, 52, 19, 53, 50].

BVP is derived from the waveform generated by the photoplethysmography (PPG). Again, heart rate and heart rate variability are extracted from BVP and also frequently used in stress-detection studies. Increased heart rate is typically seen as an indicator of increased levels of anxiety [27], while HRV has an inverse relationship with anxiety [54, 28]. One word of caution regarding using HRV for short time analysis (named ultra-short term HRV, defined as less than 5 minutes) has been found to be less reliable for frequency-based characteristics (and general characteristics) compared to longer duration (>5 minutes) analysis windows [55].

Speech is widely used in emotion recognition, making it a relevant medium to transfer to stress detection [56, 57]. There are features of speech that can be used to describe the characteristics of the individual's speech and speaking patterns which we can use to help quantify the behavioral aspect of PSA.

1.1.4 Vibrotactile Biofeedback

Recent work has delivered vibrotactile biofeedback in the form of continuous pulses (similar to a heartbeat pattern) applied to the wrist area using wrist-worn devices (similar to a watch) [14, 13, 58, 49, 59]. This rate varies for some studies, but the agreed value appears to have landed on 60 beats per minute. Indeed, a calming effect has been observed at that rate by multiple studies [14, 13, 58]. The vibrations are applied as a constant stimulus during stress-inducing tasks or after exercise, and has been shown to reduce some physiological indicators of stress like EDA peaks and heart rate. Choi & Ishii, 2020 observed that individuals receiving vibrotactile feedback at a rate of 60 bpm had faster heart rate decreases over time compared to control [14]. In the case of [13], individuals that received the stimulus in this way also reported lower levels of state anxiety differences between before and after anticipating giving a public speech compared to a control group.

For the question of why vibrotactile biofeedback has a calming effect, the answer to that remains unclear. One theory is that it could be related to 'entrainment', or the theory that the human body tends to subconsciously synchronize to external rhythms [60]. With audio and visual stimulus, this is an effect that has been observed in [61], but results are mixed [62]. To our knowledge, this has not yet been investigated for vibrotactile stimulus.

Regardless of a possible entrainment effect, if we are to assume the calming effect is due to the perceived magnitude of the stimulus (or at least in some part due to the sensation of the stimulus), then vibrotactile adaptation needs to be considered in developing vibrotactile biofeedback further as a feasible intervention. If the user habituates to the vibrotactile sensation, then the calming effects may diminish over time. Vibrotactile adaptation does occur rather quickly, and the perceived magnitude is maximal after 5 seconds and then declines over a period of a few minutes [63].

Vibrotactile biofeedback has an edge over visual or auditory biofeedback since the medium of its delivery doesn't require user interaction as demonstrated by previous studies. The nature of the stimulus is tactile – therefore, the user just has to passively *feel* it, which they can do in parallel with whatever task they have at hand. None of the studies demonstrating the vibrotactile biofeedback's effect required the user to focus on the feeling of the stimulus. Visual stimuli may visually impair the user, while auditory stimuli may draw too many attentional resources from the user. In fact, tactile stimuli were rated as having the lowest disturbance levels [14]. Audio and visual stimulus have the drawback of being perceived by external users, which may draw too much attention to the targeted user. This is especially prevalent in a task with high cognitive load like public speaking.

On calling the delivery of vibrotactile stimulus 'biofeedback', this is technically incorrect for some studies, strictly speaking [64]. Biofeedback typically incorporates the perception of the target signal into the feedback loop. A change in the signal affects the user, which in turn affects the change in the signal, and this effect continues in a loop. Many of these studies don't inform the user about the vibrotactile stimulus or what it represents, except for [49] who told participants the stimulus was their actual heart rate. A more accurate term may be 'subconscious' biofeedback, where the stimulus does affect some physiological measure of the individual, but what that variable is is unknown. However, the term vibrotactile biofeedback has already been used in previous work, so to remain consistent we use the same.

It is important to keep real-life feasibility in mind when designing ecological intervention systems. Interventions that may perform well in a controlled experiment have no outlook on transferring to a longitudinal study as the user goes about their life. For instance, visual-based interventions would require an electronic display for the user to look at which may distract them from their current task. Audio interventions may be audible for other people to hear, unless the user is constantly wearing headphones. The user's perception of social acceptability of the intervention is another key factor in intervention success. Devices worn on the wrist are ranked as the most socially acceptable compared to other locations on the body [43]. An intervention that draws attention to

the user from outside observers may actually inhibit its effectiveness and elevate feelings of social anxiety [44]. Wrist-worn devices can minimize this risk, since they can continuously monitor one's physiological signals and at the same time provide vibro-tactile feedback in an unobtrusive manner.

1.1.5 Estimation of Moments of Stress

While there has been a considerable amount of work done in predicting stress, very few studies perform estimation of stress within the stressful event itself. In fact, there is little differences between these studies with predicting stress aside from their use of physiology and speech as modalities for inputs of prediction models. For example, Healey and Picard (2005) performed real-time stress detection using data collected from people driving [65]. The stress labels were decided based on the area the participant was driving (e.g. highway, city, countryside) and weren't decided based on self-reported stress from the individual. Further studies that predict stress have the same problems [66] where stress is decided based on whether or not the participant is performing the *task*, rather than querying the participant for their own stress levels. Ciabattoni et al. (2017) does the same thing with participants taking part in a logic task. Other work also involves *assuming* the introduction of a stimulus increases stress, which again lacks the methodology of querying the user for their cognition of stress [67]. Martinez et al. (2017) labeled levels of stress using rule-based decisions on physiological data and their own analysis [68]. Kyriakou et al. (2019) exposed participants to an airhorn sound, assuming those moments were stress-inducing [69].

Within the domain of stress prediction, there have been studies that apply this stress to public speaking, but they again suffer the same methodology issues as Healey and Picard. Lu et al. (2012) performed stress detection within a public speaking domain, but the affect labels were done based on the speaking task (e.g. job interview and marketing presentation), but no labeling was done within those tasks [70].

Fortunately, a small number of studies have been done which estimate moments of stress *within* the stressful event or task itself using affect labeling [71]. Soury and Devillers (2013) used expert annotators for continuous annotations of stress levels of speakers in a simulated job interview [72].

Kimani and Bickmore (2019) proposed a sensor-based framework for anxiety detection based on affect labeling where the participant watches a video of their public speaking presentation [34]. Wen et al. (2020) had annotators label moments of externally-perceived stress when watching the participant perform their thesis defense [73]. Gjoreski et al. (2016) also used affect labeling, but had longer durations of analysis windows to take advantage of [74]. To our knowledge, no studies have attempted to perform estimation of state anxiety during public speaking using affect labeling.

1.2 Research Contributions and Proposed Approach

1.2.1 Proposed Approach

The first part of this thesis will expand on related work investigating the efficacy of vibrotactile biofeedback. In contrast to other studies delivering continuous stimulus throughout the stress-inducing task, we will experiment with delivering vibrotactile biofeedback in shorter segments to avoid vibrotactile adaptation in the participant. Two algorithms for delivering the intervention will be evaluated and compared to a control group who receives no stimulus. One of these algorithms will randomly deliver the intervention for a short segment (Random), and the other will be a heuristic-based algorithm using physiological indicators from the participant (Heuristic), captured within a small analysis window relative to the overall duration of the public speaking session. The participant's physiological state will be captured with a wrist-worn device on the non-dominant handed wrist of the participant to prevent motion artifacting in the captured signals. The heuristic-based algorithm will therefore use EMA to assess the state of the user. The algorithms are described in detail in section 3.3. The vibrotactile biofeedback device will be applied when each algorithm decides to enable it. The device will then vibrate at 60 bpm in a 'rhythmic' pattern consistent with [14, 49] and be applied on the dominant handed wrist of the participant.

The effectiveness of the proposed vibrotactile feedback will be compared at the micro- and macro-level. At the micro-level, we will compare the moments before and after the vibrotactile stimulus is introduced amongst the groups. At the macro-level, we will examine the overall effect of each experimental group to observe the calming effect of vibrotactile biofeedback seen in pre-

vious studies. We will also examine these sessions longitudinally to catch any habituating effects of the stimulus groups.

The second part of this thesis will explore methods for personalizing adaptive machine learning models for detecting heightened stressful moments for real-time applications. 'Real-time' is defined in this sense as within a few seconds of time granularity using very short duration analysis windows (30 seconds). This contrasts with other claims of 'real-time' which actually perform analysis over the entire session without regarding the necessity of true proximal state captured of the individual.

We will investigate offsetting models with the trait anxiety scores of each participant, as well as using baseline physiological data as context to the models. We will also utilize the 'v1' VerBIO dataset [75] to explore the effectiveness of training the model to a general population versus and individual-specific dataset, as well as fine-tuning with both.

1.2.2 Research Contributions

The contributions of this thesis are evaluating the effectiveness of vibrotactile biofeedback delivered as an in-the-moment intervention for alleviating anxiety. It also evaluates the differences in delivering a physiology-informed intervention compared with a randomly-delivered intervention. These contributions differ from previous work in that the vibrotactile biofeedback isn't delivered continuously, like in [14, 49, 13], and instead delivered in short, targeted moments.

The latter part of this thesis advances efforts in the task of detecting onsets of self-reported stressful moments from public speaking using short-duration windows of user state information. The moments will be labeled using affect labeling and will be done by the user themselves to obtain a granular ground-truth label for their moment-to-moment cognition. This differs from previous work in stress detection which uses global annotations of stress levels or relatively large analysis windows for prediction of stress, and doesn't operate within the domain of public speaking. The stressful moments *within* the stressful event (public speaking) will be estimated based on proximal context of the individual

1.2.3 Research Questions

The research questions this thesis will answer are as follows:

1. Does physiology-informed delivery of vibrotactile biofeedback better alleviate proximal measures of state anxiety compared to random delivery or no delivery?

2. What impact does vibrotactile biofeedback have on self-reported anxiety, physiology, and acoustic measures of speech across public speaking sessions and overall at the end of all sessions?

3. To what extent are personalized machine learning models effective for predicting moment-to-moment self-reported stress in real-time?

2. USER STUDY

2.1 Study Population

We recruited 30 participants ($N = 30$) to take part in this study using a campus-wide e-mail. All participants were students at Texas A&M University and between the ages of 18 and 30. Each participant took no longer than 3 hours for the experiment, and upon completion of the entire procedure were compensated with a $25 Amazon gift card.

Characteristic	Population Details
Number of participants	30
Age	18 – 27
Sex	13 female, 17 male
Education	20 undergraduate, 10 graduate/post-graduate
Ethnicity	14 Asian, 4 Hispanic/Latino, 7 White/Caucasian, 2 Black/African American, 3 Other/Prefer Not To Say

2.2 Devices

2.2.1 Wrist-Worn Devices

To measure BVP and EDA, we use the Empatica E4 [76], which is a research-grade device used to recording physiological signals. It can capture EDA, BVP, skin temperature, and 3-axis accelerometer data. EDA is sampled at a rate of 4 Hz while BVP is sampled at 64 Hz. The E4 also has the capability to send interbeat interval (IBI) derived from the BVP signal, but this was intentionally ignored by our recording software since the computation of IBI blocks the TCP stream. As stated before, the E4 is worn on the participant's non-dominant hand.

To deliver the vibrotactile biofeedback at a rate of 60 bpm, we use the Soundbrenner Pulse [1]. The Pulse is a metronome-like device used by musicians to keep time, and delivers each 'click' of the metronome as a vibration instead. It is capable of rates between 20 and 400 bpm,

(a) (b)

Figure 2.1: Soundbrenner Pulse (Left) and Empatica E4 (Right)

but was locked at 60 bpm for this study. The app that controls the Pulse also offers control over the strength and duration of each beat. For future reference, we chose the medium strength and medium duration settings. The Pulse is connected via Bluetooth to the experimenter's cell phone, where the experimenter can control the delivery of the intervention from the outside of the room. Images of both devices are presented in Figure 2.1.

2.2.2 Other Devices

To display the virtual environment during the speech, we use the Oculus Rift VR headset [77]. The participant wears this over their head and the environment is projected through the optical lenses in the headset. During the speech, we also record the participant's speech in real time using a lapel microphone from FIFINE [78]. The microphone is worn on the collar of the participant's shirt.

2.3 Software

To generate the virtual audience in the VR environment, we use the Virtual Orator software [79]. Virtual Orator offers options to change the environment being display (e.g. meeting room, classroom, large hotel), the number of people in the audience, and the 'difficulty' of the audience

Figure 2.2: App interface for the Pulse [1]

(loud background noises, disinterest, facial expressions, etc.).

We use the Soundbrenner app to control the Pulse (see Figure 2.2). The interface allows us to control the rate at which the Pulse vibrates, and to start or stop the vibrations.

The E4 is connected to the E4 streaming server [80], which serves a TCP connection with which we connect to with the *e4stream* Python package [81]. To record the speech data, we use Audacity [82] mono-sampled at a rate of 16 KHz in 16-bit PCM format.

After every speech, each participant listens to a playback of their speech recording and annotates moments in which they felt particularly stressed. We use the CARMA software [2] to do this on a 5-point scale, with '1' being no stress and '5' being stress, similar to prior work [74, 34, 72]. The participant uses a slider to label these on a continuous scale while the speech is playing. See Figure 2.3 for an example of the interface.

While we use many Python packages, the majority packages for feature extraction on EDA and

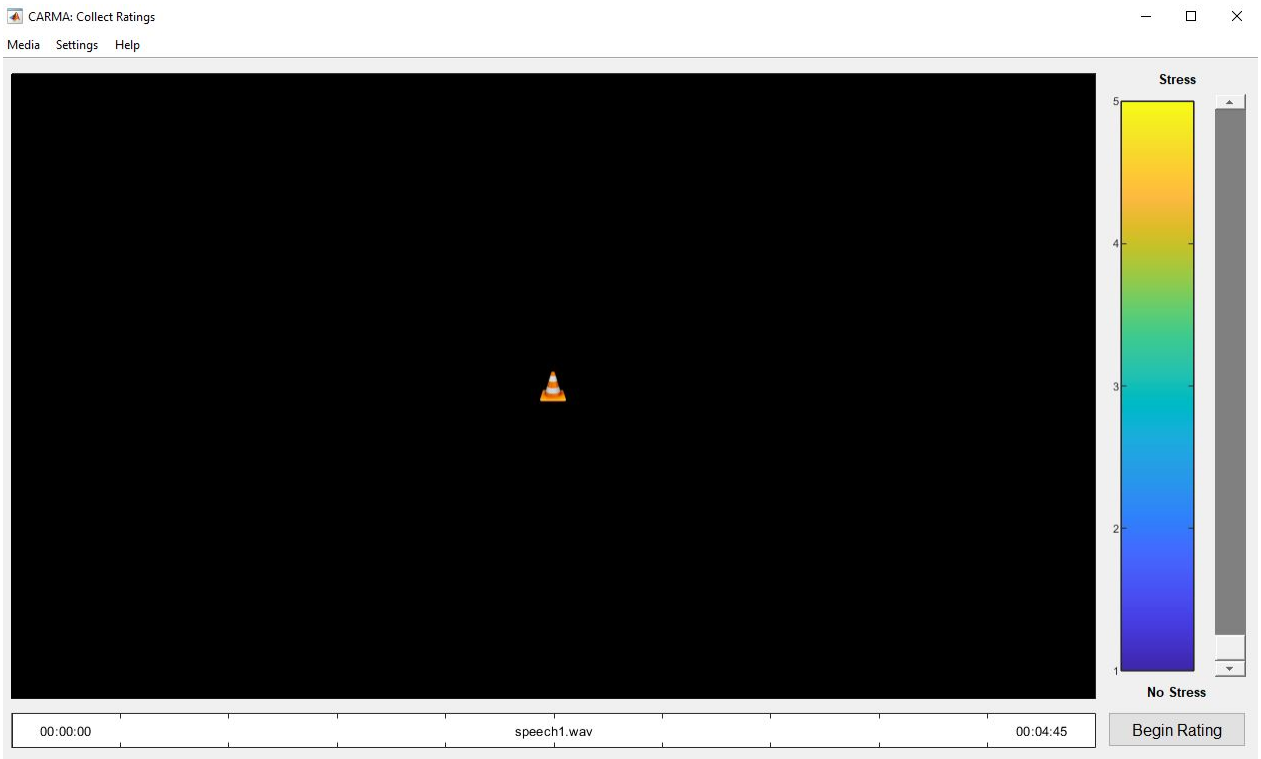


Figure 2.3: CARMA configuration and interface [2]

BVP is a modified version of NeuroKit2 [83], which we adapted to better handle real-time systems. For speech, we use the OpenSMILE package from [84] and the ComParE2016 feature set [85].

2.4 Self-Reports

We use the following self-reports to capture various trait and state information about the user as well as other psychological characteristics. The full procedure description is in section 2.6.

2.4.1 PRE Surveys

The following surveys are taken at the very beginning of the experiment:

- *Big Five Inventory* (BFI): A measure of the big five personality traits, which include extraversion, agreeableness, openness, conscientiousness, and neuroticism [86].
- *Brief Fear of Negative Evaluation* (BFNE): Measures the degree to which the participant fears that their performance will be evaluated negatively [87]. A higher score indicates

higher levels of apprehension of negative evaluation.

- *Trait-Scale of Communication Anxiety Inventory* (CAI-Trait): Measures the trait (generalizable behavior) level of the participant's fear in a particular communication setting [88].
- *Daily Experiences Questionnaire*: A custom questionnaire asking questions about the participant's current day conditions like if they've consumed any caffeine or alcohol.
- *Demographics*: A custom questionnaire asking demographical questions like education, age, gender, etc.
- *Personal Report of Public Speaking Apprehension* (PRPSA) measures the participant's fear of public speaking. Also referred to as the Personal Report of Public Speaking Anxiety [89]. Higher scores indicates higher general apprehension for public speaking.
- *Reticence Willingness to Communicate* (RWTC): Another indicator of the participant's trait communication anxiety [90]. Higher scores indicate higher levels of reticence and low willingness to communicate (essentially how reserved is the individual in speaking).
- *Trait-Scale of State-Trait Anxiety Inventory* (STAI): Measures trait level of anxiety in the participant [91]. A high score indicates higher levels of trait-anxiety.

We use the BFNE, PRPSA, and RWTC to capture the participant's feelings about public speaking, as well as trait scales of the CAI and STAI to assess that particular participant's disposition towards anxiety and public speaking. The BFI may also show some personality traits lead individual's to being more reactive towards public speaking. We also collect population-specific information through the demographics survey to observe any differences between those groups and assess the background of each participant. The daily experiences questionnaire is mainly to validate any factors in the participant's day that may confound results (e.g. like too much caffeine before the experiment).

2.4.2 TEST Surveys

The following surveys are taken at the end of each speech given in the VR environment (every TEST session):

- *Presentation Preparation Performance* (PPP): Measures the participant's level of knowledge and preparedness of the presentation that was just given [30].
- *State Anxiety-Enthusiasm Behavior Scale* (AEBS): Measures anxious behaviors during speech [90]. A higher score indicates higher anxiety and lower levels of enthusiasm regarding the task just performed.
- *Vibrotactile Questionnaire*: A custom questionnaire asking the participant's to give their best guess at the number of times they noticed the vibrotactile feedback begin to vibrate.

For each TEST phase, the state scale of the AEBS is used to measure the state of anxiety in the participant just after the speech. The PPP will help reveal whether any effects were related to the preparedness of the speech and the assigned article. The vibrotactile questionnaire is used to asses how conscious the participant was about the vibrations and how noticeable the vibrotactile sensations were. We can compare this estimation to the true number of intervention toggles in the speech.

2.4.3 POST Surveys

The following surveys are taken after the experiment has concluded:

- *State-Scale of Communication Anxiety Inventory* (CAI-State): Measures the state (current behavior) level of the participant's fear in a particular communication setting [88].
- *State-Scale of State-Trait Anxiety Inventory* (STAI-State): Measures the state level of the participant's anxiety [91].
- *VR Presence* (VRP): Measures level of immersiveness of the VR environment and experience [92].

- *Body Sensations Questionnaire* (BSQ): Measures the level of fear in agoraphobia of the participant [93].
- *Post Experiment Feedback & Post Experiment Vibrotactile Feedback*: A custom questionnaire asking participant questions about the experiment, how they felt during it, and how they felt about the vibrotactile feedback.

At the end of the experiment in the POST phase, we measure the state levels of CAI and STAI for all participants to compare their anxiety levels after the study has concluded. BSQ will also inform us about their levels of state anxiety as well as their internal cognition about their physical symptoms. VRP will measure the overall immersiveness and presence of the VR environment for all speech sessions. Finally, the post experiment feedback and post experiment vibrotactile feedback questionnaire is for gathering subjective thoughts about the experiment and any comments or insights the participant might have.

2.5 Groups

We defined three experimental groups for this study – *Control*, *Random*, and *Heuristic*. The Control group wears the Pulse but the Pulse stays deactivated throughout the entire speech. The Random group wears the Pulse and receives the stimulus according to the Random algorithm (Algorithm 2). The Heuristic group wears the Pulse and receives the stimulus according to the Heuristic algorithm (Algorithm 1). See Table 2.1 for a breakdown of the number of participants per group and sessions per group. Every participant has 4 TEST sessions. Note that for the Control group and Heuristic group, 1 participant from each did not consent to us saving their speech recording, so the audio files are missing for those participants.

2.6 Experiment Procedure

The experiment has three phases, which are the *PRE* phase, the *TEST* phase(s), and the *POST* phase. We cover the details for each phase below.

Intervention Group	# Participants	Total # TEST Sessions
Control	10	40
Heuristic	10	40
Random	10	40
Total	30	120

Table 2.1: Intervention group breakdown

2.6.1 PRE Phase

The PRE phase begins with bringing the participant into the presentation room and briefing them about what to expect for the experiment overall. Nothing is revealed about the purpose of the study or hypotheses, only that the participant will be giving four public speaking sessions in the VR environment, filling out surveys, and reflecting on their speech. After the participant consents to the experiment, we attach the E4 device to the participant's wrist on the non-dominant hand and the Pulse on the other wrist. We briefly explain what the E4 is measuring and provide a brief demonstration of what the Pulse's vibrotactile stimulus will feel like. This demonstration is given to all participants regardless of intervention group. We also inform all participants that they may feel that same sensation during their speeches today.

Next, the participant fills out the demographics and daily experiences questionnaire at a computer, as well as the BFI, BFNE, CAI-Trait, STAI-Trait, PRPSA, and RWTC. Afterwards the participant performs a brief memory test on the MemTrax website [94]. When the memory test is finished, the participant watches a 5 minute relaxation video [95]. During this time we collect physiology measurements from the E4 that will be used as baseline measurements. This data is later referred to as the *RELAX* data. After the relaxation video is finished, this concludes the PRE phase of the experiment.

2.6.2 TEST Phase(s)

Each participant completes four *TEST* phases over the entire experiment. Each TEST phase consists of a preparation (*PREP*) component and a presentation (*PPT*) component.

2.6.2.1 PREP Component

The participant is provided a randomly assigned paper article from a pool of 30 articles. For each participant, each of the articles is different for the four TEST phases. We hand the article to the participant and inform them that they have 10 minutes to read the article and prepare a speech in their head. They aren't allowed to take notes and are also informed that they won't be able to have the article with them during the presentation. While they read and prepare their speech, we collect further data from the E4 to use as a second baseline. This preparation data is referred to as the *PREP* data.

2.6.2.2 PPT Component

After 10 minutes has passed, the presentation component begins. The participant is instructed to attach the lapel microphone to their shirt collar. Then, we place the VR headset on the participant with the built-in headphones over their ears and initialize the VR environment. The audience and environment type are also randomized from a pool of 12 configurations and different for each session for a particular participant. Some environments that include small conference rooms allow the participant to be seated instead of standing.

After confirming the participant can see the environment, we inform the participant that after the recording has started, there will be a 30 second delay from when they can begin their speech. This gives time for the experimenter to leave the room and also collect enough data for the initial 30 second analysis window for the heuristic algorithm. This 30 second period happens to all participants regardless of their assigned group. After 30 seconds, the participant is provided with a single buzz from the Pulse letting them know that they can begin. From this point, interventions are enabled and can be triggered if the participant is in the Random or Heuristic group. Based on the output of the intervention software, the experimenter triggers the Pulse accordingly from outside of the room (i.e. a 'Wizard of Oz experiment'). The intervention software is monitoring the data coming from the E4 in real-time, and the feature extraction for the Heuristic algorithm is also running with online data.

After the speech has reached a natural conclusion or 5 minutes has passed (whichever happens first), the VR headset is removed and the recording of the speech and E4 data stops. The participant answers the POST self-reports (the PPP, AEBS, and Vibrotactile survey). When the surveys are completed, we export the speech from Audacity into CARMA so the participant can annotate their stress levels for each second of the speech. After annotations are finished, this concludes the PPT component, and also one TEST phase. The data collected from this component is also referred to as the *PPT* data.

2.6.3 POST Phase

Once four TEST sessions are completed, the participant completes the state scale of the STAI and CAI, the VRP, BSQ, and the Post Experiment Feedback questionnaire. After the POST surveys are completed, the individual is compensated with a $25 Amazon gift card for their participation.

3. METHODS

3.1 Data Pre-Processing and Feature Extraction

We use the aforementioned modalities to extract physiological and speech features from the participant in real time. In addition to specific feature extraction, some of these features may use functionals like the arithmetic mean or the coefficient values of regression models that were fit based on the data, as a method of condensing the information of the corresponding signals.

3.1.1 EDA Features

The raw EDA signal is measured by driving a micro current across the skin and calculating the resistance of the skin. The inverse of this resistance is the conductance, which is EDA. The base EDA signal can be decomposed into two resultant signals. The first is the tonic component of the EDA signal, usually called the skin conductance level (SCL) and represents the slow-changing aspect of the EDA signal. The second is called the phasic component of the EDA signal, sometimes referred to as skin conductance response (SCR), though SCR typically refers to the peaks *within* the phasic component. The phasic component reflects the fast-changing aspects of the EDA signal. These changes are controlled by the sympathetic nervous system, making it very relevant to stress and anxiety detection. Special peak detection responses can be used on the phasic component to detect peaks, sometimes called SCR peaks. These SCR peaks typically reflect the introduction to an external stimulus. A higher frequency of SCR peaks within a certain window of time is representative of increased psychological arousal [96]. SCR peaks are closer to being participant-agnostic over mean SCR or SCL levels due to being mostly event-driven signals. The following features derived from EDA are used in this experiment:

- *Skin Conductance Level* (SCL): The tonic, or slow-acting component of the total EDA signal.
- *Skin Conductance Response* (SCR): The phasic, or fast-acting component of the total EDA signal, that is typically caused by stressful events or stimuli.

- *SCR Peaks*: Peaks or abnormal amplitudes within the SCR signal, which typically corresponds to events or stimulus introduction. If we divide the number of SCR peaks by the time duration in wcich they occurred, we obtain the SCR Frequency.

3.1.1.1 Extraction Method

We utilize various components of the NeuroKit2 package for EDA feature extraction. The raw EDA signal is filtered with a low-pass Butterworth filter with a cutoff frequency of 1.5 Hz (limited by the 4 Hz sampling rate). Then, a Blackman kernel of size 8 is applied to smooth the resultant signal. Finally, we use NeuroKit2's built-in peak detection algorithm which uses the method from [97]. These methods are consistent with typical pre-processing of EDA signals [98, 83].

3.1.2 BVP Features

The PPG generates a waveform which is the result of an optical measurement of light when shone through the skin from the surface of the skin. This allows for a non-invasive method of measuring heart rate and heart rate variability [99]. Increased heart rate is a typically an indicator of increased levels of stress [27], but it can be difficult to compare across individuals due to their relative differences in elevated heart rate. Heart rate variability has found success in quantifying individuals with high anxiety. Within HRV there are time-based and frequency-based metrics, both of which can be used to estimate anxiety, though time-based are often more appropriate when using shorter duration analysis windows.

Each heartbeat is calculated from the BVP signal. The root-mean square differences of successive R-R intervals (RMSSD) is the typical metric people use when discussing low or high heart rate variability in the time domain. The standard deviation of interbeat intervals (SDNN) is also commonly discussed [54]. RMSSD, SDNN, and pNN50 are time-domain features and viable in shorter duration windows [100]. Prior work indicates that high heart rate variability correlates negatively with stress and anxiety, since high heart rate variability indicates high levels of adaptive variability or homeostasis of the autonomic nervous system [28]. Ultra-short time HRV is defined as any HRV derivation using less than 5 minutes for the duration of the analysis window. Typi-

cally, RMSSD and SDNN are of the few that are valid with this constraint [54]. The following list describes the features we extract from the BVP signal:

- *Heart Rate* (HR): Heart rate of the participant, measured in beats per minute.
- *RMSSD*: Root-mean-square of successive differences between heartbeats (RMS of differences of IBI).
- *SDNN*: Standard deviation between NN (equivalent to IBI) intervals
- *pNN50*: Percentage of adjacent NN intervals that differ by 50%
- *HF*: Absolute power of the high frequency band (0.15 – 0.4 Hz)
- *LF*: Absolute power of the low frequency band (0.04 – 0.15 Hz)

3.1.2.1 Extraction Method

We use NeuroKit2 to detect the peaks in the PPG/BVP signal and also to extract frequency- and time-domain HRV features. After the peaks in the BVP signal are determined, we can estimate heart rate by measuring the distance between peaks, also known as the interbeat intervals (IBI). Continuous heart rate is then estimated with cubic interpolation. HF and LF are computed by calculating the absolute power of their relative bands based on heart rate oscillations.

3.1.3 Speech Features

Acoustic markers extracted from speech signals are commonly used in emotion recognition applications, since they capture spectrotemporal and prosodic variations of speech that are typically indicative of changes in affect [56, 57, 101]. Speech is also much more reliable in short duration analysis windows [72], making the extracted features accurate for comparison. We extract the following features from the speech signal:

- *ZCR*: The rate at which the speech signal changes signs (i.e. positive to negative or negative to positive)

- *RMS Energy*: The energy of a discrete signal is the sum of its squared values. RMS energy is this sum divided by the number of samples in the signal, then square rooted.
- *F0*: Sometimes regarded as pitch, the fundamental frequency reflects the frequency of oscillation of the vocal folds and is defined as the average number of such oscillations per second, measured in Hertz.
- *Jitter*: The variation from the fundamental frequency in the signal.
- *Shimmer*: The variation in amplitude when comparing neighboring amplitudes within three periodic amplitudes in the signal.

3.1.3.1 Extraction Method

We use the ComParE2016 features set from the OpenSMILE toolkit [84]. OpenSMILE computes these features over the entire speech window and applies functionals (like averaging) to summarize the statistics gathered in each segment of the window.

3.2 VerBIO Dataset

The previous study [30] with which this thesis extends had participants partake in four days of public speaking sessions. One day of real audience speaking, two days of VR audience speaking, and one final day of real audience speaking. The goal of the study was to measure the improvement of real public speaking as a result of the subsequent exposure to the VR public speaking.

The dataset which was released as a result of this study has the same contextual signals used in this thesis (EDA, BVP, Speech) as well as manual annotations from four third-party annotators. These annotators ranked the perceived stress of the participant on a scale of 1 to 5 (5 being high stress) by listening to a playback of the participant's speech. These ratings are at a resolution of every second. Each participant completed 8 VR speeches. The dataset also contains the baseline measurement levels (without speech) recorded while the participants watched a relaxation video.

This dataset was used extensively to guide the design of the intervention algorithms in this thesis, as well as in the latter part of the thesis which focuses on personalization. The dataset, here-

after referred to as the ‘v1’ dataset, contains a relatively large number of samples making it useful as general measure of the population for assessing the conditions of an individual experiencing stressful moments.

3.3 Intervention Design

It is unclear when the optimal moment to deliver an intervention is. Jaimes and Steele (2018) hypothesized three points at which to apply an intervention; before, during, and after the heightened moment of stress [102]. However, there has been little work in evaluating the effectiveness of interventions for the aforementioned three different timings. However, interventions delivered "just-in-time" are indeed effective [103, 104]. Furthermore, King et al. (2000) showed that providing immediate feedback for speech performance was ideal [105]. Therefore, for the purposes of this study, the interventions will be delivered during the *onset* of the stressful moment in a "just-in-time" manner.

For very short duration sessions like the one in this experiment, to our knowledge the optimal time to deliver an intervention has not been evaluated. Therefore, there is a need to explore algorithms that are aware of the state of the user and act accordingly. Instead of developing a machine learning model trained on the annotations from the ‘v1’ VerBIO dataset, we instead opted to develop a simple rule-based algorithm to investigate the differences between delivering an intervention while the user is in a heightened state of arousal compared to delivering the intervention at random moments in the session. A rule-based algorithm has the benefit of being highly interpretable and allows future work to build on it. The two algorithms developed are called *HEURISTIC* and *RANDOM*.

3.3.1 Heuristic Algorithm

The HEURISTIC algorithm takes the physiological state of the participant into account during the public speaking session. It accomplishes this by estimating the number of SCR peaks (‘SCR frequency’) within a 30 second analysis window. Note that these would be considered non-specific skin conductance responses (NS-SCR) as they aren’t a reaction to any sort of introduced stimulus.

If it observes at least 10 peaks, the intervention is turned on (or stays on). Otherwise, the intervention is turned off (or stays off). This is not unlike other algorithms that deliver prompts based on a breathing rate threshold of 7 breaths per minute [106]. HEURISTIC updates its count of peaks every second by ingesting new data from the E4, but always stays limited to 30 seconds, recycling old data as necessary. This is to ensure we have a fairly accurate context of the user's state, as keeping information from the very beginning of the session wouldn't reflect the current state of the user.

The decision of using at least 10 peaks over 30 seconds was determined based off of analysis of the v1 VerBIO dataset. We observed that windows with at least 10 SCR peaks occurred in 30% of the possible analysis windows in the dataset for each participant. We empirically decided that 30% was a reasonable tradeoff between stimulus and non-stimulus time. With this percentage we will not overstimulate the participant but will have enough delivery time such that we may observe the effects of the stimulus. The frequency of NS-SCR has been shown to indicate increased physiological arousal and stress [107], and 20 NS-SCR in a minute appears to be a general 'rule-of-thumb' for high physiological arousal according to [107, 96].

The pseudocode for HEURISTIC is contained in Algorithm 1. Also, see Figure 3.1 for an example of the execution of HEURISTIC, which has the HEURISTIC algorithm being executed in action and is colored with the self-reported stress annotations of the participant. Note that the HEURISTIC algorithm doesn't depend on the annotation levels in any way, this is just for a demonstration.

3.3.2 Random Algorithm

We designed the RANDOM algorithm to match the same amount of intervention delivery time as the HEURISTIC algorithm. To determine when to trigger the intervention, we simulated the HEURISTIC algorithm on the VerBIO dataset and observed the positive-edge (i.e. turning the intervention on) and the negative-edge (turning the intervention off) triggers for each participant. On average, when the intervention turns on, it stays on for 5 seconds, and then stays off for 15 seconds. Therefore for our random algorithm, it has a 30% chance of triggering for each second,

Algorithm 1 HEURISTIC: Algorithm which triggers intervention if at least 10 SCR peaks are detected in latest 30 second window.

$t_i \leftarrow 0$ seconds
$t_{start} \leftarrow 30$ seconds
$t_{max} \leftarrow 530$ seconds
$x_{win} \leftarrow \emptyset$
$intervention \leftarrow$off
while $t_i < t_{start}$ **do**
    $t_i \leftarrow$${TIME} ▷ Poll system time
end while
while $t_i < t_{max}$ **do**
    if x_{win} is filled **then**
        $n_{peaks} \leftarrow eda\_peaks(x_{win})$
        if $n_{peaks} \geq 10$ **then**
            $intervention \leftarrow$ on
        else
            $intervention \leftarrow$ off
        end if
        $x_{win} \leftarrow x_{win} \setminus x_{win}[0]$ ▷ Pop first second of data off window
    end if
    $x_{win} \leftarrow x_{win}+$incoming data
    $t_i \leftarrow$${TIME}
end while

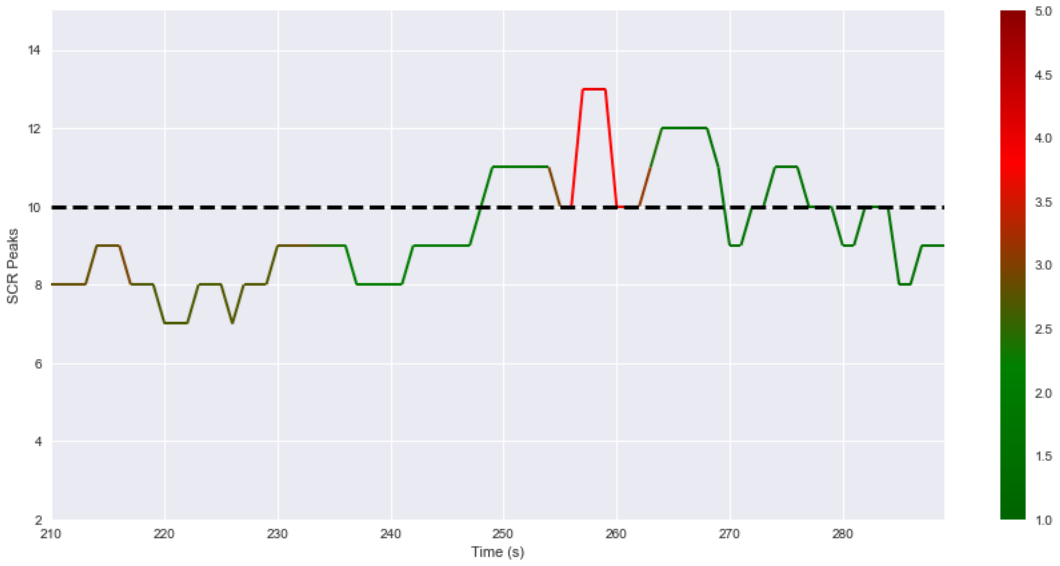


Figure 3.1: Example of Heuristic execution with the 'heat' of the line being the self-reported annotation. Intervention triggers when SCR peaks goes above the dashed line.

and if it triggers, it stays on for 5 seconds, and then stays off for 15 seconds, after which it is again possible to trigger. The enforced 15 seconds of off-time allows us to observe physiological after-effects of the stimulus without bleeding into the next intervention. The pseudocode for RANDOM is contained in Algorithm 2.

3.3.3 JITAI Representation

Following the JITAI framework, we define the following attributes of our interventions:

- Our *distal* outcome is defined as an improved overall decrease of PSA compared to a control without the intervention while using the VR platform to be exposed to the possibility of PSA.
- Our *proximal* outcome is defined as an improvement in stress-related physiological signals when comparing the difference of these signals before and after the intervention is delivered.
- The *tailoring variables* correspond to the physiological signals of the user. For the heuristic group, these are SCR peaks, but could be extended in future work as decision rules and thresholds become more adaptable.

Algorithm 2 RANDOM: Algorithm which has a 30 percent chance of triggering every second. If it triggers, the intervention is enabled for 5 seconds, with a cooldown period of 15 seconds.

$t_i \leftarrow 0$ seconds
$t_{start} \leftarrow 30$ seconds
$t_{max} \leftarrow 530$ seconds
$intervention \leftarrow$off
$t_{next} \leftarrow t_{start}$ ▷ Next available time intervention can trigger
while $t_i < t_{start}$ **do**
    $t_i \leftarrow$${TIME} ▷ Poll system time
end while
while $t_i < t_{max}$ **do**
    if $t_i > t_{next}$ **then** ▷ We are allowed to enable the intervention
        if ${RANDUNIFORM} ≤ 0.3 **then** ▷ $P(trigger) = 0.3$
            $intervention \leftarrow$ on for 5 seconds
            $t_{next} \leftarrow t_i + 20$ ▷ Next trigger is blocked until 20 second have passed
        else
            $intervention$ stays off
            $t_{next} \leftarrow t_{next} + 1$
        end if
    end if
    $t_i \leftarrow$${TIME}
end while

- *Decision thresholds* are defined for the heuristic group as greater than or equal to 10 SCR peaks in a 30 second analysis window.
- *Decision points* are defined as every second in the speech session since our analysis window has a stride of 1 second.
- *Intervention options* are the delivery of the vibrotactile stimulus at a rate of 60 bpm while the decision threshold holds.

3.4 Personalization

Given that the magnitude of physiological data varies from person to person, we can achieve much better accuracy in detecting moments of stress if we apply personalization techniques to our dataset that tailor user-specific models [59]. The dataset that emerges from this experiment (and previous experiments like [30]) is sparse per-participant, and after generating analysis windows

for proper feature extraction, we only end up with less than 100 samples per participant. For the samples that we *do* have for the participant, allowing us to make these samples comparable to the other samples from the general experiment population may improve model performance by allowing us to train models with more data. We will assess the effects of personalization techniques on the performance of Gradient Boosting Trees [108] using the XGBoost [109] library. We will evaluate the boosted trees on two tasks – the first is classifying high levels of stress from the self-reported annotations, and the second is classifying increasing levels of stress from decreasing (or non-changing) levels of stress, which we can estimate by doing regressions on the self-reported annotations.

For each participant, we will fine-tune/train/validate using the first two TEST sessions (TEST01 and TEST02) and evaluate the final performance on a hidden test set constructed from TEST03 and TEST04. Performance will be evaluated per-participant and then averaged for an overall estimate across the experiment population. We define three personalization techniques that will be used in combination and individually so we can access their cumulative or solo effects on the performance of these classification tasks. The strategies are listed as follows:

- *Technique 1*: Z-score normalization of PPT data using the mean and standard deviation of any corresponding features captured in the RELAX data (this is done on a per-participant basis).
- *Technique 2*: Fine-tuning pre-trained base models on TEST01 only or TEST01 and TEST02. The base models will be trained on data from the v1 VerBIO dataset.
- *Technique 3*: Including the participant's trait-scale anxiety scores taken from the PRE phase into the model as a feature for context. The idea behind this is that participants with different trait characteristics may depict distinct physiological and acoustic patterns of anxiety.

3.4.1 Technique 1

Due to individual differences in physiology levels, technique 1 allows us to realign the data of each participant to be comparable with other participants. This has the additional benefit of

providing the model with context on how reactive the participant is at that moment relative to how their reactivity in a relaxed state.

Formally, for each participant i and baseline feature f, we can define a mean μ_{if} and standard deviation σ_{if}. For every later PPT data sample x_{if} for participant i and feature f, if f is a baseline feature, we will rescale x_{if} such that

$$\bar{x_{if}} = \frac{x_{if} - \mu_{if}}{\sigma_{if}} \tag{3.1}$$

This is effectively calculating the z-score for each sample relative to the mean and standard deviation of the baseline features in the RELAX data. Baseline features are those derived from EDA and BVP and are computed by averaging the 30 second windows in the RELAX component. This allows us to have a mean window and a standard deviation among the windows. Note that we won't be able to apply this to speech, since the participant doesn't speak during the RELAX component.

3.4.2 Technique 2

The second strategy follows a similar reasoning behind the first, but this has the benefit of training the model on captured features from the participant rather than relying on a global population estimate. It also makes the model more adaptive by capturing recent changes in the individual's anxiety. Models that are trained on individual-specific data have been shown to be more accurate than those trained on data from a larger population [59], so we will observe whether this idea holds for stressful moments with a shorter analysis duration.

Since we're using gradient boosting trees, the 'fine-tuning' step just involves training a pre-trained (or empty) tree on additional data from the target user for additional boosting rounds. This way, a fraction of the decision stumps in the ensemble are tailored towards predicting for the individual, and the remaining are tuned to predict for a general population (if pre-trained). The number of additional boosting rounds varies based on the amount of individual-specific data used (10 rounds for just TEST01, 20 rounds for TEST01 and TEST02). Specific parameters of the gradient boosting ensemble will be presented in the Results section after optimization has been

performed.

3.4.3 Technique 3

The idea behind the third strategy is that the trait scores for each participant may be used to offset the model's expectations of reactivity for the participant. If the trait levels for anxiety of an individual are higher, the model may become more sensitive to classifying more instances as high anxiety. In practice, we calculate trait scores and append them to the end of each input vector on a per-participant basis. So, all vectors for that participant would have the same tail for their samples, but each participant has a different tail (unless the participants score the same on the surveys).

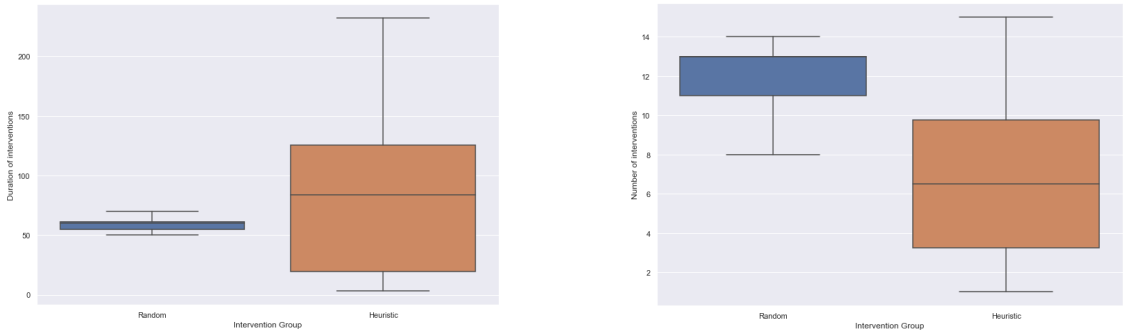
3.4.4 Personalization Operators

In order to simplify naming later on, we will define the following personalization operators that we can apply to customize our model:

- *BASE*: Model trained on the v1 VerBIO dataset for 50 boosting rounds
- *MOD*: Model trained with Z-score normalization for the participant's PPT data using the RELAX data (Technique 1)
- *TEST01*: Model fine-tuned with TEST01 PPT data of the participant for 10 boosting rounds (Technique 2)
- *TEST**: Model fine-tuned with TEST01 and TEST02 PPT data of the participant for 20 boosting rounds (Technique 2)
- *TRAIT*: Model trained with the additional context of the CAI-Trait and STAI-Trait scores per participant for every sample input of that participant (Technique 3).

4.1 Number and Duration of Interventions

Figure 4.1 shows the differences in the total number of interventions delivered and total duration of stimulus between the Random and Heuristic groups for each session. We can observe that in practice the Random group had lower overall duration of intervention delivery time, but a slightly higher number of intervention triggers throughout their sessions. These number are still comparable, though. The number of interventions and duration for the Heuristic clearly has a much higher distribution spread compared to the Random group, most likely due to individual differences in SCR frequencies.



(a) Random: 56.7 ± 9.2; Heuristic: 85.0 ± 62.3 (b) Random: 11.7 ± 2.1; Heuristic : 7.8 ± 6.6

Figure 4.1: Box plot of (a) intervention duration and (b) total number of intervention activations

To estimate an effect between the number of interventions and intervention duration received by a participant and overall physiology outcomes for that session, we performed a repeated measures correlation for the Heuristic group. We use a repeated measures correlation [110] over a regular correlation because each overall feature value for each TEST session is not independent to the rest of the participant's in the group. Surprisingly, most of the significant effects we observe are based on the number of intervention activations in the session, and not the duration (our expectation was

that more intervention activations might interrupt the participant more). In fact, the only near-significant effect we observe for duration was found for heart rate with a correlation coefficient of $r(29) = -.35, p = 0.071$, possibly confirming previous effects of other studies observed on heart rate, though more data would be needed to confirm the statistical significance. The results for the number of intervention activations are summarized in Table 4.1.

Feature	Correlation
LF	$r(29) = -.45, p = .070$
SDNN	$r(29) = .36, p = .015$
RMSSD	$r(29) = .35, p = .062$
Shimmer	$r(29) = .41, p = .036$
F0	$r(29) = .51, p = .008$
RMS Energy	$r(29) = .57, p = .003$
ZCR	$r(29) = .42, p = .032$

Table 4.1: Repeated measures correlation results for number of interventions within the Heuristic group. Note that degrees of freedom are calculated as $N(k-1)-1$, where $N = 10$ is the number of unique participants and $k = 4$ is the average number of samples per participant

To better understand how the interventions were delivered across time, Figure 4.2 has two examples of where the interventions start and stop, as well as when they're enabled for the Heuristic group. Figure 4.3 also has an example of the intervention being delivered for the Random group.

4.2 Proximal Effects of Vibrotactile Feedback

To analyze the proximal, or *local* effects of vibrotactile biofeedback between the Random and Heuristic groups, we observe the physiological state of the participant and their speech characteristics before the intervention starts, during the intervention, and after the intervention ends. Note that the size of the analysis windows for these vary by the feature being analyzed. We define the following window types:

- *before*: The end of the window is at the point at which the intervention starts, and extends backwards in time according to the analysis window duration.

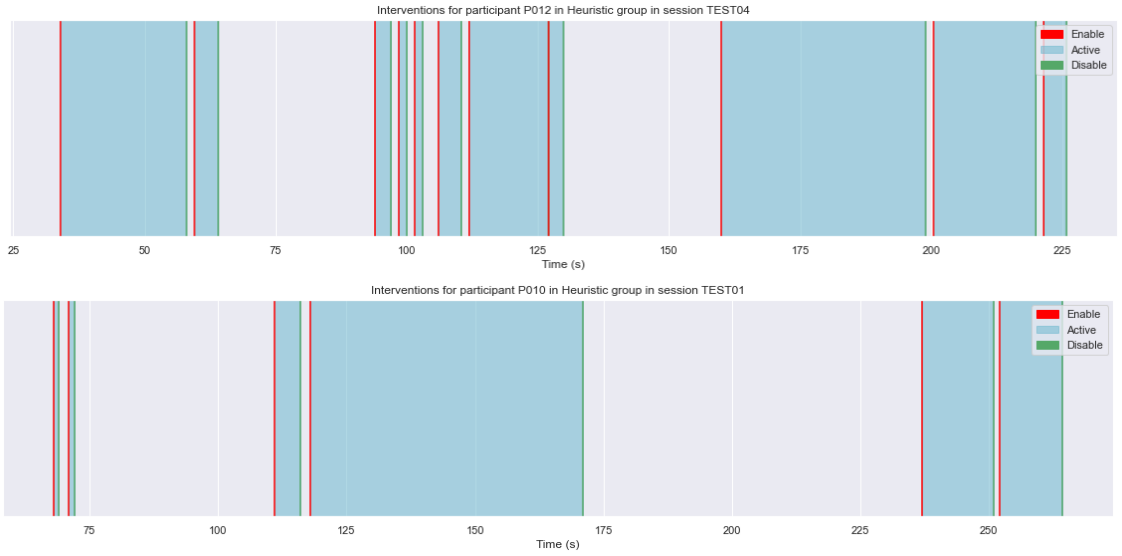


Figure 4.2: Two plots of activations and triggers for Heuristic intervention. Top is for P012, bottom is for P010

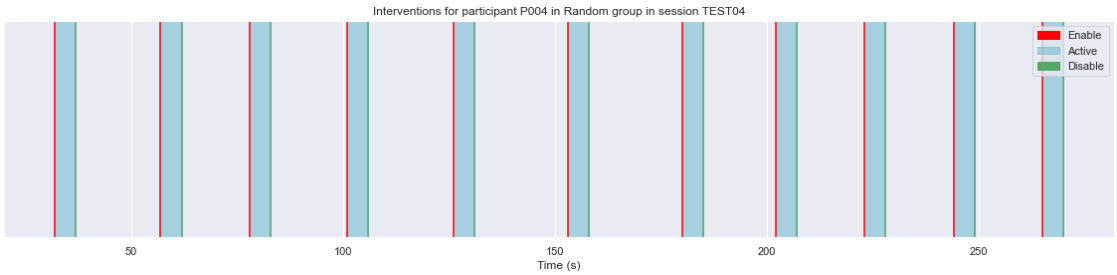


Figure 4.3: Plot of activations and triggers for Random intervention for P004

- *during*: The beginning of the window starts 3 seconds after the intervention starts, and extends forwards in time according to the analysis window duration.
- *after*: The beginning of the window starts at the point at which the intervention ends, and extends forward in time according to the analysis window duration.

The analysis window duration varies according to the following rules:

- BVP- and EDA- related features use a 15-second analysis window, since both of those signals often require longer analysis windows for accurate analysis [96, 54]. Frequency-related

characteristics of HRV are not valid for this duration, so only RMSSD and SDNN (as well as HR) are computed.

- Speech-related features use a 10-second analysis window. Stress-related changes in speech works better with shorter analysis windows according to [72], which is why this window is shorter than the one for EDA and BVP.

- We use a 5-second window and apply functionals like mean, median, max, and slope to the annotations to observe the immediate impact on the participant's stress self-reported stress levels.

To analyze the differences between groups and across interventions within each session, and across the sessions themselves, we model the nested factors with a linear mixed effects model. First, we define the following variables:

- *igroup*: The intervention group the participant was assigned to (Control, Random, Heuristic)

- *session*: The numerical index of the session (1, 2, 3, and 4 for TEST01, TEST02, TEST03, TEST04)

- *ivn_no*: The numerical index of the intervention within a particular session (e.g. 3 for the 3rd intervention seen so far in TEST02)

- *win_type*: The type of analysis window the feature was computed over (e.g. before, during, or after)

- *pid*: The unique participant identifier code

- *value*: The numerical value of the feature computed over the analysis window

igroup, session, ivn_no, and *win_type* are all fixed effects in the mixed effects model, as we wish to observe the differences in physiological signals as the sessions progress and as more interventions are received, and compare them between window types and intervention groups. *pid*

is used as a random effect to account for individual variance in participants. *value* is the dependent variable the mixed effects model is trying to estimate. One thing to note is that if there is a categorical variable as a fixed effect, the variable is automatically coded to a binary variable by the software (e.g. for *igroup* this would be something like *Random* and *Heuristic* as separated variables being true or false relative to the *Control*). Also note that using the '*' symbol for any two fixed effects a and b like $a * b$ expands automatically to $a + b + a : b$ to include both the main effects of a and b as well as their nested interactions. Equation 4.1 denotes the abbreviated formula using the $*$ symbol for the mixed effects model, and Equation 4.2 denotes the full formula. We use the *lmer* model from the *lme4* R package [111]. All tables were generated by the *texreg* R package [112] and modified according to our needs.

$$value \sim igroup * session * ivn\_no * win\_type + (1 \mid pid) \tag{4.1}$$

$$\begin{aligned} value \sim\ & igroup + session + ivn\_no + win\_type + igroup : session + \\ & igroup : ivn\_no + igroup : win\_type + igroup : session : ivn\_no + \\ & igroup : session : win\_type + igroup : session : ivn\_no : win\_type + session : ivn\_no + \\ & session : win\_type + session : ivn\_no : win\_type + ivn\_no : win\_type + (1 \mid pid) \end{aligned} \tag{4.2}$$

For each significant effect, we also perform a power sensitivity analysis to determine the minimum detectable effect size (MDES) with a power threshold of 80%, $\alpha = 0.05$, and $N = 30$.

4.2.1 Differences between Random and Heuristic

We ran the mixed effects model on the experimental data which we filtered to just include the Heuristic and Random group to minimize the length of the output results, and the *before* and *after*

analysis windows to better isolate the effect of the pre/post effect of each intervention. For the after-effects of the intervention, we observe that as more interventions are received for the Heuristic group, the heart rate at the end of the intervention decreases (-0.72) more over time compared to the Random group, which actually shows an increase the more interventions are received (1.03). However, as sessions increase, this effect starts to diminish for the Heuristic group (0.27) up to a point where the heart rate starts increasing as more interventions are received. The heart rate increase per intervention experienced by the Random group starts to decrease across sessions as well (-0.29). These effects are captured in table 4.2. A sensitivity analysis suggests that the effect on session for both the Heuristic and Random group, which had a minimum detectable effect size (MDES) of -0.55 and 2.70, respectively, is underpowered as the observed effect size is smaller than the MDES. The effects from the number of interventions are robust from the sensitivity analysis, however.

These results indicate that, at least in earlier sessions, the targeted interventions offered by the Heuristic algorithm produce a benefit to heart rate reduction over time. However, as the participant performs more sessions, this benefit fades away, and in contrast, the Random group starts to benefit from a reduction in heart rate over time. This suggests that the Random group may just need more time to get used to the somewhat frequent delivery of the vibrotactile sensations before experiencing a calming effect.

Regarding vocal characteristics, we observe an effect for shimmer as well. For the Heuristic group, across interventions we observe no significant change in shimmer, but for the Random group, we do indeed observe a significant effect across interventions ($2.0e$-3), potentially indicating a decrease in vocal stability for the participants speaking in the random group as each intervention is received. This across-interventions effect decreases across sessions however ($-1.1e$-3), again indicating that as the participant receives more random interventions, their voice stability improves. However, for the Heuristic group, speech stability across interventions appears to be unchanged. The effect sizes of these results are presented in 4.3. A sensitivity analysis reveals the diminishing effect across sessions to be robust, but the effect across interventions is less than the MDES

Fixed Effect	Estimate	Std. Error
(Intercept)	79.84^{***}	2.39
Random	-5.76	3.39
Session	-1.67^{*}	0.69
Intervention #	-0.72^{***}	0.19
Before	0.23	2.47
Random: Session	1.88^{*}	0.96
Random: Intervention #	1.03^{**}	0.34
Session: Intervention #	0.27^{**}	0.10
Random: Before	-0.25	3.56
Session: Before	0.19	0.96
Intervention #: Before	0.31	0.27
Random: Session: Intervention #	-0.29^{*}	0.15
Random: Session: Before	-0.68	1.35
Random: Intervention #: Before	-0.27	0.48
Session: Intervention #: Before	-0.19	0.14
Random: Session: Intervention #: Before	0.24	0.20

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.2: Mixed effects results on HR with Heuristic group and *after* window type as reference.

$(2.3e\text{-}3)$.

4.2.2 Differences between Control and Heuristic

To understand the differences between delivering vs. not delivering vibrotactile feedback according to the HEURISTIC algorithm, we 'simulate' HEURISTIC on the Control group's physiological data to gather the windows in which we *would have* delivered the intervention if the participants were in a different group. We also use the same configuration for the mixed effects model that was used in the Random and Heuristic comparisons, but instead filter the data for just the Heuristic and Control group to reduce the output size of the table, as well as just the *before* and *after* windows for a pre/post comparison.

While the disabling of the intervention delivery by the HEURISTIC algorithm is contingent on a reduction in SCR peaks (i.e. less than 10 SCR peaks), we still notice significant differences in the SCR peaks after the intervention ends. For the Control group, as the participant experiences more sessions, the number of peaks detected after the intervention ends increases (0.347), however

Fixed Effect	Estimate	Std. Error
(Intercept)	$169.5e$-3^{***}	$8.3e$-3
Random	$-10.3e$-3	$11.6e$-3
Session	$-5.2e$-3^{**}	$1.7e$-3
Intervention #	$-0.1e$-3	$0.4e$-3
Before	$-1.3e$-3	$5.8e$-3
Random:Session	$6.5e$-3^{**}	$2.4e$-3
Random:Intervention #	$2.0e$-3^{*}	$0.8e$-3
Session:Intervention #	$0.3e$-3	$0.2e$-3
Random:Before	$3.6e$-3	$8.7e$-3
Session:Before	$1.6e$-3	$2.3e$-3
Intervention #:Before	$0.9e$-3	$0.6e$-3
Random:Session:Intervention #	$-1.1e$-3^{**}	$0.3e$-3
Random:Session:Before	$-2.7e$-3	$3.3e$-3
Random:Intervention #:Before	$-0.9e$-3	$1.1e$-3
Session:Intervention #:Before	$-0.7e$-3^{*}	$0.3e$-3
Random:Session:Intervention #:Before	$0.8e$-3	$0.5e$-3

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.3: Mixed effects results on shimmer with Heuristic group and *after* window type as reference.

for the Heuristic group, this number *decreases* across sessions (-0.362). This implies that the presence of the interventions decreases the number of SCR peaks after the intervention concludes, and this reduction improves as more sessions are experienced by the participant in the Heuristic group. We can reasonably conclude that the disabling of the intervention provides some relief for physiological arousal (SCR peaks in this instance). The effect sizes for these results are reported in Table 4.4. However, neither of these effect sizes are robust according to the sensitivity analysis. The Control group MDSE calculated was 0.40, and -0.48 for the Heuristic effect, both of which are larger than their corresponding observed effects.

We also observe significant differences in SDNN and RMSSD, which are proxy indicators to HRV. While the intercepts of RMSSD for the Heuristic group trend higher (196.60) compared to the Control (220.82), both SDNN (-69.87) and RMSSD (-102.28) show little improvement across sessions compared to the Control group, which has increases in both SDNN (71.38) and RMSSD

Fixed Effect	Estimate	Std. Error
(Intercept)	3.709^{***}	0.451
Heuristic	0.051	0.608
Session	0.347^{**}	0.133
Intervention #	-0.010	0.053
Before	1.820^{***}	0.519
Heuristic:Session	-0.362^{*}	0.183
Heuristic:Intervention #	-0.037	0.063
Session:Intervention #	-0.007	0.019
Heuristic:Before	-1.344	0.686
Session:Before	-0.392^{*}	0.186
Intervention #:Before	-0.100	0.074
Heuristic:Session:Intervention #	0.041	0.026
Heuristic:Session:Before	0.473	0.255
Heuristic:Intervention #:Before	0.139	0.088
Session:Intervention #:Before	0.035	0.026
Heuristic:Session:Intervention #:Before	-0.068	0.0371

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.4: Mixed effects results on SCR peaks with Control group and *after* window type as reference

(102.66) across sessions. This implies delivering the vibrotactile stimulus briefly supports one's parasympathetic response to speaking anxiety, but this increased homeostasis is not sustained for the entire session. These results are captured in Tables 4.5 and 4.6. The intercept for RMSSD of the Heuristic group differs from the MDES (270) from the sensitivity analysis, as well as the across session effect of SDNN in the Control group (-74.5). The remaining effects are robust from the sensitivity analysis.

Similar to the effects observed in 4.2 between the Heuristic and Random group, we analyze only the *during* window between the Heuristic and Control group. Heart rate again shows to have a significant difference in the Heuristic group compared to the Control group. Heart rate tends to increase as interventions increase for the Control group (0.74), but for the Heuristic group it tends to decrease as interventions increase (-1.19). This suggests that as interventions are delivered in each session for the Heuristic group, their heart rate decreases more *during* the intervention itself,

Fixed Effect	Estimate	Std. Error
(Intercept)	173.78^{**}	53.23
Heuristic	127.45	70.67
Session	71.38^{***}	18.19
Intervention #	18.32^{*}	7.22
Before	81.17	71.02
Heuristic:Session	-69.87^{**}	24.91
Heuristic:Intervention #	-18.79^{*}	8.66
Session:Intervention #	-7.81^{**}	2.59
Heuristic:Before	-73.97	93.54
Session:Before	-39.91	25.49
Intervention #:Before	-14.42	10.07
Heuristic:Session:Intervention #	7.65^{*}	3.61
Heuristic:Session:Before	33.77	34.86
Heuristic:Intervention #:Before	14.42	12.03
Session:Intervention #:Before	5.68	3.59
Heuristic:Session:Intervention #:Before	-4.70	5.04

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.5: Mixed effects results on SDNN with Control group and *after* window type as reference.

Fixed Effect	Estimate	Std. Error
(Intercept)	220.82^{**}	73.44
Heuristic	196.60^{*}	97.55
Session	102.66^{***}	25.02
Intervention #	23.80^{*}	9.94
Before	123.02	97.68
Heuristic:Session	-102.28^{**}	34.27
Heuristic:Intervention #	-24.81^{*}	11.92
Session:Intervention #	-10.16^{**}	3.56
Heuristic:Before	-118.92	128.64
Session:Before	-56.32	35.06
Intervention #:Before	-20.13	13.85
Heuristic:Session:Intervention #	10.46^{*}	4.97
Heuristic:Session:Before	46.48	47.94
Heuristic:Intervention #:Before	20.98	16.55
Session:Intervention #:Before	7.86	4.94
Heuristic:Session:Intervention #:Before	-6.80	6.93

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.6: Mixed effects results on RMSSD with Control group and *after* window type as reference.

Fixed Effect	Estimate	Std. Error
(Intercept)	72.06^{***}	2.53
Heuristic	6.49	3.40
Session	0.40	0.79
Intervention #	0.74^{*}	0.31
Heuristic:Session	-1.59	1.09
Heuristic:Intervention #	-1.18^{**}	0.38
Session:Intervention #	-0.28^{*}	0.11
Heuristic:Session:Intervention #	0.42^{**}	0.15

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.7: Mixed effects results on HR with Control group as reference. Results are for the *during* window type.

and is a positive result in reducing state anxiety. The effects for these interactions are outlined in Table 4.7. Comparing these effect sizes from the MDES calculated from the sensitivity analysis, thte effect for the Heuristic group is robust but the effect for the Control group is less than the MDES (0.83).

We also see an effect for pNN50 during the intervention where the pNN50 measurement is generally higher for the Heuristic group (10.43) but decreases across sessions for the *during* analysis window (-2.72). These effects are in Table 4.8. This is similar to the previous effects we observed for HRV-related effects where HRV for the Heuristic group was higher earlier on but decreased (or had no improvement) across sessions. The intercept for the Heuristic group is not robust from the sensitivity analysis (12.7), but the across session effect is robust.

4.3 Overall Effects of Vibrotactile Feedback

We use a similar mixed effects model structure compared to the proximal differences but we remove the *win_type* and *ivn_no* variables since we're computing the feature values over the data collected in the entire session. As before, we also perform a sensitivity analysis for relevant effects.

We did not observe significant differences in any measure derived from the overall sessions. We also observed no significant differences between the slopes of the heart rate or EDA time-series. Differences between the Random and Control groups for SDNN and RMSSD were approaching

Fixed Effect	Estimate	Std. Error
(Intercept)	71.00^{***}	3.64
Heuristic	10.43^{*}	4.94
Session	1.74	0.97
Intervention #	0.54	0.38
Heuristic:Session	-2.72^{*}	1.33
Heuristic:Intervention #	-0.67	0.46
Session:Intervention #	-0.03	0.14
Heuristic:Session:Intervention #	0.18	0.19

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.8: Mixed effects results on pNN50 with Control group as reference. Results are for the *during* window type.

significance with $p = 0.0509$ and $p = 0.0654$ but did not reach our threshold. The effect size of SDNN (-384.78) and RMSSD (-460.21) would indicate lower overall levels of HRV for the Random group, though, implying that the Random interventions may have negatively impacted participants while the Heuristic had no such negative impact.

One way we do observe a significant effect is if we assume the first sessions was a *training* session (i.e. the user may have been startled and not yet accustomed to receiving the vibrotactile biofeedback in the first session, so feature measures may have been inflated). If we run the mixed effects model on the experiment data but omit the data collected from the first session (TEST01), we observe significant effects on SDNN, RMSSD, and HR between the Random and Control groups. The Random group has *significantly* lower SDNN (-676.39) and RMSSD (-831.14) compared to the Control group (830.16 and 1026.17), indicating that the Random delivery of the vibrotactile biofeedback leads to overall lower HRV metrics for the participants in the Random group, which suggests their overall state anxiety is worse. From the sensitivity analysis, the intercepts for the Random group for SDNN and RMSSD weren't robust with MDES values of -710 and -900, respectively.

For the slope of the heart rate computed over each session, the Heuristic group has a lower slope (-0.0921) compared to the Control group (0.0935), which means the heart rate increases

Fixed Effect	Estimate	Std. Error
(Intercept)	830.16^{***}	201.24
Heuristic	-259.47	277.39
Random	-676.39^{*}	277.39
Session	-61.56	60.87
Heuristic: Session	22.13	83.91
Random: Session	171.90^{*}	83.91

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.9: Results for overall effects of SDNN without TEST01 for mixed effects model with Control group as reference

Fixed Effect	Estimate	Std. Error
(Intercept)	1026.17^{***}	263.56
Heuristic	-302.02	363.29
Random	-831.14^{*}	363.29
Session	-59.79	79.52
Heuristic:Session	20.88	109.61
Random:Session	210.44	109.61

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.10: Results for overall effects of RMSSD without TEST01 for mixed effects model with Control group as reference

faster during each session for the Control group compared to the Heuristic group. This suggests that the Heuristic delivery of the intervention has a calming effect with respect to heart rate if we assume the first session was a training session. Both of these effects, however, are less than the MDES from the sensitivity analysis, which for the Heuristic was computed to be -0.15 and for the Control was 0.10.

4.4 Reductions in Anxiety

We observed no significant differences in any of the PRE or POST surveys using an ANOVA test. Since there was no difference in the VRP survey taken at the end, it is likely that we can say any resultant effects from the stimulus groups weren't confounded by any differences in overall

Fixed Effect	Estimate	Std. Error
(Intercept)	0.0935^{**}	0.0323
Heuristic	-0.0921^{*}	0.0445
Random	-0.0820	0.0445
Session	-0.0204^{*}	0.0102
Heuristic:Session	0.0242	0.0141
Random:Session	0.0274	0.0141

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.11: Results for overall effects of HR Slope without TEST01 for mixed effects model with Control group as reference

VR immersion despite receiving stimulus in the real world, and the stimulus did not affect the reality of the VR environment compared to the Control.

We also used a mixed effects model on the AEBS scores across sessions to observe any effects over time for each group. Fixed effects were session index and intervention group as well as effects between them. While we observe a change across the session index for the AEBS scores, none of these show a significant difference between intervention groups. That is, the AEBS scores between groups change approximately equally over time. Average AEBS scores also show no significant differences between groups. Regardless, the AEBS series across time is shown in Figure 4.4.

One potential explanation behind the lack of differences in AEBS scores is that the vibrotactile biofeedback didn't have an effect on the participant's anxious cognition from public speaking (i.e. the participant didn't think believe they felt less anxious). Given the results from the micro-analysis, this suggests that vibrotactile biofeedback has a physiological impact on the individual rather than a cognitive one.

4.4.1 Weak Estimation of AEBS Differences

While we did not collect initial state anxiety scores for any anxiety-related surveys in the PRE phase to calculate before-after differences in state anxiety scores, we can make a weak estimation of any significant differences by taking the difference between the AEBS scores for TEST04 and subtracting them from TEST01, thereby 'base-lining' them to that particular participant. We found

Figure 4.4: AEBS scores across each TEST session

no significant differences from an ANOVA, but the boxplot of these differences can be seen in Figure 4.5. Similarly, the boxplots for TEST01 and TEST04 for each intervention group are shown in Figure 4.6. The results of the ANOVA suggest that receiving the vibrotactile biofeedback did not lead to a significant difference in self-reported anxiety scores for a pre/post comparison.

4.5 Subjective Ratings

At the conclusion of the experiment in the POST phase, the last surveys the participants answered were asking about their comments and subjective experiences about the vibrotactile sensations. We report the subjective comments from the participants in the Random and Heuristic groups as well as graphs of their ratings about questions regarding the vibrotactile sensations. Many reported that the introduction of the stimulus interrupted their train of thought and distracted them. Some participants also claimed to have notice the sensations at the beginning of each session, but

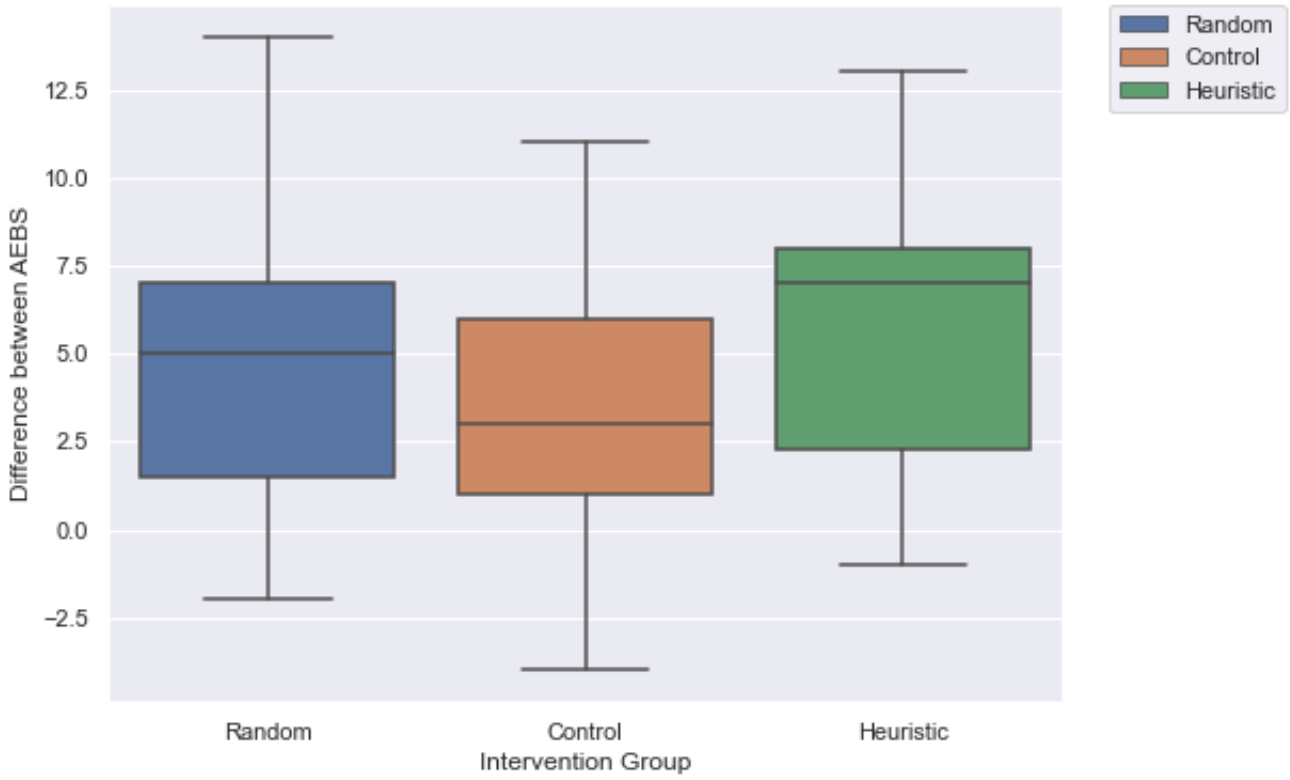


Figure 4.5: Differences in AEBS scores from TEST04 and TEST01

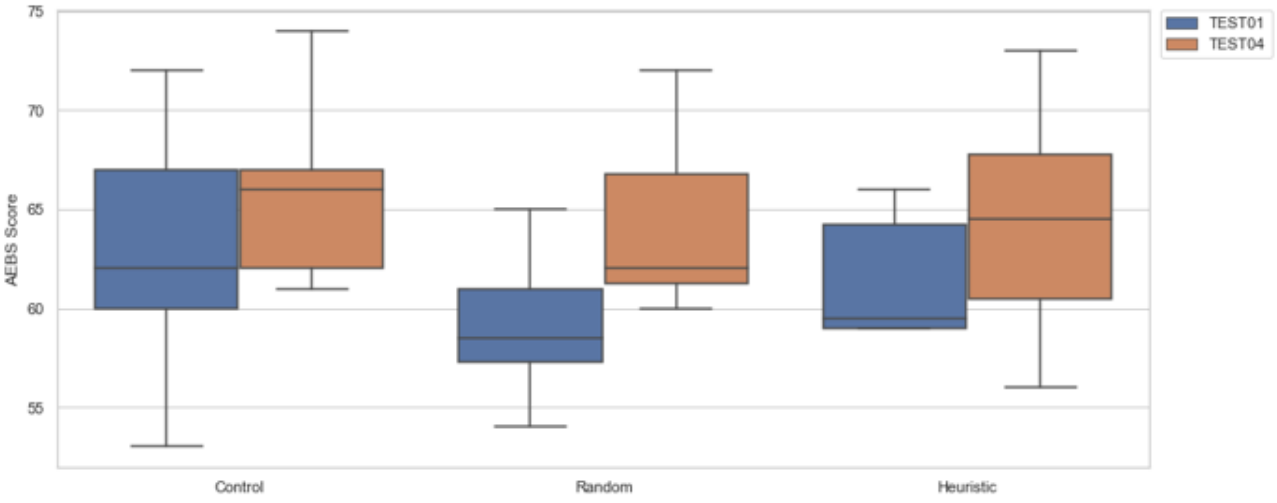


Figure 4.6: AEBS scores for TEST01 and TEST04 for each intervention group

near the end of the session noticed them less as they became more engaged in the speech. We list a selection of relevant responses here for discussion purposes. Additionally, the separate bar graphs for the Random and Heuristic groups are contained in 4.7.

- P011 (Random):*"[The sensations] threw off my train of thought but I was sometimes able to tune [them] out when I was really getting into a speech."*
- P004 (Random): *"I noticed [the sensations] more in the beginning [of the session], but did not feel it at all by the end of all my speeches."*
- P025 (Heuristic): *"[The sensations] distracted my thought process but not my speaking ability or my overall speech. They more or less brought me back to reality as I was paying more attention to [giving] the speech versus analyzing the world around me to the fullest extent."*
- P023 (Heuristic): *"...it interrupted my train of thought very well. It made it very hard to focus and I had difficulty learning to tune it out."*

4.6 Personalization

4.6.1 Baseline Differences

As expected, some of the baseline measurements captured in the RELAX component of the PRE phase exhibit high variance as shown in Table 4.12. Examples of the distributions are captured in Figure 4.8.

4.6.2 Data Description

For this analysis, we expand the EDA and BVP feature extraction windows to 30 seconds, maintaining a window of 10 seconds for speech. Following that, we examine two window sizes on the target prediction for the model, depending on the prediction task. These tasks are:

1. *Task 1*: Predicting high state anxiety levels within a 5-second window. The self-reported state anxiety annotation scores are binarized based on the mean annotation level within this

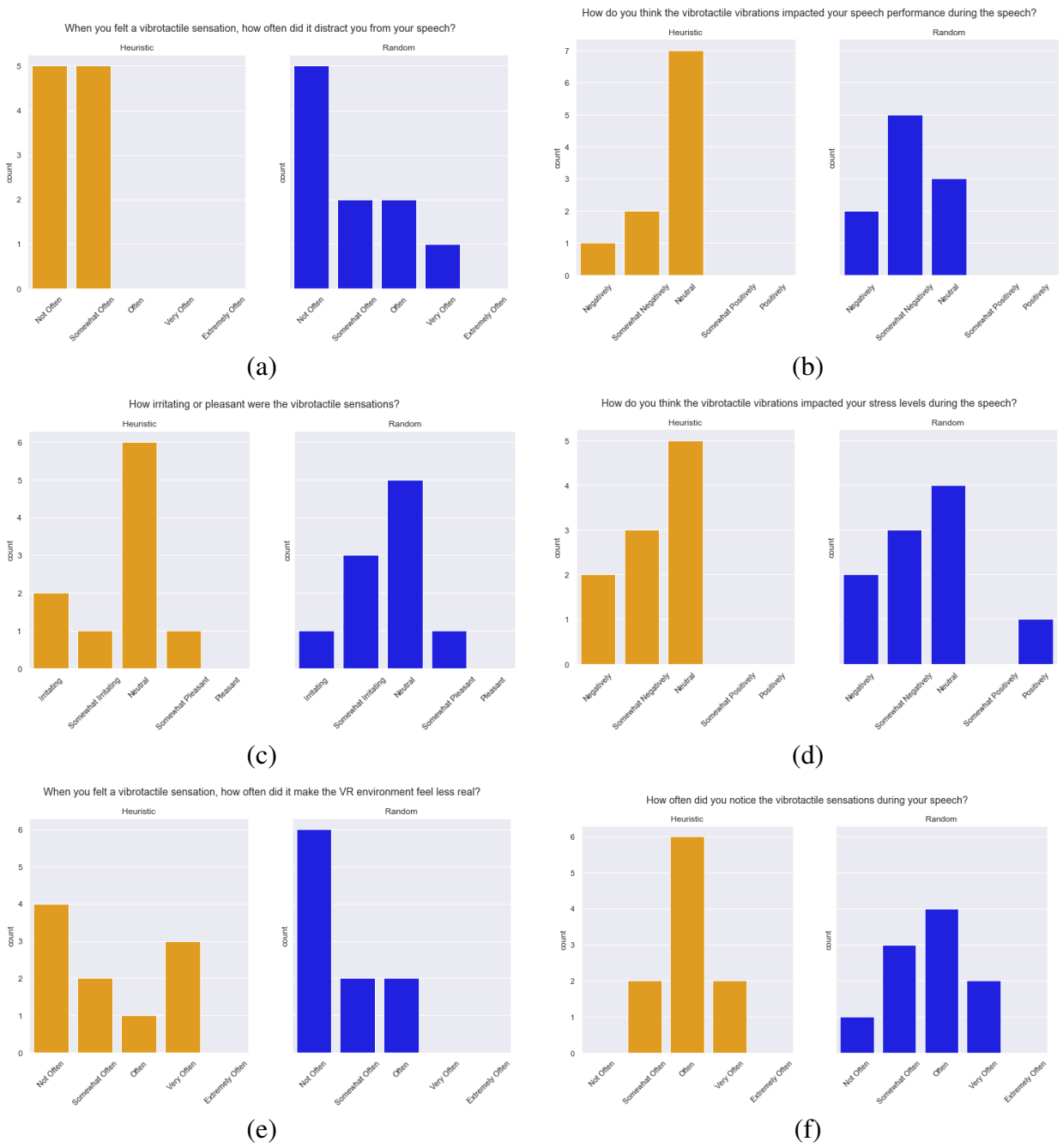
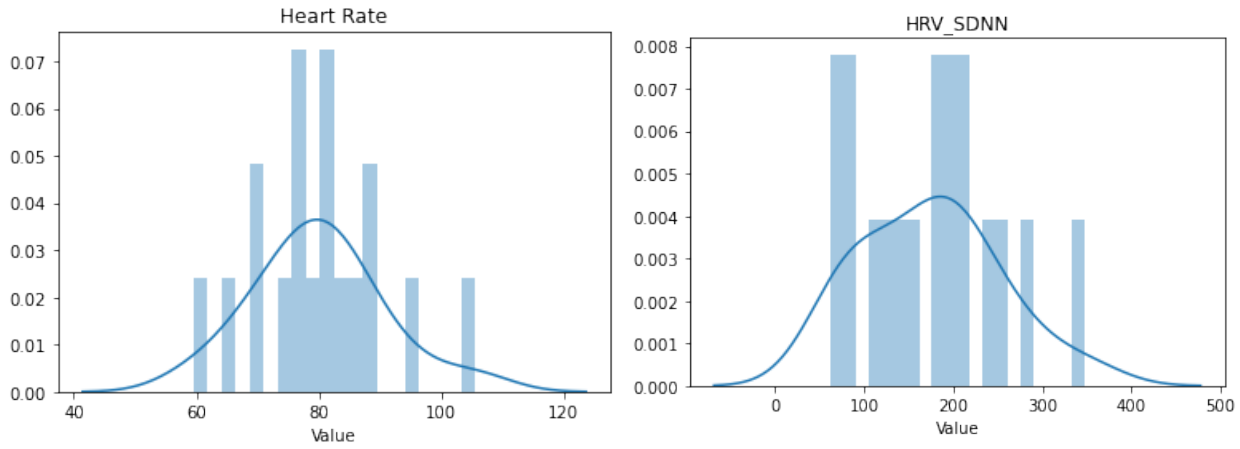


Figure 4.7: Subjective survey responses about vibrotactile stimulus from the Heuristic (Orange) and Random (Blue) groups

window, with a threshold of 3.0 (the mean must be *greater than* 3.0 to indicate a positive sample).

Feature	μ	σ	σ^2
SCR Frequency	0.328	0.105	0.011
Mean EDA	0.642	1.216	1.479
SDNN	174.1	75.58	5712.1
RMSSD	233.3	103.23	10656.1
HR	79.95	10.45	109.16

Table 4.12: Mean, standard deviation, and variance of each baseline feature captured in the RELAX component



(a) (b)

Figure 4.8: Distributions of (a) Heart Rate and (b) SDNN for measurements in the RELAX component

2. *Task 2*: Predict increasing annotation levels within a 10-second window. We calculate the slope over the 10 second window for annotations. If that slope is strictly positive, the label is positive, otherwise the label is negative.

The samples are collected with a maximum window length of 30 seconds and at a stride of 15 seconds to prevent significant overlap.

4.6.3 Model Performance

We investigate all possible combinations of personalization operators for the data collected from the current participants. We used a maximum tree depth of 6 for each boosted tree. For pre-

dicting high levels of stress from the annotations, we observed better performance when leaving out speech and just including the BVP- and EDA-related features. The metrics we gathered were accuracy, f-score, precision, and recall. The results are captured in four tables, each table containing the results on fine-tuning. Table 4.13 has the results with only fine-tuning operators, Table 4.14 with fine-tuning and MOD operators, Table 4.15 with fine-tuning and TRAIT operators, and Table 4.16 with fine-tuning, TRAIT, and MOD operators.

Note that for decision trees, the magnitude of the variables isn't typically a factor. So, the MOD operator has no effect if we are fine-tuning an empty model on the current participant's TEST01 and/or TEST02 data. If we're including the BASE data, then the MOD operator will have an impact. The operators for each table are the results of applying MOD, TRAIT, or both to the fine-tuning options. BASE trains the model on the 'v1' dataset, TEST01 further trains the model on the participant's TEST01 data, and TEST* trains the model on the participant's TEST01 and TEST02 data.

For predicting high moments of stress (Task 1), our best model appears to be when we pretrain the model on the 'v1' dataset and fine-tune on the participant's TEST01 and TEST02 data, taking care to baseline-norm both sets of data from each participant's relaxation sessions (MOD(BASE + TEST*) + TRAIT). We also found our best model resulted from omitting speech features, possibly due to the dimensionality of the input growing too high with speech. This could be because of the personalization efforts altering the trees to handle the individual more accurately. Table 4.17 contains the top 3 features with respect to feature importance metrics, which are Weight (number of times that feature is used to split the data across all trees), Gain (average reduction in entropy for all splits for that feature), and Cover (number of observations used in all splits for that feature) [109].

Weight	Gain	Cover
Mean SCL	STAI-Trait	CAI-Trait
STAI-Trait	CAI-Trait	STAI-Trait
SCR Amplitude	Mean SCL	SCR Frequency

Table 4.17: Top 3 features for weight, gain, and cover for Task 1

For predicting increasing levels of stress (Task 2), our best model appears to be when we only pretrain the model on data which has been baseline-normed from the 'v1' dataset (MOD(BASE)). Speech features also helped improve performance for Task 2. This could be due to modeling increases and decreases being much easier than high or low stress moments because increases or decreases are more generalizable. Table 4.18 contains the top 3 features similar to those described for Task 1.

Weight	Gain	Cover
HR Slope	RMSSD	RMS Energy
Shimmer	ZCR	Shimmer
Mean SCR	F0	ZCR

Table 4.18: Top 3 features for weight, gain, and cover for Task 2

Operator	Task 1: Predict moments of high stress				Task 2: Predict increasing stress			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
BASE	0.491	**0.420**	0.383	**0.608**	0.603	**0.229**	**0.197**	**0.603**
BASE + TEST01	0.470	0.337	0.377	0.531	0.662	0.196	0.180	0.279
BASE + TEST*	**0.617**	0.398	**0.435**	0.368	0.730	0.098	0.094	0.084
TEST01	0.423	0.346	0.362	0.543	0.693	0.185	0.166	0.205
TEST*	0.586	0.376	0.392	0.289	**0.741**	0.118	0.137	0.089

Table 4.13: Results of gradient boosting trees without using any operators aside from the fine-tuning on individual-specific data

Operator	Task 1: Predict moments of high stress				Task 2: Predict increasing stress			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
BASE	0.479	**0.424**	0.363	0.459	0.319	**0.355**	**0.244**	**0.821**
BASE + TEST01	0.475	0.359	0.389	**0.534**	0.703	0.116	0.118	0.139
BASE + TEST*	**0.592**	0.371	**0.451**	0.296	**0.730**	0.098	0.094	0.084

Table 4.14: Results of gradient boosting trees with baseline norming each participant's data based on their relaxation data (MOD)

Operator	Task 1: Predict moments of high stress				Task 2: Predict increasing stress			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
BASE	0.439	0.310	0.357	0.402	0.410	**0.212**	0.148	**0.655**
BASE + TEST01	0.430	0.348	0.331	0.532	0.692	0.151	0.151	0.166
BASE + TEST*	**0.611**	**0.429**	**0.456**	0.380	0.721	0.077	0.074	0.071
TEST01	0.423	0.346	0.362	**0.543**	0.693	0.185	**0.166**	0.205
TEST*	0.586	0.376	0.392	0.289	**0.741**	0.118	0.137	0.089

Table 4.15: Results of gradient boosting trees with appending the Trait-anxiety scores for a given individual to each input sample for that individual (TRAIT)

Operator	Task 1: Predict moments of high stress				Task 2: Predict increasing stress			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
BASE	0.450	0.359	0.335	0.418	0.320	**0.226**	**0.148**	**0.744**
BASE + TEST01	0.433	0.310	0.325	**0.491**	0.693	0.149	0.138	0.185
BASE + TEST*	**0.623**	**0.452**	**0.467**	0.402	**0.739**	0.102	0.115	0.092

Table 4.16: Results of gradient boosting trees with appending the Trait-anxiety scores for a given individual to each input sample for that individual (TRAIT), combined with baseline norming each participant's data based on their relaxation data (MOD)

5. DISCUSSION

5.1 Vibrotactile Biofeedback

Results from this thesis indicate that vibrotactile biofeedback delivered for short durations may have a calming effect for the individual. We saw that the timing of the vibrotactile biofeedback had a difference in the proximal effects of the intervention in the short-term – individuals in the Heuristic group had lower heart rate after each intervention in earlier sessions compared to the Random group. However, the Random group experienced an improvement in vocal stability (shimmer) effects in later sessions. When we compare the Heuristic with the Control group, we saw that delivering the interventions (as opposed to withholding them in the Control) led to a reduction in SCR peaks after the intervention ended, suggesting that the disabling of the intervention is what leads to a proximal calming effect (lower SCR peaks). The Heuristic group also observed higher HRV for earlier sessions, but this HRV did not improve across sessions like it did with the Control group. We further observed heart rate *during* the interventions between the Control and Heuristic group and found that while the vibrotactile biofeedback is being administered, participants experienced a lower heart rate across sessions *and* across interventions within each session. So, the vibrotactile biofeedback appears to lower SCR peaks and HR, which may help one's ability to manage stress, but also decreases HRV, which may reduce one's ability to effectively manage stress.

Regarding the distal effects of the vibrotactile biofeedback, the repeated correlations revealed an increase of HRV-related time metrics for the number of interventions received over the sessions, suggesting that an individual may just need to become used to the initial triggering of the interventions to start receiving any benefits. We also saw a negative correlation in the energy within the low frequency band for HRV, reinforcing this idea. When we used the mixed effects model with the feature values computed across the entire session, we saw no significant differences in any of the feature values unless we assume the first TEST sessions is a *training* session, in which case we observe negative effects regarding HRV for the Random group compared to the Control

which indicate the Random group experienced increased levels of overall state anxiety. However, we saw a slower change in heart rate over time for the Heuristic group compared to the Control group, which indicates a positive calming effect for the Heuristic group with respect to overall state anxiety measures.

While we did not observe any significant differences in any of the POST state anxiety self-reports or any of the AEBS surveys after each TEST session, the interventions delivered using the Heuristic algorithm still appear to benefit the participant over the Random delivery and over no delivery, but the distal effects are minimal. Since we failed to observe significant distal effects of the vibrotactile biofeedback, this suggests that distal effects are in some way caused by vibrotactile adaptation, as it is clear the stimulus wasn't as effective in the short term compared to other studies that delivered the stimulus continuously.

5.2 Personalization

Attempting to detect the onset of high stressful moments and increasing moments of stress still appears to present some challenges. While the task of stress prediction is challenging, the performance of models is moderate, depicting F1-score between 0.3-0.45. This suggests that there needs to be additional work for deploying such continuous prediction system in real-life.

Generally it appears that fine-tuning leads to improvements in accuracy. This may be due to informing the model of recent trends, but it also runs the risk of overfitting. The MOD and TRAIT operators appear to not have much effect on their own, but when combined for Task 1 give us our best results. The performance on Task 2 needs serious improvement. While accuracy is high, the F1 measure is abysmally low.

5.3 Limitations

One major limitation of this study was lack of baseline State-scale scores of each participant before beginning any of the TEST sessions. The lack of this baseline score makes it difficult to compare the State-scale scores of anxiety and communication anxiety after the TEST sessions to ideally observe a difference between the three intervention groups.

Furthermore, the sensation delivered by the Pulse intuitively feels *too* epidermal. Other devices should be considered that deliver a vibration that feels *in* the wrist.

The delivery time of the vibrotactile biofeedback for the Heuristic group also didn't match with the Random group, which could have confounded any results observed between the groups. Future work could instead use a yoked control design between one participant in the Random group and one participant in the Heuristic group to ensure the delivery of the intervention for the Heuristic group is matched by the Random group exactly.

With respect to algorithm design, it would have made more sense to design a heuristic algorithm based on heart rate, since heart rate changes have been the primary physiology effect from vibrotactile stimulus. This was eliminated from the initial design choice due to difficulties in comparing heart rates between participants. However, with proper baseline procedures, it might be the case that using heart rate as a triggering mechanism would provide better results. Alternatively, future work could use mean SCR amplitude, which appears to have more support of indication of stress [107]. This was an initial candidate for use in the rule-based Heuristic algorithm, but finding a comparable SCR amplitude amongst the population in the v1 dataset proved to be difficult, and indeed according to [107], group means for this metric (and other EDA-related measures) can vary based on experimental and environmental conditions. In the resulting data from this study, mean SCR amplitude computed over the session was the only EDA-related feature that had a positive correlation ($.067$) with the AEBS score at the end of each session. Other EDA-related features had negative correlations, including SCR frequency ($-.136$) and mean SCL level ($-.125$).

Finally, while the goal of this experiment was to access the *subconscious* effect of the vibrotactile biofeedback on the stress of the participant, the interventions may be more effective if the user is conditioned on them in a true biofeedback loop. For example, the participant could have engaged in a practice session where they practice trying to lower their heart rate in accordance with receiving the vibrotactile stimulus. Alternatively, beforehand we could condition the participant to instead slow their speaking rate down if the intervention is enabled. This strategy involves informing the user that the vibrotactile stimulus is delivered during perceived physiological arousal,

therefore they become aware of the reason behind the initial triggering of the intervention though they may not necessarily understand its purpose.

6. CONCLUSION

This thesis investigated the effectiveness of vibrotactile feedback by deploying two algorithms during a public speaking VR session. We compared the effects of delivering the vibrotactile stimulus according to these algorithms with a Control group. We found that the vibrotactile biofeedback delivered with the Heuristic algorithm reflected better state anxiety in earlier sessions, but habituated and converged with the Control group in later sessions, but still revealed some overall positive distal effects for controlling symptoms of state anxiety. Furthermore, the randomly delivered vibrotactile biofeedback resulted in lower overall HRV compared to a Control, indicating that receiving the vibrotactile biofeedback at random moments leads to more state anxiety for the individual. We can therefore conclude that the timing of the intervention matters to an extent.

We also assessed the effectiveness of personalized machine learning models for detecting onsets of high moments of stress and increasing moments of stress based on affect labeling. We say that baseline norming, fine-tuning, and providing the trait scores for certain anxiety self-report surveys can all be used to effectively augment models that detect high moments of stress.

6.1 Recommendations for Future Work

The results from this thesis will hopefully guide those researching vibrotactile biofeedback as an anxiety intervention to further investigate the habituation effect. Since we could not find overwhelming evidence of in-the-moment vibrotactile biofeedback being effective in reducing stress, this is highly suggestive that vibrotactile adaptation is causing the calming effect. Future work could include intentionally delivering varying durations of vibrotactile stimulus to individuals experiencing stress and search for the point at which a calming effect occurs, and then connect these results with vibrotactile adaptation.

Furthermore, other vibrotactile stimulus devices should be explored, as the Pulse may not be an effective candidate at delivering the vibrotactile sensations. Based on the market survey that was conducted before the study, we found that there are no affordable alternatives to delivering

the vibrotactile stimulus at a controllable rate. However, we expect that this functionality may be included in the future, as smart watch manufacturers continue to develop their products.

While the calming effect was not observed in the wrist, other avenues of research could investigate alternative locations for delivering the stimulus, like the upper arm. Similarly, the idea of compression [113] may be in some way related to vibrotactile biofeedback, delivering a possible relieving effect discretely like the vibrotactile stimulus.

Finally, research should continue investigating true real-time in-the-moment detection of anxious moments within stressful 'sessions', instead of classifying entire sessions as stressed or not stressed. The challenges that come with designing more granular detection for stressful moments will guide future research in attempting to find the optimal moments through EMA to deliver interventions.

REFERENCES

[1] "Soundbrenner pulse." `https://www.soundbrenner.com/pulse`.

[2] J. M. Girard, "Carma: Software for continuous affect rating and media annotation," *Journal of open research software*, vol. 2, no. 1, 2014.

[3] L. F. Parvis, "The importance of communication and public-speaking skills," *Journal of Environmental Health*, vol. 63, no. 9, pp. 44–44, 2001.

[4] J. R. Johnson and N. Szczupakiewicz, "The public speaking course: Is it preparing students with work related public speaking skills?," *Communication Education*, vol. 36, no. 2, pp. 131–137, 1987.

[5] G. R. VandenBos, *APA dictionary of psychology.* American Psychological Association, 2007.

[6] S. Abuse and M. H. S. Administration, "Dsm-5 changes: Implications for child serious emotional disturbance [internet]," 2016.

[7] R. C. Kessler, M. B. Stein, and P. Berglund, "Social phobia subtypes in the national comorbidity survey," *American Journal of Psychiatry*, vol. 155, no. 5, pp. 613–619, 1998.

[8] V. Tejwani, D. Ha, and C. Isada, "Public speaking anxiety in graduate medical education—a matter of interpersonal and communication skills?," *Journal of graduate medical education*, vol. 8, no. 1, pp. 111–111, 2016.

[9] F. Raja, "Anxiety level in students of public speaking: Causes and remedies.," *Journal of education and educational development*, vol. 4, no. 1, pp. 94–110, 2017.

[10] J. A. Daly, A. L. Vangelisti, and S. G. Lawrence, "Self-focused attention and public speaking anxiety," *Personality and Individual Differences*, vol. 10, no. 8, pp. 903–913, 1989.

[11] E. Kimani, T. Bickmore, H. Trinh, and P. Pedrelli, "You'll be great: Virtual agent-based cognitive restructuring to reduce public speaking anxiety," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 641–647, IEEE, 2019.

[12] D. A. Clark, "Cognitive restructuring," *The Wiley handbook of cognitive behavioral therapy*, pp. 1–22, 2013.

[13] R. T. Azevedo, N. Bennett, A. Bilicki, J. Hooper, F. Markopoulou, and M. Tsakiris, "The calming effect of a new wearable device during the anticipation of public speech," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.

[14] K. Y. Choi and H. Ishii, "ambienbeat: Wrist-worn mobile tactile biofeedback for heart rate rhythmic regulation," in *Proceedings of the fourteenth international conference on tangible, embedded, and embodied interaction*, pp. 17–30, 2020.

[15] J. T. Grossman, M. R. Frumkin, T. L. Rodebaugh, and E. J. Lenze, "mhealth assessment and intervention of depression and anxiety in older adults," *Harvard review of psychiatry*, vol. 28, no. 3, p. 203, 2020.

[16] M. S. Marcolino, J. A. Q. Oliveira, M. D'Agostino, A. L. Ribeiro, M. B. M. Alkmim, and D. Novillo-Ortiz, "The impact of mhealth interventions: systematic review of systematic reviews," *JMIR mHealth and uHealth*, vol. 6, no. 1, p. e23, 2018.

[17] A. B. Labrique, L. Vasudevan, E. Kochi, R. Fabricant, and G. Mehl, "mhealth innovations as health system strengthening tools: 12 common applications and a visual framework," *Global health: science and practice*, vol. 1, no. 2, pp. 160–171, 2013.

[18] S. Bae, T. Chung, D. Ferreira, A. K. Dey, and B. Suffoletto, "Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions," *Addictive behaviors*, vol. 83, pp. 42–47, 2018.

[19] T. Rahman, M. Czerwinski, R. Gilad-Bachrach, and P. Johns, "Predicting" about-to-eat" moments for just-in-time eating intervention," in *Proceedings of the 6th International Conference on Digital Health Conference*, pp. 141–150, 2016.

[20] W. Choi, S. Park, D. Kim, Y.-k. Lim, and U. Lee, "Multi-stage receptivity model for mobile just-in-time health intervention," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–26, 2019.

[21] S. Akter and P. Ray, "mhealth-an ultimate platform to serve the unserved," *Yearbook of medical informatics*, vol. 19, no. 01, pp. 94–100, 2010.

[22] S. P. Whiteside, "Mobile device-based applications for childhood anxiety disorders," *Journal of Child and Adolescent Psychopharmacology*, vol. 26, no. 3, pp. 246–251, 2016.

[23] W. Nilsen, S. Kumar, A. Shar, C. Varoquiers, T. Wiley, W. T. Riley, M. Pavel, and A. A. Atienza, "Advancing the science of mhealth," *Journal of health communication*, vol. 17, no. sup1, pp. 5–10, 2012.

[24] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," *Communication education*, vol. 59, no. 1, pp. 70–105, 2010.

[25] . D. J. A. Beatty, M. J., "Physiological assessment," in *Avoiding communication: Shyness, reticence, and communication apprehension* (J. A. T. H. . D. M. A. J. A. Daly, J. C. McCroskey, ed.), pp. 217–229, Cresskill, NJ: Hampton Press, 1997.

[26] C. Maaoui and A. Pruski, "Emotion recognition through physiological signals for human-machine communication," *Cutting Edge Robotics*, vol. 2010, no. 317-332, p. 11, 2010.

[27] Z. Zhang, H. Su, Q. Peng, Q. Yang, and X. Cheng, "Exam anxiety induces significant blood pressure and heart rate increase in college students," *Clinical and experimental hypertension*, vol. 33, no. 5, pp. 281–286, 2011.

[28] J. A. Chalmers, D. S. Quintana, M. J. Abbott, A. H. Kemp, *et al.*, "Anxiety disorders are associated with reduced heart rate variability: a meta-analysis," *Frontiers in psychiatry*, vol. 5, p. 80, 2014.

[29] M. Birket-Smith, N. Hasle, and H. Jensen, "Electrodermal activity in anxiety disorders," *Acta Psychiatrica Scandinavica*, vol. 88, no. 5, pp. 350–355, 1993.

[30] M. Yadav, M. N. Sakib, K. Feng, T. Chaspari, and A. Behzadan, "Virtual reality interfaces and population-specific models to mitigate public speaking anxiety," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, IEEE, 2019.

[31] M. E. Owens and D. C. Beidel, "Can virtual reality effectively elicit distress associated with social anxiety disorder?," *Journal of Psychopathology and Behavioral Assessment*, vol. 37, no. 2, pp. 296–305, 2015.

[32] H. S. Wallach, M. P. Safir, and M. Bar-Zvi, "Virtual reality cognitive behavior therapy for public speaking anxiety: a randomized clinical trial," *Behavior modification*, vol. 33, no. 3, pp. 314–338, 2009.

[33] E. Kimani, "A sensor-based framework for real-time detection and alleviation of public speaking anxiety," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 55–59, IEEE, 2019.

[34] E. Kimani and T. Bickmore, "Addressing public speaking anxiety in real-time using a virtual public speaking coach and physiological sensors," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 260–263, 2019.

[35] S. R. Harris, R. L. Kemmerling, and M. M. North, "Brief virtual reality therapy for public speaking anxiety," *Cyberpsychology & behavior*, vol. 5, no. 6, pp. 543–550, 2002.

[36] P. L. Anderson, E. Zimand, L. F. Hodges, and B. O. Rothbaum, "Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure," *Depression and anxiety*, vol. 22, no. 3, pp. 156–158, 2005.

[37] S.-C. Yeh, Y.-Y. Li, C. Zhou, P.-H. Chiu, and J.-W. Chen, "Effects of virtual reality and augmented reality on induced anxiety," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1345–1352, 2018.

[38] M. S. Patel, D. A. Asch, and K. G. Volpp, "Wearable devices as facilitators, not drivers, of health behavior change," *Jama*, vol. 313, no. 5, pp. 459–460, 2015.

[39] J. Gepperth, "Smart things: Wearables & clothing," *Smart Things*, vol. 3, no. 2012, pp. 41–48, 2012.

[40] K. Bodine and F. Gemperle, "Effects of functionality on perceived comfort of wearables," in *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pp. 57–57, Citeseer, 2003.

[41] R. W. Picard and J. Healey, "Affective wearables," *Personal technologies*, vol. 1, no. 4, pp. 231–240, 1997.

[42] E. Dagan, E. Márquez Segura, F. Altarriba Bertran, M. Flores, R. Mitchell, and K. Isbister, "Design framework for social wearables," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, pp. 1001–1015, 2019.

[43] N. K. Dim and X. Ren, "Investigation of suitable body parts for wearable vibration feedback in walking navigation," *International Journal of Human-Computer Studies*, vol. 97, pp. 34–44, 2017.

[44] H. P. Profita, J. Clawson, S. Gilliland, C. Zeagler, T. Starner, J. Budd, and E. Y.-L. Do, "Don't mind me touching my wrist: a case study of interacting with on-body technology in public," in *Proceedings of the 2013 International Symposium on Wearable Computers*, pp. 89–96, 2013.

[45] K. E. Heron and J. M. Smyth, "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments," *British journal of health psychology*, vol. 15, no. 1, pp. 1–39, 2010.

[46] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[47] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support," *Annals of Behavioral Medicine*, vol. 52, no. 6, pp. 446–462, 2018.

[48] S. P. Goldstein, B. C. Evans, D. Flack, A. Juarascio, S. Manasse, F. Zhang, and E. M. Forman, "Return of the jitai: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors," *International journal of behavioral medicine*, vol. 24, no. 5, pp. 673–682, 2017.

[49] J. Costa, A. T. Adams, M. F. Jung, F. Guimbretière, and T. Choudhury, "Emotioncheck: leveraging bodily signals and false feedback to regulate our emotions," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 758–769, 2016.

[50] V. Hollis, A. Pekurovsky, E. Wu, and S. Whittaker, "On being told how we feel: how algorithmic sensor feedback influences emotion perception," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–31, 2018.

[51] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 261–268, 2011.

[52] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behavioural brain research*, vol. 341, pp. 50–53, 2018.

[53] M. S. Goodwin, C. A. Mazefsky, S. Ioannidis, D. Erdogmus, and M. Siegel, "Predicting aggression to others in youth with autism using a wearable biosensor," *Autism research*, vol. 12, no. 8, pp. 1286–1296, 2019.

[54] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[55] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *2007 29th annual international conference of the ieee engineering in medicine and biology society*, pp. 4656–4659, IEEE, 2007.

[56] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker classification I*, pp. 108–137, Springer, 2007.

[57] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[58] P. Paredes and M. Chan, "Calmmenow: exploratory research and design of stress mitigating mobile interventions," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 1699–1704, 2011.

[59] R. Cuijpers, E. K. van Dijk, S. Longhi, E. Marchegiani, and A. Monteriu, "Psychophysiological stress control via heart rate entrainment," in *2019 Zooming Innovation in Consumer Technologies Conference (ZINC)*, pp. 9–10, IEEE, 2019.

[60] M. Thaut, *Rhythm, music, and the brain: Scientific foundations and clinical applications*. Routledge, 2013.

[61] V. S. Anishchenko, A. G. Balanov, N. B. Janson, N. B. Igosheva, and G. V. Bordyugov, "Entrainment between heart rate and weak noninvasive forcing," *International Journal of Bifurcation and Chaos*, vol. 10, no. 10, pp. 2339–2348, 2000.

[62] H. Mütze, R. Kopiez, and A. Wolf, "The effect of a rhythmic pulse on the heart rate: Little evidence for rhythmical 'entrainment'and 'synchronization'," *Musicae Scientiae*, vol. 24, no. 3, pp. 377–400, 2020.

[63] U. Berglund and B. Berglund, "Adaptation and recovery in vibrotactile perception," *Perceptual and motor skills*, vol. 30, no. 3, pp. 843–853, 1970.

[64] M. S. Schwartz and F. Andrasik, *Biofeedback: A practitioner's guide*. Guilford Publications, 2017.

[65] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[66] J. Minguillon, E. Perez, M. A. Lopez-Gordo, F. Pelayo, and M. J. Sanchez-Carrion, "Portable system for real-time detection of stress level," *Sensors*, vol. 18, no. 8, p. 2504, 2018.

[67] V. Khullar, R. G. Tiwari, A. K. Agarwal, and S. Dutta, "Physiological signals based anxiety detection using ensemble machine learning," in *Cyber Intelligence and Information Retrieval*, pp. 597–608, Springer, 2022.

[68] R. Martinez, E. Irigoyen, A. Arruti, J. I. Martín, and J. Muguerza, "A real-time stress classification system based on arousal analysis of the nervous system by an f-state machine," *Computer methods and programs in biomedicine*, vol. 148, pp. 81–90, 2017.

[69] K. Kyriakou, B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. H. Wilhelm, T. Osborne, and J. Pykett, "Detecting moments of stress from measurements of wearable physiological sensors," *Sensors*, vol. 19, no. 17, p. 3805, 2019.

[70] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 351–360, 2012.

[71] J. B. Torre and M. D. Lieberman, "Putting feelings into words: Affect labeling as implicit emotion regulation," *Emotion Review*, vol. 10, no. 2, pp. 116–124, 2018.

[72] M. Soury and L. Devillers, "Stress detection from audio on multiple window analysis size in a public speaking task," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 529–533, IEEE, 2013.

[73] W. Wen, G. Liu, Z.-H. Mao, W. Huang, X. Zhang, H. Hu, J. Yang, and W. Jia, "Toward constructing a real-time social anxiety evaluation system: Exploring effective heart rate features," *IEEE transactions on affective computing*, vol. 11, no. 1, pp. 100–110, 2018.

[74] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, pp. 1185–1193, 2016.

[75] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Transactions on Affective Computing*, 2020.

[76] "Empatica e4 wristband." `https://www.empatica.com/research/e4/`.

[77] "Oculus rift." `https://www.oculus.com/rift/`.

[78] "Fifine lapel microphone." `https://fifinemicrophone.com/collections/lavalier-mic-system/products/lavalier-microphone-for-pc-recording-k053`.

[79] "Virtual orator." `https://virtualorator.com/`.

[80] "E4 streaming server." `https://developer.empatica.com/windows-streaming-server.html`.

[81] "e4stream." `https://pypi.org/project/e4stream/`.

[82] "Audacity." `https://www.audacityteam.org/`.

[83] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, pp. 1–8, 2021.

[84] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[85] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, pp. 2001–2005, 2016.

[86] O. P. John, S. Srivastava, *et al.*, *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*, vol. 2. University of California Berkeley, 1999.

[87] M. R. Leary, "A brief version of the fear of negative evaluation scale," *Personality and social psychology bulletin*, vol. 9, no. 3, pp. 371–375, 1983.

[88] S. Booth-Butterfield and M. Gould, "The communication anxiety inventory: Validation of state-and context-communication apprehension," *Communication Quarterly*, vol. 34, no. 2, pp. 194–205, 1986.

[89] J. C. McCroskey, "Measures of communication-bound anxiety," 1970.

[90] M. Pörhölä, "Trait anxiety, experience, and the public speaking state responses of finnish university students," *Communication research reports*, vol. 14, no. 3, pp. 367–384, 1997.

[91] C. D. Spielberger, "State-trait anxiety inventory for adults," 1983.

[92] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence*, vol. 7, no. 3, pp. 225–240, 1998.

[93] D. L. Chambless, G. C. Caputo, P. Bright, and R. Gallagher, "Assessment of fear of fear in agoraphobics: the body sensations questionnaire and the agoraphobic cognitions questionnaire.," *Journal of consulting and clinical psychology*, vol. 52, no. 6, p. 1090, 1984.

[94] "Memtrax memory test." `https://memtrax.com/test/`.

[95] "Beautiful jellyfish aquarium for relaxation in 4k - sleep relax meditation music 2 hours screensaver." `https://www.youtube.com/watch?v=95Tc8qIJRuI&ab_channel=Balu-RelaxingNaturein4K`.

[96] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[97] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[98] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan, "Sparse representation of electrodermal activity with knowledge-driven dictionaries," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 960–971, 2014.

[99] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[100] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term hrv features as surrogates of short term hrv: a case study on mental stress detection in real life," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[101] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, 2019.

[102] L. G. Jaimes and R. Steele, "Mobile stress interventions: mechanisms and implications," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 4, no. 13, 2018.

[103] J. M. Smyth and K. E. Heron, "Is providing mobile interventions" just-in-time" helpful? an experimental proof of concept study of just-in-time intervention for stress management," in *2016 IEEE Wireless Health (WH)*, pp. 1–7, IEEE, 2016.

[104] E. Howe, J. Suh, M. B. Morshed, D. McDuff, K. Rowan, J. Hernandez, M. I. Abdin, G. Ramos, T. Tran, and M. Czerwinski, "Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing," in *CHI 2022*, April 2022.

[105] P. E. King, M. J. Young, and R. R. Behnke, "Public speaking performance improvement as a function of information processing in immediate and delayed feedback interventions," *Communication Education*, vol. 49, no. 4, pp. 365–374, 2000.

[106] M. A. Zafar, B. Ahmed, R. Al Rihawi, and R. Gutierrez-Osuna, "Gaming away stress: Using biofeedback games to learn paced breathing," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 519–531, 2018.

[107] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[108] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[109] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[110] J. Z. Bakdash and L. R. Marusich, "Repeated measures correlation," *Frontiers in psychology*, vol. 8, p. 456, 2017.

[111] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[112] P. Leifeld, "texreg: Conversion of statistical model output in R to LaTeX and HTML tables," *Journal of Statistical Software*, vol. 55, no. 8, pp. 1–24, 2013.

[113] E. Foo and B. Holschuh, "Dynamic compression in affective haptics," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pp. 577–583, 2018.