

This file is part of the following work:

Arushi (2025) *A virtual reality approach to improving public speaking skills*. PhD Thesis, James Cook University.

Access to this file is available from:

<https://doi.org/10.25903/fc1w%2Dd353>

Copyright © 2025 Arushi

The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owners of any third party copyright material included in this document. If you believe that this is not the case, please email

researchonline@jcu.edu.au

A Virtual Reality Approach to Improving Public Speaking Skills

*A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy
(Information Technology) in the College of Science and Engineering, James Cook University
Australia.*

14th January 2025

Student name: Arushi

Degree: Doctor of Philosophy (Information Technology)

Student number: 13546276

Advisory panel: A/Prof Roberto Dillon (primary adviser)

A/Prof Ai Ni Teoh (secondary adviser)

A/Prof Denise Dillon (adviser mentor)

Acknowledgments

I would like to extend my heartfelt and humble gratitude to everyone whom I came across in this journey of PhD and until the completion of this thesis. It has been a beautiful learning experience, in terms of academic accomplishment and in ways it has opened my heart to the beauty of human existence and its purpose.

From first being inspired by Professor Liz Tynan for her joyful spirit, my supervisors, Associate Professors Roberto Dillon, Ai Ni Teoh and Denise Dillon whose unwavering support, care, joy, encouragement and enthusiasm have been indispensable throughout this journey. They provided guidance during moments of darkness contributing to both academic growth and personal development. I am deeply moved by their message of joy, continuous learning, growth, resilience, strength, support, togetherness and light, which I will carry in my heart for the rest of my life. I would like to thank Graduate Research School (GRS) Australia, research support Singapore (Ms Belinda and assistants), learning support advisers (Mr Daniel O'Sullivan, Dr Kunchana (Gi) Chouynuu), librarian (Ms Jemina Shanthi), IT department (Mr Vinay, Mr Louie), and Counsellor (Mr Caleb) who provided invaluable support.

Also, I am very thankful to friends, colleagues and teachers, I came across: Xiaofeng, Marc, Alif, Barry, Rushan, Yong Jie, Jonathan, Zewei, James, Himalshi, Sherin, Divya, Kana, Kenan, Kanchan, Rajneesh, Pragya, Ram, Ms Karen, Ms Manisha, Mr Vanan for their invaluable support. I am grateful to JCU Singapore, JCU Australia for welcoming students and individuals with less opportunities and giving them a platform to build themselves.

I would also like to thank my family (mother, father, brother) for sending me to Singapore and my relatives (uncle, aunt) for providing a house, food and supporting me through this journey into unknown with me. I dedicate this work especially to my supervisors and to the universal love that has supported me in countless forms and names, of which I was both a witness and a part.

Statement of Contribution of Others

This thesis was completed by Arushi under the supervision of Associate Professor Roberto Dillon, Associate Professor Ai Ni Teoh and Associate Professor Denise Dillon.

Nature of Assistance	Contribution	Co-contributors
Intellectual Support	Supervision, Conceptual Guidance, Editorial Assistance, Experimental Design and Data Analysis Guidance	A/Prof Roberto Dillon, A/Prof Ai Ni Teoh, A/Prof Denise Dillon and other teachers in university
Financial Support	Tuition Fee Waiver	JCU Research Fee Waiver
Infrastructure Assistance	Use of HDR office space, for providing virtual reality (VR) headset and lab space for conducting experiments	JCU Singapore IT and psychology Department
Ethics Approval	Providing feedback and editorial work in ethics application	Dr Randy Zhou, Dr Simona Azzali
Participant recruitment	Assisted in showcasing participant recruitment research posters in classes	JCU Lecturers/Sessional Staff, Friends (Xiaofeng and many others)
Professional editor	Provided copyediting and proofreading services according to the protocols in the university-endorsed National Guidelines for Editing Research Theses	Dr John Cokley PhD, of EduPreneur Services International

Copyright Declaration

This thesis includes original materials created by me as part of my research and professional activities, as well as certain materials sourced from publicly available websites. These external materials have been used under the principles of fair use for purposes of education, research, and commentary. Proper attribution has been provided wherever applicable in form of references. All rights to the original content remain with me and no infringement on third-party copyrights is intended. If any copyright owner has been inadvertently overlooked or improperly acknowledged, please contact me at arushi@my.jcu.edu.au so that appropriate corrections can be made.

Abstract

Public speaking is a critical skill for effective communication, personal development and advocating societal causes. However, stress is a major barrier to effective public speaking, manifesting in increased heart rate, elevated blood pressure and changes in breathing patterns and tone of voice. While many studies have focused on detecting affective states such as happiness, anger and fear in real-time using algorithmic models, there has been limited research on the real-time detection of stress during public speaking through voice analysis, particularly in virtual reality (VR) settings. The lack of a comprehensive stress-related voice database in the context of public speaking has further limited the identification of voice features and the development of machine learning models for stress detection in this scenario. The present thesis addresses these gaps by exploring the development and integration of a voice-based machine learning model for stress detection in VR environments, specifically in public speaking scenarios. The study was conducted in two phases.

Phase I includes Experiment 1 (E1), which utilises VR to create controlled public speaking environments, featuring Neutral-Positive (NP) and Neutral-Negative (NN) audience behaviours to induce varying levels of stress. E1 successfully generated stress-related voice data, which was used to build the virtual reality public speaking (VRPS) database. This experiment demonstrates that VR can reliably induce measurable stress responses, both physiologically and psychologically, providing a replicable framework for stress induction in performance-based tasks.

Phase II involves the identification of stress indicative voice features, and development of a Stress Detection Model (SDM) using a feedforward neural network (FFNN) architecture, designed to analyse voice features such as Mel-frequency cepstral coefficients (MFCCs) extracted from VRPS voice samples. The Phase II research identified MFCCs to be a base voice feature that can be used to distinguish stress and nonstress voice files, while frequency domain features including spectral-rolloff, spectral centroid, and spectral bandwidth can be secondary features that can accurately be used to create ensemble-learning-based stress prediction and detection models.

The developed SDM classifies testing voice samples as stressed or nonstressed with an accuracy of 69.44%. Experiment 2 (E2) tests the effects of supportive feedback (SF) and competitive feedback (CF) modes on public speaking performance and stress management. The findings reveal that while there was not any buffering effect of the supportive feedback in comparison to CF mode, the SF mode increased arousal, similar to E1 NP condition, suggesting a positive relationship between feedback and stress responses in social contexts which can be utilised for stress management with large datasets and with longitudinal use of SDM integrated VR for stress management. The real-time stress detection and adaptive feedback from the SDM demonstrates the potential for personalised, immersive learning environments that enhance user engagement.

The practical implications of this research extend to fields such as mental health, where VR could be used for exposure therapy in cases of social anxiety and public speaking fears, and education, where real-time feedback systems could improve public speaking and leadership skills. Additionally, the present study offers valuable insights for AI and VR developers, highlighting the potential of affective computing in creating responsive, emotion-driven VR experiences. The adaptability of the SDM suggests applications beyond public speaking, including customer service training and health care, where managing stress is critical for performance.

This thesis contributes to the growing body of literature on affective computing, voice analysis and VR-based training tools, demonstrating the effectiveness of real-time stress detection and feedback in enhancing performance and stress management. Future research should focus on improving and expanding these models to diverse populations and stress-inducing contexts to further refine their applicability and robustness.

Publications

Peer-Reviewed Publications

Chapter 2: Literature Review

Publication 1: Arushi, R. Dillon, A. N. Teoh, and D. Dillon (2023, March). “Detecting Public Speaking Stress via Real-Time Voice Analysis in Virtual Reality: A Review”. In *Sustainability, Economics, Innovation, Globalisation and Organisational Psychology Conference* (pp. 117–152). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-2909-2_7

Chapter 4: Preliminary Work Related to Stress Detection Model

Publication 2: Arushi, R. Dillon, and A. N. Teoh (2021). “Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills”. *2021 IEEE Conference on Games (CoG)*, Copenhagen, Denmark, pp. 1–4. <https://doi.org/10.1109/CoG52621.2021.9618989>

Manuscripts for Publication

Chapter 3: Validation of Stress Induction Capabilities of Virtual Reality

Publication 3: Arushi, R. Dillon, A. N. Teoh, and D. Dillon. “Validation of Social Support in a Virtual Reality Public Speaking Scenario: An Experimental Study on Stress Arousal”. (To be submitted to *Virtual Reality*).

Chapter 4: Development of Stress Detection Model

Publication 4: Arushi, R. Dillon, A. N. Teoh, and D. Dillon. “Voice Feature Analysis and Development of Real-Time Stress Detection Model”. (To be submitted to *Computers in Human Behaviour*).

Chapter 5: Stress Management during Public Speaking with VR Feedback Modes

Publication 5: Arushi, R. Dillon, A. N. Teoh, and D. Dillon. “Effectiveness of Real-Time Stress Detection Model in VR Feedback Modes for Public Speaking”. (To be submitted to *Computers in Human Behaviour*)

Glossary of Terms for this Thesis

Affect: A broad experience of feelings or emotions. Shouse (2005) defined affect as a nonconscious experience of intensity, akin to a raw feeling that precedes any personal experience of emotion. Russell (2003) defines it as a neurophysiological state that varies along two dimensions: valence (pleasantness-unpleasantness) and arousal (activation-deactivation), often leading to emotional responses. According to Russell, affect can be triggered by social, cultural and situational factors, leading to an emotional response.

Emotion: Emotions are specific reactional patterns that vary in nature, intensity, duration and persistence. Scherer (2005) defines emotions as brief episodes of synchronised responses to significant external events. Various theories propose different mechanisms for the emergence of emotions, including physiological responses (James-Lange theory), cognitive evaluation (Lazarus' Cognitive Appraisal theory), or neural processes (Damasio's theory of emotion). Basic emotions like fear, anger, happiness, and sadness have been classified by theorists such as Ekman (1999) and Plutchik (2001). These emotion theories have influenced models used in speech emotion recognition and emotion detection, categorising emotions into discrete or dimensional models for more precise analysis.

Emotional State: An intense and short-lived emotional experience. Emotional states are triggered by external stimuli and manifest more visibly than affective states. These states are accompanied by overt emotional expressions (e.g., fear, joy, anger) and clear physiological responses, such as increased heart rate or blood pressure. Emotional states are triggered by external stimuli and are easily noticeable in speech, facial expressions and body gestures. Emotional states can vary contextually and typically represent more intense emotional experiences, as compared to the broader, more subtle nature of affective states.

Affective State: A subtle, underlying psychological perception of feelings or emotions, typically less intense than emotional states, but often leading to physiological changes. Defined

by Russell (1980) as a psycho-physiological construct, varying along the dimensions of valence and arousal. Affective states, such as stress, may not always be visible but manifest through physical symptoms and can lead to emotions. This thesis adopts Russell's definition of affective states, focusing on their subtle and persistent nature, particularly in the context of public speaking.

Stress: Stress is a complex affective state that arises in response to challenging or demanding situations, often characterised by psycho-physiological changes. It is triggered by external factors and leads to various bodily responses, such as increased heart rate, elevated blood pressure, changes in breathing patterns and alterations in voice. Stress is considered a subtle affective state in this thesis, meaning it often persists below the level of consciousness but can escalate into visible emotional states. Stress-related responses can be detected through voice features during public speaking, making it a key factor in understanding how individuals cope with such high-pressure situations.

Fear: Fear is classified as a basic, specific intense emotion that arises in response to an immediate threat, an aversive stimulus in any given present moment situation. It leads to bodily changes, provoking mobilisation or leading to self-preservation actions in the organism (American Psychological Association, 2018b). It is often typically associated with a specific, identifiable object or cause, making it easier to link the emotional response to a source of danger.

Anxiety: Anxiety on the other hand, is an emotion that is experienced in response to a potential or future oriented threat and it is characterised by feelings of tension and worrisome thoughts along with physical changes such as high blood pressure (American Psychological Association, 2018a). Unlike fear, anxiety may not always have a clear, identifiable object, as it often stems

from uncertain or diffuse threats and anticipation of possible future risks, which may or may not be clearly defined.

Table of Contents

Chapter 1: Introduction to Context, Scope and Significance of the Research	20
1.1 Background.....	20
1.2 Existing Research and Identified Challenges	24
1.3 Purpose of the Research.....	28
1.3.1 Research Objectives and Questions	31
1.4 Significance of the Research.....	33
1.5 Structure of the Thesis	38
Chapter 2: Foundations of the Study: Literature Review and Research Design	41
2.1 Method for Article Screening of Search 1 and Search 2.....	41
2.2 Affective States	44
2.2.1 Role of Stress within Public Speaking and Its Impact on Voice	47
2.3 Voice Analysis.....	50
2.3.1 Voice Features for Emotion Detection	51
2.4 Virtual Reality for Public Speaking	58
2.4.1 The Use of Virtual Reality for Public Speaking.....	60
2.4.2 VR for Assessing, Training and Evaluating Performance to Improving Public Speaking.....	61
2.5 Algorithmic Models for Detecting Affective States.....	70
2.6 Discussion of Findings from Search 1 and Search 2	89
2.7 Identification of Key Gaps and Research Opportunities	93
2.8 Outline of the Present Research and Addressing Anticipated Barriers	94
Chapter 3: Evaluating and Validating Stress Arousal Capability of Virtual Reality for Public Speaking via Experiment 1	99
3.1 Rationale and Aim of Experiment 1.....	99
3.2 Research Questions and Hypotheses of Experiment 1	102

3.3 Design of Experiment 1	103
3.4 Methods of Experiment 1.....	104
3.5 Development of VR prototypes for E1	106
3.5.1 The Virtual Environments for Public Speaking	106
3.5.2 Virtual audience	110
3.5.3 Animation.....	116
3.5.4 Audios	119
3.5.5 Scripts	120
3.5.6 Method	124
3.5.7 Procedure	128
3.6 Results.....	132
3.6.1 Data Preparation.....	132
3.6.2 Assumption Testing.....	133
3.6.3 Preliminary Analysis.....	134
3.7 Main Results: Hypothesis Testing.....	141
3.7.1 Physiological Measures	141
3.7.2 Self-Reported Questionnaire Responses.....	142
3.8 Discussion	143
3.9 Theoretical and Practical Implications.....	145
3.10 Summary	146
Chapter 4: Development of Voice Analysis based Stress Detection Model.....	148
4.1 Preliminary Work	149
4.1.1 Database.....	149
4.1.2 Feature Extraction.....	150
4.1.3 Hypothesis.....	151

4.1.4 Preliminary Results.....	152
4.1.5 Feature Selection.....	154
4.1.6 Training and Testing of Model Development	155
4.1.7 Final Results from the RAVDESS Database Model Development.....	156
4.2 Development of Stress Detection Model	157
4.2.1 Virtual Reality Public Speaking (VRPS) Database	157
4.2.2 Feature Extraction using VRPS Database.....	160
4.2.3 Results from the Feature Extraction	161
4.2.4 Feature Selection.....	166
4.2.5 MFCCs extraction Procedure.....	171
4.2.6 Development of Stress Detection Model (SDM) for VR Integration	176
Chapter 5: Affective Computing Integrated VR Application for Public Speaking Training via Experiment 2	186
5.1 Rationale and Aim of Experiment 2.....	186
5.2 Research Questions and Hypothesis of Experiment 2	187
5.3 Design of Experiment 2	189
5.4 Methods of Experiment 2.....	192
5.5 Development of the Game Modes for E2	193
5.5.1 Game Modes	193
5.5.2 Score System.....	195
5.5.3 Animation and Audio	197
5.5.4 Scripts	202
5.6 Experiment 2 (E2).....	205
5.6.1 Participants.....	205
5.6.2 Measurements and Equipment.....	207

5.6.3 Procedure	209
5.7 Results.....	212
5.7.1 Data Preparation.....	212
5.7.2 Assumption Testing.....	212
5.7.3 Preliminary Analysis.....	213
5.7.4 Main Results Hypothesis: Mixed MANOVA	221
5.8 Discussion	223
5.9 Theoretical and Practical Implications.....	227
5.10 Summary	228
Chapter 6: Discussion and Conclusion.....	229
6.1 Key findings.....	229
6.1.1 Phase I Findings.....	230
6.1.2 Phase II Findings.....	231
6.2 Theoretical Implications	233
6.2.1 Technical Implications: Voice Analysis based Stress Detection	234
6.2.2 Experiment 1 Implications: VR for Public Speaking Stress Induction.....	238
6.2.3 Experiment 2: SDM integrated VR for Stress Reduction	240
6.2.4 E1 and E2 replicability: Implications for Public Speaking Training	241
6.3 Practical Implications.....	243
6.4 Limitations and Suggestions for Future Work	246
6.5 Conclusion	250
References	252
Appendices	290

List of Abbreviations

HCI	Human-Computer Interaction
VR	Virtual Reality
SDM	Stress Detection Model
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
SVM	Support Vector Machine
ANN	Artificial Neural Network
ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
FFNN	Feedforward Neural Network
MFCCs	Mel-Frequency Cepstral Coefficients

List of Figures

Figure 1.1 <i>Thesis Structure</i>	40
Figure 2.1 <i>Circular stretched out clusters using GMM</i>	70
Figure 2.2 <i>Oblong Stretched Out Clusters using GMM</i>	71
Figure 2.3 <i>Possible Hyperplanes</i>	74
Figure 2.4 <i>Neural Network or Artificial Neural Network (ANN) with Layers</i>	76
Figure 2.5 <i>Steps in Development of Stress Detection Model to be Integrated in VR application</i>	95
Figure 2.6 <i>Developed VR application for Public Speaking</i>	97
Figure 2.7 <i>The Overall Process of the Research Design</i>	98
Figure 3.1 <i>Overall Methods of Experiment 1</i>	105
Figure 3.2 <i>Neutral-Positive (NP) Virtual Audience (Positive Behaviour)</i>	106
Figure 3.3 <i>Neutral-Negative (NN) Virtual Audience (Negative Behaviour)</i>	106
Figure 3.4 <i>The stage speaker can walk around and move to right</i>	107
Figure 3.5 <i>The stage speaker can walk and move to front</i>	107
Figure 3.6 <i>The stage speaker can walk around and move to left</i>	108
Figure 3.7 <i>The ceiling of the VR auditorium</i>	108
Figure 3.8 <i>The ceiling of the VR auditorium kept high</i>	109
Figure 3.9 <i>Full auditorium view</i>	109
Figure 3.10 <i>An animation model of woman telling off</i>	113
Figure 3.11 <i>An animation model of woman telling off</i>	113
Figure 3.12 <i>A static animation model of man using his mobile</i>	114
Figure 3.13 <i>An animation model of man clapping</i>	114
Figure 3.14 <i>An initial female character prefab (left) and changed female character prefab (right)</i>	115

List of Figures

Figure 3.15 <i>Neutral positive (NP) animation controller</i>	118
Figure 3.16 <i>Neutral negative (NN) animation controller</i>	119
Figure 3.17 <i>NP audios triggered with animations</i>	119
Figure 3.18 <i>NN audios triggered with animations</i>	120
Figure 3.19 <i>Overall Development for Conducting E1</i>	123
Figure 4.1 <i>Overall significant and nonsignificant values with Preliminary dataset from RAVDESS</i>	152
Figure 4.2 <i>Total number of features showing a significant difference for each actor</i>	154
Figure 4.3 <i>MFCCs (overall for all 40 coefficients), RMS, AE respectively showed 23%, 18%, 16% significant difference between their normal calm (confident) and normal fearful (stressed) vocal expressions</i>	155
Figure 4.4 <i>Comparison of Random Forest, KNN and SVM for detection of stress levels for each of the actor.</i>	157
Figure 4.5 <i>Significance Percentage of each Voice Feature</i>	167
Figure 4.6 <i>Stress Detection Model visualisation with all layers, weights and activations</i>	178
Figure 4.7 <i>Results of stress detection model classification</i>	182
Figure 5.1 <i>Methods for Experiment 2</i>	192
Figure 5.2 <i>Supportive Feedback and Competitive Feedback Game Mode Scene</i>	194
Figure 5.3 <i>Animation States of the SF modes based on the stress levels predicted</i>	199
Figure 5.4 <i>Animation controller in the Unity game engine for the CF mode</i>	200
Figure 5.5 <i>Parameters of SF mode on left and CF mode on right</i>	201
Figure 5.6 <i>Audio clips for SF mode on left and CF mode on right</i>	202
Figure 5.7 <i>Overall work done to develop the prototype for Experiment 2</i>	204

List of Tables

Table 2.1 <i>Different types of Voice Feature Representations</i>	55
Table 2.2 <i>Various Applications of Virtual Reality (VR)</i>	59
Table 2.3 <i>Relevant Studies of Public Speaking in VR</i>	68
Table 2.4 <i>Types of Artificial Neural Network Architecture</i>	77
Table 2.5 <i>Various Models for Affective State Detection from Human Voice</i>	83
Table 3.1 <i>Virtual Reality Scenario Conditions</i>	103
Table 3.2 <i>Correlation Results</i>	136
Table 3.3 <i>T-Test and ANOVA Results</i>	137
Table 3.4 <i>Correlation Results for Posttest Questionnaire Items for NP condition</i>	138
Table 3.5 <i>Correlation Results for Posttest Questionnaire Items for NN Condition</i>	139
Table 3.6 <i>Mean and Standard Deviations: Physiological Measures</i>	142
Table 3.7 <i>Mean and Standard Deviations: Self-Reported Questionnaire Responses</i> ...	143
Table 4.1 <i>t-Test Analysis Results on Extracted Voice Features</i>	164
Table 4.2 <i>MFCC features bands Significance Value</i>	172
Table 5.1 <i>SDM Integrated Game Modes in VR for Public Speaking</i>	189
Table 5.2 <i>Live Stress Detection on Two separate on Voice Samples</i>	196
Table 5.3 <i>Correlation Results</i>	216
Table 5.4 <i>T-Test and ANOVA Results</i>	217
Table 5.5 <i>Correlation Results for Posttest Questionnaire Items in CF condition</i>	218
Table 5.6 <i>Correlation Results for Posttest Questionnaire Items in SF condition</i>	219
Table 5.7 <i>Systolic Blood Pressure, Diastolic Blood Pressure and Heart rate Results</i> ..	222
Table 5.8 <i>Stress and Arousal Results</i>	223

List of Appendix

Appendix A.....	290
Appendix B <i>Poster, and Email to JCU staff for recruiting participants and Social Media Post for Experiment 1</i>	294
Appendix C <i>Scales and Demographic Questionnaires</i>	297
Appendix D <i>Devices Used in Experiment 1 and Experiment 2</i>	301
Appendix E <i>Ethics Approval for Experiment 1</i>	303
Appendix F <i>Feature Extraction Code for all the Voice Samples for RAVDESS and VRPS database</i>	304
Appendix G <i>SDM code</i>	324
Appendix H <i>Count Down Timer C# Code</i>	330
Appendix I <i>Real-time testing Code of the SDM</i>	332
Appendix J <i>Real-time Audio Recording Using Scipy</i>	335
Appendix K <i>Score Adjustment C# Script</i>	337
Appendix L <i>SF Score Adjustment C# Code</i>	340
Appendix M <i>SF and CF Modes Scripts</i>	343
Appendix N <i>Experiment 2 Poster, Email to Lecturers and Social Media Post</i>	349
Appendix O <i>Ethics Approval for Experiment 2</i>	355

Chapter 1: Introduction to Context, Scope and Significance of the Research

1.1 Background

Public speaking is a vital means of communication (Cicero, 1986; Schreiber, 2011; Walters et al., 1993; Whately, 1857). As the name suggests, public speaking encompasses a formalised communication, generally involving purposeful, organised and face-to-face verbal delivery of ideas or information to a large audience. Proficiency in public speaking skills is crucial for advancement across three primary contexts: professional settings, educational environments and broader social initiatives (Lucas et al., 2020; Tucker et al., 2019). However, stress associated with public speaking stands out as a major hindrance and key obstacle, more than other potential challenges leading to ineffective communication. In both professional and educational settings, a large population experiences negative affective states due to public speaking (Sawyer et al., 1999). Over 60-70 per cent of the population experiences the negative affect induced by public speaking (Furmark et al., 1999; Marinho et al., 2017; Raja, 2017) and 64 per cent of college students experience physical symptoms of stress, when faced with public speaking (Marinho et al., 2017). Therefore, stress induced by speaking to a large audience during the development of public speaking skills requires targeted strategies to overcome it. Addressing this stress reaction is therefore essential for maximising the benefits of public speaking and enabling individuals to harness the full potential of their public speaking skills and perform successfully in diverse settings (Gareis, 2006).

Undeniably, mastery of public speaking skills has several elements to it (McCroskey, 2015). A successful delivery requires a comprehensive grasp of the topic, a tailored message for an audience, and logically organised information. Beyond these, the speaker needs to be adept in

employing vocal skills such as speech pace, clear pronunciation and rhythm while maintaining posture, gestures and eye contact to engage the audience. Moreover, the speaker's awareness of feedback from the audience is important; both positive cues such as nodding, posture or clapping and negative cues such as fidgeting, lack of eye contact or checking devices play an important role in a persuasive speech delivery. This interactive loop enables the speaker to adjust the delivery of their speech, thereby enhancing the impact of their speech and enriching the overall communication process (Curtis et al., 2015).

Firstly, in professional contexts, possessing proficient public speaking skills is imperative to help convey ideas, present proposals in meetings and conduct skilful negotiations (Van Ginkel et al., 2015). Such competencies in the workplace enable individuals to be efficient team players, build professional networks, advance in their careers and take leadership roles (Arnold, 2018). Mastery of persuasive skills empowers individuals to confidently articulate pitches in entrepreneurial businesses, enhance career prospects and open doors to new opportunities. In workplace scenarios, where delivery of presentations is necessarily required and obligated, severe stress can impact the message delivery and lower the speaker's confidence. Factors such as fear of judgement, self-doubt and pressure to perform contribute significantly to stress in professional public speaking. This stress can reduce confidence and impede message clarity, weakening a speaker's ability to network, collaborate and lead effectively. By managing stress, individuals enhance their ability to present ideas clearly, build professional relationships and confidently take on leadership roles, which are key components in career advancement.

Secondly, public speaking in educational settings promotes critical thinking, research and organisational skills, essential for students' academic and personal growth. These skills further encourage students to engage with diverse perspectives, opinions and articulate their own

viewpoints coherently. Skilful public speaking contributes to one's personal growth in the academic journey (Burrus et al., 2013; Kallet, 2014) by boosting self-esteem, and enabling individuals to build connections, and maintain and expand social circles. On the other hand in academia, an inability to deal with the stress of public speaking might affect self-esteem and hinder learning and growth opportunities (Feldman et al., 2004). The stress of public speaking often causes students to experience negative affective states, impacting their ability to participate fully in discussions, deliver well-structured presentations, and learn effectively (Dwyer et al., 2012). Factors contributing to this stress include performance anxiety, fear of peer evaluation and physical symptoms like trembling or sweating. These findings show the widespread nature of issues with regard to public speaking among young adults in educational settings (Grieve et al., 2021). By addressing these stress factors, students can gain confidence, engage more actively and develop essential academic skills that foster lifelong learning.

Thirdly, it can also be argued that effective public speaking skills serve as a powerful tool for social change. These skills enable individuals to advocate for critical issues like human rights, environmental sustainability and public health. Historically, individuals skilled in public speaking have influenced societal shifts by articulating complex ideas, leading to changes in communities and inspiring actions. However, the stress linked to high stakes speaking, such as the pressure to deliver a compelling speech, can hinder a speaker's ability to advocate confidently and persuasively. Factors such as apprehension about presenting to a large audience, the pressure to deliver a well-constructed compelling speech, self-doubts regarding one's capabilities, potential for social evaluations and the sudden onset of physical symptoms might contribute to perceptions of stress arousal. Addressing stress in this context empowers individuals to drive social movements and effect meaningful change, illustrating the broader societal impact of effective public speaking.

By honing public speaking abilities, individuals are empowered to confidently articulate and express their ideas, contributing to the advancement of broader societal causes, ultimately shaping opinion and influencing collective action.

Stress has been found to amplify the perceived difficulty of public speaking. Furthermore, despite all the benefits that can be harnessed by acquiring public speaking skills, it is perceived by many individuals as an activity that specifically triggers stress. The experience of stress preceding or during a public speech is an inherently subjective phenomenon (Elfering et al., 2011). Specifically, stress in contrast to other factors that may influence public speaking, such as anxiety surrounding negative social evaluations, self-focused attention, rumination, insufficient rehearsal or unfamiliarity with the topic, interferes with the ability to engage with the audience and deliver a coherent presentation (Schlenker et al., 1982). The associated stress of public speaking is challenging because while inadequate preparation can be mitigated with practice, the effects can manifest suddenly and unexpectedly, leading to a cascade of physiological changes and impacting the voice (Ayres, 1996). These responses undermine even the most well-rehearsed speakers. In the long term, the effects of stress can have several implications for the individual. Repeated stressful experiences in public speaking contexts may lead to avoidance behaviours, diminished self-efficacy, and a persistent fear of public speaking performances creating a cycle of stress that diminishes their performance (Stein et al., 1996). The success or failure of a public speaking performance across varying contexts depends largely on an individual's ability to handle stress during or preceding the speaking performance, making it even more important to develop effective strategies to manage and overcome it.

While there are several benefits that can be derived from having good public speaking skills in various settings including professional, educational and social environments, many individuals

are not naturally adept at it due to difficulties in managing associated stress, its symptoms and significant impact it causes. Stress associated with public speaking can have a severe impact on individual's career progression, missed opportunities, lessened engagement in learning activities and overall self-esteem. It affects even those who are well-prepared, and skilled, making it a primary barrier to effective communication. By learning to address and mitigate the stress, speakers can maintain their composure, deliver the message and information more confidently, and connect more effectively with their audience, ultimately harnessing the full potential of public speaking skills for individual growth, progress and social contributions.

1.2 Existing Research and Identified Challenges

Current literature has consistently shown that public speaking frequently induces a stress response (Koroleva et al., 2014), which can compromise the quality of speech being delivered (Kotlyar et al., 2008; Westenberg et al., 2009; Wiemers et al., 2015). Unlike other factors, research has suggested that affective states such as of stress can have a profound and often immediate impact on public speaking, primarily due to its effects on both the psychological and physiological states of the speaker (Goyal et al., 2016).

Several studies revealed that stress is driven by the psychological state, which is primarily intensified by the activation of the autonomic nervous system (ANS) shown to regulate the body's stress mechanisms (Hellhammer et al., 2010; Wirtz et al., 2013; Won et al., 2016). Stress substantially influences a speaker's psychological state, intensifying anxiety and reducing cognitive processing, contributing to poor public speaking performance. In addition to psychological impact, stress-induced ANS activation contributes to changes in physiology such as increased heart rate, elevated blood pressure, and altered cortisol levels (Dowd et al., 2010; Droppleman et al., 1971; Tecles et al., 2014). Various physiological biomarkers of stress such as

heart rate, blood pressure, finger sweat prints, salivary cortisol (Pisanski et al., 2016) and galvanic skin response, have been well-studied across multiple disciplines such as of biology (Cannon, 1934; Sapolsky, 2007) and psychology (Pisanski et al., 2018; Selye, 2013) and studies have extended the research in the area of detecting stress utilising physiological features (Stephanie Gillespie et al., 2017). Together these changes significantly impair a speaker's ability to deliver coherent and confident presentations (Mandel, 2000).

Research has further explored how combined psychological and physiological changes impact the speakers' voice and its main role in effective communication. Several studies have postulated a convergence between physiological parameters and voice. Among other stress indicators, several research studies demonstrated the human voice's ability to recognise at least 13 different emotions from brief vocalisations, highlighting the expressive power of the human voice (Cowen et al., 2019). These alterations in vocal quality due to varying affective states can negatively impact how a message is perceived by an audience, leading to a diminished persuasive impact and overall effectiveness, thus reducing the effectiveness of speech delivery (Cummings et al., 1992; Hansen et al., 2007). Despite well-documented physiological indicators, their specific impact on voice in high-stress contexts like public speaking is less explored. However, the absence of a stress voice database specific to the context of public speaking remains a major limitation, restricting further research in this area.

Experimentally, stress has been shown to affect different voice features (Tolkmitt et al., 1986). Evidence from several research studies have asserted that voice occupies a more central role in stress detection (Van Puyvelde et al., 2018) as stress marks the voice output negatively (Giddens et al., 2013). Research argues that the voice is extremely susceptible to stress with changes in the features frequently observed under emotionally dysregulating situations (Hansen &

Patil, 2007; Pisanski et al., 2016; Pisanski et al., 2021), making voice a potentially valuable measure for detecting affective states. Research further argues that vocal indices are impacted by stress. Many studies have confirmed that spectral, temporal and prosodic features could be utilised for detection and recognition of emotional states from human voice (Bhavan et al., 2019; Busso et al., 2009; Hao et al., 2020; Hu et al., 2007; Jain et al., 2020; Nogueiras et al., 2001; Nwe et al., 2003a; Ooi et al., 2014; Schuller et al., 2003b; Semwal et al., 2017). For example, changes in fundamental frequency have long been recognised as indicative of stress in a speaker's voice and recently the Teager Energy Operators (TEOs) have been employed to capture subtle stress variations in voice and reported in research (Hansen & Patil, 2007; Rahrkar et al., 2002). Despite a continuously growing body of research recognising the critical role played by the human voice as a good indicator of affective state (Juslin et al., 2003; Mitchell et al., 2013; Scherer, 1995; Scherer et al., 1991) research on recognising stress through voice remains convoluted and unexplored, specifically in the context of public speaking, leaving a critical gap in understanding how vocal features respond to stress in these contexts. Very few studies have explored stress detection during a stressful event like public speaking using nonverbal features (Pfister et al., 2011). The effect of stress on vocal features during social evaluative tasks such as public speaking has remained insufficiently studied. The gap in the literature highlights the need for focused research on how stress manifests vocally within voice features in public speaking contexts, given the unique pressures associated with being evaluated by an audience.

Several previous studies predominantly focused on stress detection via physiological affect measurements in the context of computing device usage (Paredes et al., 2013), and during motor activities (Karthikeyan et al., 2023). Researchers have also employed numerous algorithmic models such as Gaussian mixture models (GMM; Li et al., 2007), Hidden Markov Models (HMM;

Nogueiras et al., 2001; Nwe et al., 2003a), or Support Vector Machines (SVM; Lin et al., 2005) for analysing the human voice to detect different emotions and affective states (for example disgust, sadness, happiness, fear, anger and surprise). Most recently, work has utilised Artificial Neural Network (ANN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN), Deep Neural Network (DNN), Machine Learning (ML) algorithms to recognise human affects (Dahl et al., 2011) in a wide range of studies. However, these effects are not consistently observed across all studies, and there is a notable lack of comprehensive research specifically addressing these indices in the context of public speaking (Giddens et al., 2013). Furthermore, research on algorithmic models specifically designed to detect stress during public speaking remains relatively unexplored and less investigated. The limited development of stress detection models specific to the public speaking context, especially those using context-specific voice databases, highlights the need for a focused exploration of computational methods that can detect stress in real-time.

Present research has highlighted that interventions for stress reduction during public speaking have largely focused on training programs (Fredrick et al., 2018; Teoh et al., 2018; Thorsteinsson et al., 1999; Uchino et al., 1997). Additional methods such as cognitive behavioural treatments (Heimberg et al., 1993), self-preparation, and behavioural rehearsals (Menzel et al., 1994) as well as development of techniques have been employed to develop public speaking skills (Fawcett et al., 1975; Rhodes et al., 1975). Other approaches such as speech language pathology programs (Breakey, 2005), curriculum tailored to the voice and speech mechanism (Hancock et al., 2010) have also been found to reduce apprehension and increase feelings of confidence and competency in undergraduate students. Face-to-face short courses (Mowbray et al., 2015) and group training tutorials (Knight et al., 2016) offered by organisations such as Toastmasters

International, remain popular for improving public speaking among students (Toastmasters International, 2024; Yu-Chih, 2008). Awareness training to decrease nervous habits (Spieler et al., 2017) and self-hypnosis (Ilmi et al., 2017) to improve self-confidence were also reported in research literature enhancing public speaking skills. Recently technological advancements particularly VR for rehearsing of public speaking skills have emerged as novel training tools academic, as highlighted in literature (Bachmann et al., 2023; Palmas et al., 2019; Vallade et al., 2021). While the use of VR in public speaking training has been found to be effective, (North et al., 1998; Poeschl, 2017; Takac et al., 2019), limited studies using VR (Kothgassner et al., 2019) evaluated its usage in alleviating specifically psychosocial stress based on real-time and interactive feedback based on voice analysis (Chollet; et al., 2015; El-Yamri et al., 2019; Hartanto et al., 2014). Hence, there remained a need to explore the integration of VR with affective computing models for voice stress detection tailored to providing real-time feedback for public speaking training, particularly focusing on stress management.

The review of the existing literature has helped identify three key gaps in the field specifically in the context of public speaking. These include the absence of a comprehensive database of stress-related voice samples, the need to identify accurate voice features indicative of stress during public speaking, the development of a machine learning (ML) model for voice analysis-based stress detection, and the seamless integration of this model into VR technology for real-time voice-based stress detection.

1.3 Purpose of the Research

The present research stems from the need for innovative solutions to address stress in real-time via focusing on voice analysis, particularly during public speaking contexts where stress can be a major impediment to effective communication. The current research has been inspired by the

recent advancements in the field of affective computing (Picard, 1997), which is a field of study focused on recognising, interpreting, simulating human emotional responses using computing systems. The rapid evolution of immersive technologies which are transforming various domains, encompassing VR have also motivated this research, offering an economically viable tool for affective computing and human-computer interaction research due to its potential to simulate immersive scenarios, deliver personalised feedback, and elicit emotions in users (Somarathna et al., 2022).

Affective computing is a field of computer science that was defined by the foundational work of Rosalind Picard in her book of the same name *Affective Computing* (Picard, 1997). Picard emphasised in her book that “computers do not need to have an understanding of emotions to become humanoids, but to respond intelligently and sensitively towards humans”, thus making holistic human-computer interactions (pp. 1, 72). Over the past 30 years, research in the field of affective computing has expanded steadily and continues to grow as applications are designed to provide better usability, enhanced user experience and human-computer interaction (HCI). Today, integration of affective computing is common across various domain applications including mental health care (Lefter et al., 2023), e-learning (Mejbri et al., 2022), training, remote detection of stress (Karthikeyan et al., 2023; Paredes et al., 2013), and music emotion recognition (Cui et al., 2022; Dillon, 2001, 2003; Xia et al., 2022). Earliest application areas of affective computing were focused on systems that were used to recognise human emotions from real-time audio cue analysis for understanding the expressions in musical performances (Dillon, 2001) and physiological parameters analysis determining the computer user’s state. Despite all the progress made, while advancements in this field have enabled applications across mental health, e-learning and user

experience design, there remains a gap in understanding and accurately detecting stress responses through voice in high-pressure situations, such as public speaking.

Applications of affective computing research within VR settings also hold immense promise (Harris et al., 2023; Kalatzis et al., 2021). VR is a human-computer interface that immerses users in a simulated environment and provides a safe and interactive space for various purposes (Burdea et al., 2017; Cummings et al., 2016; Mihelj et al., 2014; Suh et al., 2018; Wohlgenannt et al., 2020) including training (Azwar et al., 2016; Nijholt, 2014; Rosén et al., 2019). VR technology can track changes in the user and enable real-time adjustments to the virtual environment so users can learn without fearing the consequences of making mistakes (Wiederhold et al., 2005). Within the context of interpersonal skills training, while public speaking is known to induce stress in real-life situations as well as controlled laboratory settings, VR offers a dual opportunity to effectively induce and mitigate stress, providing a controlled, automated and consistent immersive environment with a virtual audience for training of stress management and skill development (Schmid Mast et al., 2018).

The present research is motivated by the need to better understand the stress experienced during public speaking, specifically its impact on the speaker's voice, aligned with the three key gaps identified in the literature. In the field of affective computing, there is a growing interest in identifying emotional markers of stress such as vocal changes that can help individuals improve communication skills. To achieve this, the research requires a specific contextual corpus of stress-related voice samples collected during public speaking. By the development of a machine learning-based voice analysis stress detection model, the present study seeks to incorporate insights from psychology and IT expertise in feature engineering techniques. It seeks to establish a new approach to enhance public speaking skills by development and integration of a voice-based stress detection

algorithmic model in a VR application. The approach is designed to address stress management during public speaking, thereby fostering confidence and self-esteem in individuals, enabling them to practice overcoming stress through consistent feedback in a safe environment, and thereby be able to deliver speech to a larger audience in real-life. By combining affective computing voice analysis advancements in ML and VR with psychological insights into stress responses, this research contributes to the field of affective computing and provides a practical solution for overcoming one of the most common barriers of stress for effective communication in educational, professional and societal environments.

1.3.1 Research Objectives and Questions

The present research aims to develop an algorithmic model for stress detection and design feedback mechanisms for public speaking training. Achieving this has required two steps: (i) developing a stress detection model and (ii) designing and testing feedback mechanisms. The research has been guided by four objectives:

- develop a context-specific stress voice database
- identify voice features that help distinguish and detect stress hindering effective communication in speakers
- use voice analysis techniques to build machine learning stress detection models for real-time automated stress analysis during public speaking
- design and develop a VR game-based training program to support individuals to self-regulate their stress during public speaking and practice, thereby empowering them to become more confident and impactful public speakers.

This research approach has combined the techniques and algorithms from affective computing and HCI to detect stress during public speaking. Affective system development is highly context-

dependent due to various influencing factors such as the type and intensity of a user's affective state, the method of state induction, and the way it is expressed (Picard, 1997). In the scenario of public speaking, stress is a subjective perception of the individual's psycho-physiological and cardiorespiratory response in the body, which is usually triggered by the activation of ANS and is expressed through voice. Specifically, identifying stress during public speaking requires an approach distinct from conventional affective computing methods for stress assessment. Thus, this research has proceeded (i) by developing a public speaking stress voice database, (ii) correlating physiological indicators of stress with vocal samples collected within VR-based public speaking scenarios, and (iii) identifying stress voice samples. In the next steps, research focuses on (iv) utilising this database for further research according to the objective 1 and objective 2 and therefore, (v) developing a voice-based stress detection model (SDM).

In this PhD thesis, current methods from affective computing and human-computer interaction are applied in the development of a VR application for public speaking training as mentioned in objective 3. The VR application for public speaking training needs a system that measures user stress from voice in real-time. Additionally, the designed VR application for public speaking should provide relevant feedback according to the identified stress levels.

The complete development of the VR application for public speaking, integrated with an SDM, addresses research questions identified from the literature review (Chapter 2). In particular:

- 1. Research Question 1 (RQ₁):** Could the different behaviours demonstrated by the virtual audience in both supportive and nonsupportive scripted VR scenarios, as provided in our specifically developed simulation, impact on a speaker's stress level irrespective of their performance, potentially leading to stress induction in the users?

- RQ₁ addresses the first objective involving creation of a stress voice database using scenarios designed in VR environments. The resulting database is then used to identify the stress vocal features for development of the stress detection model.
- 2. Research Question 2 (RQ₂):** What reliable set of voice features (e.g., prosodic, spectral) can be analysed in real-time (established in correlation to the physiological parameters e.g., blood pressure, heart rate) from recorded voice to detect stress level in a speaker?
- RQ₂ directly addresses the second objective of identifying stress-indicative features and the third objective of developing the stress detection model for recognising affective states during public speaking.
- 3. Research Question 3 (RQ₃):** How could the previously developed model, based on the identified set of voice features of stress, be effectively used to provide real-time feedback? Could this approach then test the stress-buffering effects of a virtual audience based on detected stress in speakers?
- RQ₃ examines the final objective which focuses on the integration of the stress detection model and its applicability in providing feedback mechanisms to manage stress during public speaking training through the VR system.

Respective hypotheses formulated from research questions are included in experiment Chapter3, Chapter 4 and Chapter 5.

1.4 Significance of the research

The research addresses important challenges identified in existing literature (outlined in Section 1.2) concerning stress detection in the context of public speaking training. Public speaking is an essential skill, crucial for an individual's academic learning and performance, professional

success, and personal confidence (Carnegie et al., 2017). However, stress during public speaking has remained a pervasive obstacle for many individuals, impacting an essential communication channel i.e., the voice of the speakers. By focusing on developing a SDM through voice analysis and integration into VR, this research contributes to enhancing individuals' ability to communicate confidently. Furthermore, the present research contributes to the field of affective computing by developing a voice-analysis-based SDM and integrating it with immersive VR technology.

Despite the availability of many traditional public speaking training approaches used for practicing public speaking, each method has its own drawbacks. Accessibility is of major concern, as not all methods are readily available when an individual needs them for practicing speaking skills. Many existing approaches aimed at improving public speaking training do not always provide personalised and consistent feedback or support when needed based on stress detected in speaker's voice. As a result, they might not have equipped individuals with practical tools to manage stress in the long term. Moreover, due to absence of a consistent training environment and availability of an audience, the effectiveness of these methods may vary among individuals and may have not addressed the root causes responsible for stress accurately. These approaches often show lack of strategies to detect stress. Although VR applications have been developed for training public speaking, primary focus has been on inducing affective states such as fear and anxiety in speakers (Slater et al., 2006; Slater et al., 1999), examining the impact of virtual audience (Hancock et al., 2010; Hartanto et al., 2014), investigating various feedback strategies (Chollet; et al., 2015; Wörtwein et al., 2015), and evaluating speech quality (Pellett et al., 2019; Schneider et al., 2015; Tanveer et al., 2015). Limited studies have used VR (Hartanto et al., 2014; Kothgassner et al., 2019) to explore stress and alleviate it, based on real-time and interactive feedback based on voice analysis (Chollet; et al., 2015; El-Yamri et al., 2019).

Earlier stress recognition in previous methods employed physiological affect sensing (Healey et al., 2005; McDuff et al., 2014; Sano et al., 2000), or physiological features (Stephanie Gillespie et al., 2017) but very few studies have extended the research in the area of detecting stress using either nonverbal features from databases (Hansen & Patil, 2007; Pfister & Robinson, 2011), during any stressful event like public speaking. Research lacked a comprehensive understanding of voice features that can indicate stress in a speaker's voice during public speaking (Greene et al., 2016). Detection of stress during public speaking is crucial for providing feedback to an individual learning public speaking. Real-time stress detection during public speaking has been underresearched and has received limited attention due to the challenges and limitations in creating a suitable testing environment and lack of context-specific stress voice database. While stress has been studied in various emotional contexts, its impact on voice during public speaking remains underexplored. Recent technological methods for training public speaking with VR often lack clarity on virtual audience training for affective state detection and feedback mechanisms. This research makes significant contributions to the fields of affective computing and human-computer interaction by addressing key challenges in stress detection during public speaking, with a specific focus on voice features related to stress. The contributions of this research are as follows:

1. **Development of a Context-Specific Voice Database for Public Speaking:** First, this research contributes by developing a much-needed voice database, designed to facilitate the identification of voice features indicative of stress during public speaking, and to support the advancement of stress detection algorithms.
2. **Enhanced Voice Analysis based Stress Detection Capabilities in Public Speaking:** Second, the present research contributes an understanding of how stress manifests in voice, specifically within public speaking contexts. By identifying MFCCs as the base

voice features indicative of stress, and spectral features as secondary voice features indicative of stress, it lays the foundation for developing reliable, voice-based stress detection algorithms development for the context of public speaking. These features can be further explored to create ensembled based ML models for stress detection and could be implemented for stress recognition in this contexts withing public speaking training applications and research.

3. **Creation of a Machine Learning-Based Stress Detection Model (SDM):** The research develops a SDM capable of real-time stress detection through voice feature analysis, utilising ML techniques to improve accuracy and applicability within realistic public speaking settings integrated into VR environments. This advancement contributes to affective computing research by addressing the underexplored area of stress detection in public speaking contexts, a focus that has been largely overlooked in prior studies.
4. **Integration of the SDM into a VR-Based Training Environment:** By incorporating the SDM into a VR application, this research provides a practical, immersive solution for public speaking training. The VR application offers real-time, personalised feedback based on voice analysis for stress management through two distinct training modes, allowing users to practice managing stress and improving their skills in a safe, controlled setting, which was unclear in previous studies.
5. **Exploration of Virtual Social Support (VSS) as a Stress-Buffering Mechanism:** Through the use of a virtual audience in stress induction and using real-time affective state detection to provide feedback in real-time, this study investigates the stress-buffering effects and gained insights into use of VSS for management of stress. The approach offers a new way to use VR in managing public speaking stress with responsive

virtual audience, contributing a unique perspective on how virtual audiences can help reduce stress and used for public speaking practice sessions.

6. **Provision of an Accessible and Consistent Training Tool for Skill Development:** The VR-based public speaking tool developed in this study provides personalised, consistent feedback within two different practice modes, as that was not easily achievable through traditional training methods dedicated for improving public speaking. The research supports users in developing confidence and effective speaking skills over time by offering a structured, consistent training environment with feedback mechanisms based on stress levels detected in voice, and adaptive VR simulations for public speaking.

In sum, the potential contributions of this research span both academic and practical domains.

From an academic standpoint, this study has made contributions to the field of affective computing by expanding the stress detection research via voice affect sensing. This study has stood out as a pioneering effort to develop an SDM specifically dedicated to identifying stress through voice during public speaking. The successful implementation of the SDM in a VR environment is designed to enhance public speaking skills, thus also expanding its contribution to the field of HCI. The integrated SDM enables the virtual audience to provide real-time feedback through two modes: a social supportive mode for those less confident about their speaking abilities; and a realistic feedback mode for those looking to refine their skills. This adaptive approach allows users to gradually overcome public speaking stress and improve their performance. The innovative concept of real-time public speaking training, based on voice-based stress detection, with real-time feedback from virtual audience, guides participants to modulate their voice under stress, hence uniquely contributing to merging of affective computing and HCI in immersive technology research.

Beyond academia, this stress voice analysis VR application can be integrated into public speaking training programs for students, early career professionals and individuals looking to improve their public speaking skills. The real-time feedback and stress detection features offers a personalised learning experience, allowing users to practice and refine their speaking abilities in a controlled environment. Additionally, corporations can utilise the developed VR application to train their employees. By identifying stress, staff members can work on improving their effective and clear communication skills. Mental health scenarios could utilise the SDM for analysing stress in the voice of clients as well as counsellors and therapists. The SDM can also be used to enhance public speaking training in other environments, such as Toastmasters programs, by accurately detecting stress and enhancing peer feedback. Implementing the SDM (or developed by following similar approach) can help monitor the stress levels in allied mental health professionals such as nurses and call centre operators, and even consumers. This will allow faster transfer of cases and services to be provided.

1.5 Structure of the thesis

This thesis comprises six chapters, each contributing uniquely to the overarching research objective as shown in Figure 1.1.

Chapter 1 has offered an introduction to the research topic, outlining the identified problem, background, significance and the contributions of this study.

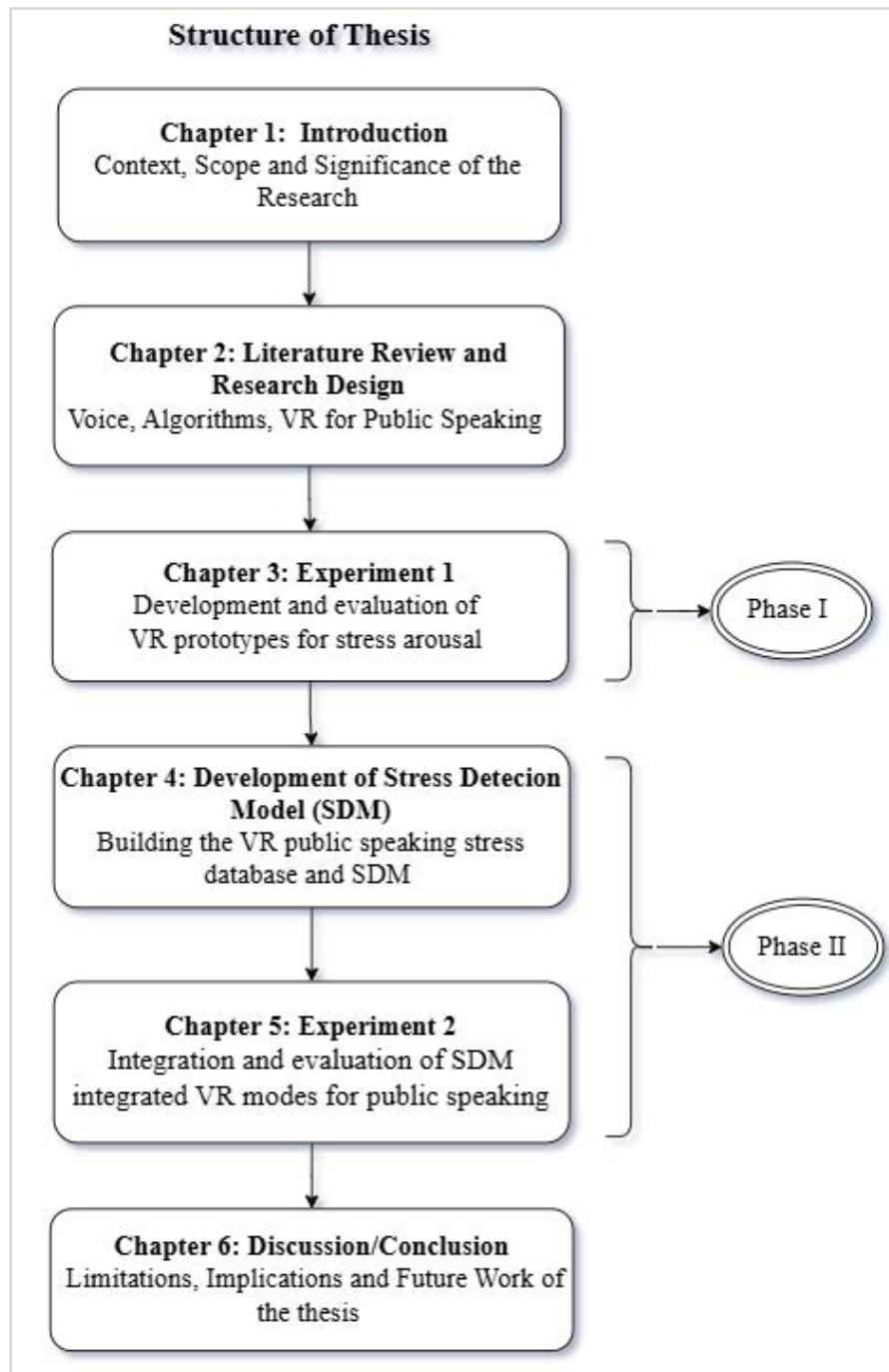
Chapter 2 delves into a comprehensive literature review, summarising studies in affective computing and human-computer interaction, focusing on voice and stress analysis, as well as algorithms used in emotion detection. It also highlights recent advancements in using VR to enhance public speaking skills and identifies existing research gaps and overall research design.

Chapter 3 presents the first experiment, which involves the development and testing of initial VR prototypes aimed at assessing stress arousal capabilities during public speaking.

Chapter 4 provides an in-depth description of the development of the ML SDM for real-time stress analysis.

Chapter 5 focuses on the second experiment, detailing the testing of final VR prototypes designed to evaluate the stress-buffering effects of a virtual audience in public speaking scenarios.

Chapter 6 concludes the thesis with a comprehensive discussion and summation of the entire research study. Finally, all the references cited throughout the thesis are included towards the references section in end.

Figure 1.1 *Thesis Structure*

Chapter 2: Foundations of the Study: Literature Review and Research Design

This current chapter provides an extensive review of the literature on affective computing and VR. It integrates research from psychology and human-computer interaction to examine public speaking, its effect on voice, voice analysis for emotion recognition, the use of VR for enhance public speaking skills and algorithmic models for affective state detection. It also provides an outline of the research design.

The review begins by outlining the method for selecting literature relevant to these key topics (Section 2.1). It then explores the concept of affective states, with a focus on stress in the context of public speaking, which is a common phenomenon that adversely impacts speaker performance and undermines self-confidence (Section 2.2). The following section reports the findings of literature that examines voice analysis and voice features related to different affective states and emotions (Section 2.3). Next, studies related to use of VR technology for improving public speaking through the integration of voice analysis techniques are explored (Section 2.4). Research related to the development of algorithmic models for emotion and affective state detection is discussed (Section 2.5). Section 2.6 includes a summary of the findings from the literature searches and Section 2.7 concludes the review. Finally, section 2.8 provides an overview of the present research's design.

2.1 Method for Article Screening of Search 1 and Search 2

The current literature review focuses on three key areas: (i) voice analysis and the use of VR in the context of public speaking, and (ii) algorithmic models for affective state and emotion recognition via voice/audio analysis. Research papers published in major indexed journals were reviewed. Two different searches were conducted over a timeframe of one year. Search 1 provided

a broad understanding of the research area and identified initial gaps regarding the impact of affective states/emotions on voice and the use of VR for public speaking. Search 2 helped uncover specific gaps in the literature, thereby laying the foundation for formulating the research aim. Articles from both searches were screened and managed using EndNote software (The EndNote Team, 2013). Peer-reviewed articles and book chapters were identified using Google Scholar, IEEE Xplore databases, ScienceDirect and Web of Science. Only English-language publications were included. Common elimination criteria for both Search 1 and Search 2 included removing articles that were less than a single page, as well as commentaries, reflective viewpoints and short papers with no data. Additionally, research articles not related to voice feature analysis, VR for public speaking or emotional and affective state detection models were excluded.

Search 1: The first search sought to identify literature that clarified the relationship between voice features indicative of various emotions or affective states and the use of VR for improving public speaking. Search 1 offered an understanding of affective states and respective voice features, and use of VR technology within the context of public speaking. Inclusion criteria for the articles from Search 1 was focused on public speaking, specifically the connection between voice, emotions and VR for public speaking. Keywords included *voice and emotion*, *public speaking and virtual reality*. A timeframe of 50 years (1970-2019) was set and this search yielded important earlier research papers such as by Scherer (Scherer, 1995), Wallbott and Scherer (Wallbott et al., 1986), and Kappas (Kappas et al., 1991), which were helpful in the basic understanding of the relationship between voice and emotional/affective state in humans. Similarly, research by Slater and colleagues (Slater et al., 2006; Slater et al., 1999) helped elaborate the earlier research methods and experiments related to public speaking in VR. Search 1 initially

resulted in a total of 686 articles. After the exclusion measures were applied, 196 articles were included in the review.

Search 2: A second search focused on identifying previous research related to real-time stress detection using voice analysis. Inclusion criteria focused on including articles with information about the use of voice and relevant models for detecting various affective states in speakers. Keywords included *audio and voice analysis*, *speech emotion detection* and *voice stress detection* were used to search the relevant articles. A timeframe of 20 years (i.e., 2000-2020) was set to capture the recent advancements in algorithmic models for emotion and affective state detection. This search provided more insights into recent technological progress made in ML technology for emotion detection. Search 2 resulted in 438 articles. Out of these, 124 articles met the inclusion criteria and were included in the review.

A total of 302 articles were screened and 133 articles which were not relevant and coherent with our research area such as with detailed focus on effective public speaking mechanisms, stress-buffering hypothesis were excluded. Following the comprehensive screening process, a total of 169 articles aligned with the study objectives were included in the literature review. Additionally, during the iterative refinement process, sources such as American Psychological Association definitions, Unity Engine, Python Libraries, several voice databases, and reference management software were also included. This resulted in the reference count of articles included in the literature review to 178.

Search 1 resulted in findings which are organised under Sections 2.2, 2.3, 2.4, providing an understanding of the key research areas relevant to this study with a focus on public speaking. Search 2 resulted in findings focused on understanding of real-time algorithmic models covered

under Section 2.5, with specific focus on voice analysis techniques for affective, emotion detection and recognition.

2.2 Affective States

Rosalind Picard's 1997 *Affective Computing* marked a turning point in the study of affective states, as computer science emerged as a significant contributor to the field by providing framework, methodologies, mechanisms for computing machines to simulate, interpret and respond to human emotions. Across all three domains, the terms 'affect', 'emotions', 'affective states', and 'emotional states' are frequently employed, despite their meanings often being ambiguous and confusing. To clarify these concepts and their relationship to the affective states commonly experienced during public speaking and their consequences, I explore the complexities of existing definitions and descriptions of affect and emotions and their respective states.

A universally accepted definition of affect is lacking, with the term generally understood to refer to a broad experience of feelings or emotions. The primary issue with philosophical discussions/explanations lies in the ambiguity and generality of the term. Shouse (2005) defined affect as a nonconscious experience of intensity, akin to a raw feeling that precedes any personal experience of emotion (Shouse, 2005). However, this initial philosophical abstraction remains unstructured and abstract. Russell (2003) offers a more structured definition, describing affect as a neurophysiological state that is a continuously present, free-floating and nonreflective feeling that integrates and varies along two primary dimensions: valence (pleasantness-unpleasantness) and arousal (activation-deactivation). According to Russell, affect can be linked to causes, which can be social, cultural, and situational, and then leads to an emotional response (Russell, 2003). His definition encompasses the pleasantness-unpleasantness and activation-deactivation aspects

experienced by an individual in any given context that can be subjective, making it comprehensive and objective, and keeping it distinct while allowing for variability at the same time.

The study of affective states has traditionally been the domain of philosophy and psychology (Scherer, 2000, 2005; Shouse, 2005). The definition of affective states remains broad brushed in literature. Prominent researchers like Scherer have used affective states as an overarching umbrella term that encompasses emotions, moods, feelings and emotional states (Scherer, 2000). However, this way of portraying affective states does not distinguish properly between subtle nuances that operate for classifying and drawing a line between visible states in any given contextual and situational fields. Russell (1980) defines affective states as a psychophysiological construct thereby linking mental and physiological processes. Building on Russell's (2003) structured definition, which describes affect as a neurophysiological state varying along the dimensions of valence (pleasantness-unpleasantness) and arousal (activation-deactivation), these affective states can be represented within these two dimensions, as outlined in Russell's model (Russell, 1980). Russell's model is an integrated approach to affective neuroscience, cognitive science, and psychological sciences, incorporating all the individualistic subjective and contextual elements of the experiences.

Emotions on the other hand remain complex, fuzzy, and poorly defined (Calvo et al., 2010). Emotions are specific reactional patterns that vary in nature, duration and persistence (Hakim et al., 2013). Scherer (2005) used the term emotion to define brief episodes of synchronised responses to significant external events (Scherer, 2005). Picard notes that various emotions may share some components, but others may be unique (Picard, 1997). There have been several theories such as the James-Lange theory (James, 1884), Cannon Bard theory (Cannon, 1927), Lazarus' Cognitive Appraisal theory (Folkman et al., 1986) and Damasio's theory of emotion (Damasio, 1998)

proposing different mechanisms, suggesting that emotions arise due to physiological responses, result from cognitive evaluation or involve complex neural processes. For instance, Ekman's Basic Emotions Theory (Ekman, 1999) and Plutchik's Wheel of Emotions (Plutchik, 2001) have attempted to explain the nature and scope of emotions and also categorised them into basic and secondary emotions. These theories have subsequently led to creation of several models (i.e., discrete and dimensional models of emotion) used to categorise emotions and are employed in several contexts and areas of speech emotion recognition and emotion detection in computer science. In affect analysis, emotional states are typically characterised along either of the two primary models of emotions: the discrete model (Lazarus, 2000) and the dimensional model (Watson et al., 1985).

Emotional states embody intense and specific emotions, characterised by their shorter duration and greater intensity, often accompanied by specific overt emotions (such as fear, joy, and anger) and distinct physiological responses (such as increased heart rate and blood pressure) making them more noticeable. However, emotional states, similar to affective states, can also manifest in speech and emotion-specific body gestures (Baltrušaitis et al., 2012), specific to any given context.

Based on several definitions of affect, affective states, emotions and emotional states, this research draws upon these definitions, adopting Russell's definition of affect and affective states. Hence, affective states are defined as subtle, underlying subjective psychological perceptions regarding feelings or emotions, accompanied by context-specific physiological phenomena. These states persist subtly, varying in intensity, pleasantness and arousal level. Affective states, like stress, are less visible and more based on self-perception, leading to various bodily responses and

ultimately manifesting as emotions. Using this definition of affective states, stress is considered an affective state rather than an emotion, underlying emotional states of fear and anxiety.

Now that a clearer definition of affective states has been operationalised for this research, it is important to delve into manifestations of these states.

Affective states such as stress and emotions such as fear and anxiety are expressed through physiological and psychological indicators such as facial expressions and vocal parameters. Understanding and managing affective states such as stress is essential for effective public speaking, as they influence both the speaker's experience and the audience's perception and response in any given context.

2.2.1 Role of Stress Within Public Speaking and Its Impact on Voice

The present section is dedicated to exploring the affective state of stress in specific context of public speaking and highlights the interaction between physiological, psychological and vocal parameters of stress for refining of affective state detection and recognition models.

Context plays a crucial role in shaping the representation, simulation, recognition, and construction of affective states (Picard, 1997). Context is known to impact the subjective experience of affect, including the personal meaning, significance and quality of the experience. In other words, the context in which an individual finds themselves significantly impacts the emergence, manifestation, intensity, and subjective experience of emotions and affective states, and therefore, must be carefully considered when studying or attempting to understand emotional experiences. During the context of public speaking, an effective delivery of a message requires speakers to possess crucial skills, including confidence and the ability to manage underlying feelings, moods and emotions. A positive affective state of the speaker such as of enthusiasm and confidence may be exhibited through more engaging body language, tone, and facial expressions,

enhancing delivery of message, engagement, and overall performance (Atkinson, 1984; Kenny, 1982). Conversely, a negative state of emotions like fear or anxiety (Gallego et al., 2022; Pribyl et al., 2001) and affective state of stress (Bassett et al., 1987; Carrillo et al., 2001) may detract from the message, as physiological responses such as rattling knees, increased heart rate, sweating and tremors, shakiness in voice, or changes in the breathing pattern become more apparent and overwhelming for the speaker (Bodie, 2010). In any public speaking scenario, stress is one of the predominant affective states for many individuals (Chanwimalueang et al., 2016; Giraud et al., 2013). The aim of this section is to understand how stress and other emotions affect voice.

The concept of stress has its origins in physics where it was initially defined by Augustin-Louis Cauchy (1827) to describe the force acting on unit area of given material (Rivlin, 1948). The same term was later adopted by Hans Selye (1950) in psychology, where it was used to define subjective psychological and physiological response to any internal or external stressors (Selye, 1950). Unlike specific emotions, such as fear and anxiety, stress is a broader concept and a multifaceted experience affecting cognition, emotions, and physiology of an individual leading to various experiences of emotional (negative emotions), cognitive (negative thoughts), and physical responses (such as palpitations, sweating, fidgeting, accelerated speech) (American Psychological Association, 2018c). Given its subjective nature and the multifaceted internal self-perception of responses to feelings, emotions and moods, stress can be classified as an affective state and is used in this research as such. Generally, in any given stressful situation, a moderate level of stress is beneficial for optimal performance, though excessive stress impairs speech quality and reduces overall performance (Dongrui et al., 2010).

Studies indicate that public speaking induces a state of high physiological arousal, characterised by increased heart rate, sweating, and other symptoms of anxiety (Krocze et al.,

2023). This high arousal state is similar to the feeling of excitement but differs in its valence or emotional tone (Brooks, 2013). Stress during public speaking is described as having negative valence – an unpleasant, aversive emotional experience with a low sense of control, quite similar to anxiety (Brooks, 2013; KroczeK & Mühlberger, 2023). In contrast, excitement has a positive valence and is considered a pleasant emotion (Brooks, 2013). So, while both stress and excitement involve high levels of arousal or physiological activation, stress during public speaking specifically involves high arousal coupled with negative valence or an unpleasant emotional state. The key distinction is that anxiety, similar to stress, has a negative emotional tone, whereas excitement has a positive emotional tone despite similar high arousal levels (Brooks, 2013).

Stress significantly impacts the voice due to its physiological and psychological effects on the body. Stress and voice processes tend to rely on the same cardiorespiratory process in alignment with ANS (Berntson et al., 1991). Consequently, stress during public speaking impacts all three fundamental processes of voice production, i.e., (i) respiration, involving airflow generated by the lungs for speech production, (ii) phonation, referring to the vibration of the vocal cords for sound production, and (iii) resonance, which determines the quality of sound shaped by the vocal tract and other cavities used for amplification. The changes in the three processes of voice production lead to increased muscle tension (Titze et al., 1998). Stress can also cause variations and asymmetries, leading to strained, unstable and a less resonant voice. Individuals may further experience vocal fatigue, involuntary pitch changes, voice tremors and a dry throat, all contributing to a weak, shaky or scratchy voice. These inconsistencies often manifested as hoarseness are worsened under stress. Stress alters breathing patterns causing shallowed breaths and even reduced voice stability (Kreiman et al., 2011).

Stress has been found to influence voice features such as jitter (the variation in the frequency of vocal fold vibrations) and shimmer (the variation in the amplitude of vocal fold vibrations), decreasing these parameters, with a more pronounced impact during emotional emergencies compared to cognitive load (Sapolsky, 2004). Features of voice, such as fundamental frequency (F_0) values, jitter, respiration rate, articulation rate, word duration, vowel duration, respiration time between words or sentences, and voice onset time (VOT) are all speaking voice features that can be related to stress (Van Puyvelde et al., 2018).

The subtle yet profound impact of stress on voice, particularly in the context of public speaking, highlights a compelling need for further investigation and exploration. While most previous studies have focused on the physiological symptoms of stress, there remains a gap in understanding its effects on voice. In the following section, literature related to voice analysis techniques is explored.

2.3 Voice Analysis

Continuing with the aim to understand and identify the relationship between different emotions, affective states and their effect on voice, this section includes research in which voice analysis has been explored.

Voice analysis is one of the specialised fields within affective computing that focuses on detection and interpretation of emotions and affective states from vocal signals (Schuller et al., 2013). Over the past decade, the analysis of both speech and nonspeech vocalisations (voice) has received much recognition to detect various emotional and affective states (El Ayadi et al., 2011; Schuller et al., 2011). The process of affective state detection through voice analysis requires identifying relevant voice features and training algorithms that can accurately interpret these features within specific contexts and datasets. Detecting emotions or affective states via voice

analysis also involves some preliminary steps such as preprocessing, signal framing, windowing, detection of voice activity, normalisation, and noise reduction of the data (Hamid, 2018). These steps are critical for extracting accurate and specific voice features that can effectively identify different affective states and for training algorithmic models to classify and recognise emotions accurately. An algorithm is then used to map input (e.g., voice features) to output (e.g., detected emotion or affective state) using a functional relationship. For instance, if a given input is X (various voice features) and an output Y (a specific emotion or affective state), then the function involves mapping the given input X to Y as $f(X) = Y$.

Feature extraction is central to the process of detecting and identifying specific emotional or affective states. However, this task becomes challenging when considering context, events and environmental factors. For example, identifying features to detect stress in real-time during the event of public speaking is complex and challenging. Common physiological markers, such as increased heart rate and blood pressure, along with psychological indications from self-reported questionnaires, may not exclusively confirm stress experiences. Thus, accurately extracting and analysing voice features for stress detection in dynamic changing situation such as of public speaking remains a challenge.

2.3.1 Voice Features for Emotion Detection

This section provides an overview of the various voice features commonly explored in research for detecting stress and related affective states, such as anxiety and fear, as well as other emotions from vocal signals. It first introduces the categories of voice features, such as prosodic, spectral and temporal features, followed by a discussion of how these features are used in the context of emotion and affective state detection. Finally, the section highlights key findings from the literature, emphasising their relevance to stress detection in public speaking scenarios.

Human voice is capable of conveying emotions and affective states through speech (Banse et al., 1996; Darwin et al., 1998; Scherer et al., 1991; Wallbott & Scherer, 1986) and nonspeech vocalisations (Cowen et al., 2019). The communication of emotions in humans takes place beyond the literal spoken words, such as in nonverbal patterns of rhythm, timbre and tone in speech. Overall, many nonlexical patterns can be identified by features such as intensity, fundamental frequency (F_0) and duration of the sounds. The temporal, frequency domain, spectral and prosodic features are of paramount importance for detecting affective states which have also been discussed in the literature (Mitchell & Ross, 2013).

Temporal features include intensity, which signifies the energy of the sound waves (signal) travelling through the air. It is the representation of a change in the amplitude of speech signals over time. Amplitude here in this process is the displacement of the sound waves produced, which determines the loudness of the voice (Kappas et al., 1991). These features are relatively simple to extract and are easier to interpret than spectral features. Frequency domain features are derived from the analysis of the signal in terms of frequency components, rather than its time-based waveform. These features represent how the energy of the signal is distributed across different frequency bands, capturing voice qualities and tone.

Spectral features specifically include features associated with the fundamental frequency (F_0) such as frequency components, formants and coefficients. Vibrations of the vocal cords cause F_0 , which denotes the lowest oscillation of sound waves to be produced, and it is related to the pitch of voice. F_0 is one of the most-used parameters for detecting emotions (Scherer, 1995). The pitch comprises characteristics related to rhythm and tone of the speech. The variation of the F_0 over time yields its other statistical properties and F_0 contour, which can also be used as features (Kamiloğlu et al., 2020). More specifically, formants indicate the location of resonance occurring

in any vocal tract. They are particularly described with their specific frequencies, their amplitude at the peak, and their bandwidth. The localisation of these formants' resonance on frequencies causes amplification or attenuation. The resonance of the formants regulates articulated sound quality and are assessed as the peak in the frequency spectrum of the sound (Koolagudi et al., 2012). Spectral features are obtained by converting the time-based signals into the frequency domain signals using a Fourier Transform (Kappas et al., 1991; Scherer, 1995).

One of the essential spectral features is the Mel Frequency Cepstral Coefficients (MFCC), which represents the speech signal's short-term power spectrum. It is obtained by division of the utterances into segments. Then a short-time discrete Fourier Transform is applied on each segment to convert it into the frequency domain. Next, the Mel filter bank is utilised for the calculation of the number of subband energies, which is then followed by taking logarithms and finally the application of inverse Fourier Transform obtains MFCC (Kuchibhotla et al., 2014). Other spectral features include Linear Prediction Cepstral Coefficients (LPCC), Log-Frequency Power Coefficients (LFPC) and Gammatone Frequency Cepstral Coefficients (GFCC) that can be acquired by a Linear Prediction Coefficient (LPC), by measurement of spectral band energies using Fast Fourier Transform and Gammatone filter bank respectively (Nwe et al., 2003a). An algorithm documented by Kaiser (Kaiser, 1993) as

$$\Psi [X(n)] = x^2(n) - x(n+1)x(n-1)$$

was also developed to measure the energy from speech by a nonlinear process, where $\Psi []$ is the Teager Energy Operator (TEO), and is applied on $x(n)$, which represents the sample of a speech signal. TEO has various features, such as TEO decomposed FM (frequency modulation), TEO decomposed FM variation (TEO-FM-Var), and TEO autocorrelation envelope area (TEO-Auto-Env), as well as a critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env).

These features explore the change in the characteristics of energy's airflow of speech under stress (Sun et al., 2011). These features can be particularly utilised for detection of stress.

Another set of features called prosodic features are a combination of both temporal and spectral features. For instance, intensity, pitch, F_0 , formants and duration are also categorised as prosodic features. Duration is defined as the time to construct vowels and other factors present in speech, such as words. Speech pace, the time duration of voiced, unvoiced and silenced region and longest voiced speech are some of the commonly used duration-related features. Some other properties of a voice include shimmer, jitter and harmonics-to-noise ratio (HNR). These features might also help differentiate the emotions produced in a speech signal. While jitter represents the unevenness of F_0 between successive oscillations, the shimmer is the unevenness of the amplitude. The HNR is the representation of the ratio between periodic to aperiodic components in a voiced speech signal (Li et al., 2007). In Table 2.1, we summarise common voice features and their representations, used in detection of emotion/affective state detection.

Table 2.1 *Different types of Voice Feature Representations*

Feature	Representation	Example
Temporal features (Time domain features)	Features pertaining to intensity	Minimum energy, zero-crossing rate, maximum amplitude, amplitude envelope, root mean square
Spectral features (Frequency domain features)	Represents the signal in terms of frequency	Fundamental frequency (F_0), formants, pitch, fundamental frequency contour, spectrum centroid, spectral bandwidth,
Mel Frequency Cepstral Coefficients (MFCCs)	Represents short-term power spectrum	Linear Prediction Cepstral Coefficients (LPCC), Log-Frequency Power Coefficients (LFPC) and Gammatone Frequency Cepstral Coefficients (GFCC), Linear prediction Coefficient (LPC)
Teager Energy Based features	Represents energy in signal by following a nonlinear process	TEO decomposed FM, TEO decomposed FM variation (TEO-FM-Var), and TEO autocorrelation envelope area (TEO-Auto-Env), as well as a critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env)
Prosodic Features	Combination of Temporal and Spectral features	Intensity, pitch, F_0 , formants, duration Unvoiced and silenced region
Some other voice features	Represents unevenness of F_0 between successive oscillations	Shimmer
	Represents unevenness in amplitude	Jitter
	Represents representation of the ratio between periodic to aperiodic components in a voiced speech signal	Harmonics to noise ratio (HNR)

It has been found that listeners can judge five types of emotions – anger, fear, happiness, sadness and sensitivity – with up to 70 per cent accuracy (Juslin & Laukka, 2003). Researchers (Kappas et al., 1991) suggested that high arousal emotions such as anger, fear, anxiety, happiness or surprise, result in increased intensity as well as a high mean F_0 . In comparison, low arousal

emotions, such as disgust and sadness, have the opposite effect. Kamiloğlu et al. (2020) suggest that pitch is high for emotions such as amusement, interest and relief, but moderate for savouring emotions like contentment and pleasure, and low for prosocial emotions such as admiration (Kamiloğlu et al., 2020). Low et al. (2010) used prosodic, spectral voice quality as well as TEO-based features for detecting clinical depression in adolescents. The results showed that TEO-based features, specifically TEO CB AutoEnv, outperformed all other features and all other combinations of features as well (Low et al., 2010).

Despite a range of voice features, most of the early studies in the last decade were focused on analysing the prosodic features including pitch, formant frequencies and intensity (Busso et al., 2009; Nogueiras et al., 2001; Nwe et al., 2003a; Schuller et al., 2003a). Recent literature surveys showed extensive use of spectral features and cepstral measurements (e.g., MFCCs), voice quality features including HNR, jitter, or shimmer, and Teager energy operators such as TEO-FM-Var and TEO-CB-Auto-Env (Bhavan et al., 2019; Hao et al., 2020; Semwal et al., 2017). This might be since various features perform differently in the detection of emotional/affective states alongside newly developed algorithms.

It is further evident from the literature search that each of the voice features have been utilised to capture specific characteristics related to emotions and affective states. For example, the time domain features (temporal features) such as minimum energy, zero-crossing rate, and maximum amplitude are useful for capturing the intensity and dynamic range of speech, indicating the changes in the emotional states. Spectral features, like fundamental frequency (F_0), formants and pitch, represent the signal in terms of frequency and are therefore crucial for identifying emotions that affect the pitch contours and formant structures of the voice. MFCCs provide a detailed representation of the short-term power spectrum, capturing subtle variations in speech that

are important for distinguishing between different emotional nuances. Teager Energy Based features, such as TEO decomposed FM and TEO autocorrelation envelope area, are efficient in capturing varying energy dynamics with emotional arousal levels. Prosodic features, which combine temporal and spectral elements are employed for comprehensive analysis of voice as they encompass variations in intensity, pitch, and duration, providing a complete profile of the emotional state. Additional voice features like shimmer, jitter, and HNR are utilised to measure the unevenness in frequency and amplitude, and the ratio of periodic to aperiodic components in voiced speech, respectively. These features are particularly useful for detecting emotions that induce vocal instability, such as fear or anxiety.

In case of stress during public speaking, high arousal is predominant when a person is giving a speech in front of a public audience and it will usually lead to high pitch, intensity (temporal features) and changes in the spectral properties as compared to a normal/calm state. Spectral features (such as spectral bandwidth, spectral roll off et cetera) and MFCCs could also provide detailed frequency-based representations, that are essential for detecting subtle emotional cues along with Teager Energy Based features for locating the changes in vocal energy. Therefore, based on the evaluation of features used in literature, a combination of temporal features, spectral features, MFCCs and TEO energy-based features might be able to facilitate accurate detection and classification of the affective state of stress.

Interestingly, emotion/affective state detection from voice has unique traits and also common elements shared with the analysis of emotional contents in music, as discussed in one of the earlier works by Scherer (Scherer, 1995). Many voice features have been analysed to identify various emotions, there remains a gap in understanding the real-time dynamics of vocal stress features, specifically during live public speaking scenarios. This further highlights the need for

exploration of specific stress indicative voice features in this context using algorithmic models-based voice analysis techniques. Notably, the literature emphasises the need for voice-based stress detection following approaches such as those discussed by Dillon (Dillon, 2001, 2003). The next section presents the results from Search 1, where we aimed to identify how VR has been utilised to induce, study, detect various states of anxiety and fear. Additionally, it explores how virtual audiences in VR environments were designed, programmed or trained to provide feedback or responses to a speaker's performance.

2.4 VR for Public Speaking

VR traces its roots back to 1838 with Sir Charles Wheatstone's stereoscope illustrating the concept of binocular vision, where the brain combines two images into a single three-dimensional (3D) image, creating depth perception (Bowers, 2001). This idea became the foundation of the first head-mounted display (HMD), called the Telesphere Mask, which provided stereoscopic 3D images and stereo sound. The first working example of a VR device named the Sensorama was created in 1956 by cinematographer Morton Heilig and was patented in 1962 (Gigante, 1993; Heilig, 1998). During the 1960s and 70s, VR technology advanced significantly and was adopted to very specific applications in flight simulation and NASA trainings (Fisher et al., 1987; Furness Iii). In 1990s, VR began to be commercialised and the early 2010s marked a significant leap with the release of consumer VR headsets such as the Oculus Rift and HTC Vive. In recent years, advanced devices like Apple Vision Pro have further refined and popularised VR technology (Jiang et al., 2022; Ko et al., 2021).

VR has a range of applications (Laver et al., 2015; Lindner et al., 2019; Miloff et al., 2016; Opriş et al., 2012; Parsons et al., 2008; Ticknor, 2018, 2019; Valmaggia et al., 2016). It is beneficial across areas such as prototyping and visualising within engineering, manufacturing and

construction (Bouchlaghem et al., 2005; Hilfert et al., 2016; Le et al., 2015). VR has also been used in industrial design and electronics assembly training (Gavish et al., 2015) and in developing educational tools for a variety of subjects including music, science (Dillon et al., 2006; Xie et al., 2023), computer architecture (Dascalu et al., 2017), as well as in training very specific skills such as military firing (Bhagat et al., 2016). VR applications are also explored for health care for providing exposure therapy and anxiety treatments. The applications of VR span several diverse areas as summarised in Table 2.2

Table 2.2 *Various Applications of Virtual Reality (VR)*

Category	Applications	Specific applications	References
Industrial	Architectural, Design Construction, Manufacturing	Automotive design, Customer service training	(Bouchlaghem et al., 2005; Hilfert & König, 2016; Le et al., 2015)
Educational	Learning/training	Social Skills Training, Public speaking training, Interview preparation	(El-Yamri et al., 2019; Hartanto et al., 2014)
		Training medical students in surgical procedures	(Jiang et al., 2022)
Health care	Psychotherapy	Diagnostic and treatment tool for anxiety, obsessive compulsive disorder, Autism spectrum disorder, Post traumatic disorder	(Carl et al., 2019; Dechant et al., 2017; Gerardi et al., 2010; Kim et al., 2009; Maskey et al., 2014; Motraghi et al., 2014)
	Physiotherapy	Motor Learning, Rehabilitation	(Laver et al., 2015; Molina et al., 2014)
General	Overcoming fears Relaxation therapies	Fear of heights, Fear of spiders, social phobia	(Garcia-Palacios et al., 2002; Gebara et al., 2016; Levy et al., 2016)
	Entertainment	Games for various purposes	(Miloff et al., 2016; Ordaz et al., 2015)

2.4.1 The Use of VR for Public Speaking

Over decades, public speaking tasks have been used as stressors in experimental and laboratory settings (Feldman et al., 2004; Jezova et al., 2016) to assess and study the effects and issues that arise during speech activities (highlighted in [Chapter 1 Section 1.2.](#) and [Chapter 2 Section 2.2.2.1](#)). These issues like psychosocial stress are generally prevalent worldwide among speakers and can significantly hinder the quality of the speech delivered (Koroleva et al., 2014; Ruscio et al., 2008).

Researchers have explored the use of VR to train public speaking skills and address these common issues, particularly, in the past 5-7 years with the advancements and availability of a variety of HMDs. VR has proven to be an effective tool in enhancing interpersonal and public speaking skills (Kimani et al., 2019; Lister et al., 2010; North et al., 1998; Schmid Mast et al., 2018; Takac et al., 2019). Although in the past, public speaking was simply used as a stressor in experiments and laboratory settings, the advantage of using VR for public speaking lies in its ability to create controlled, repeatable and safe environments, which could be tailored to specific situations and stressful scenarios, that may be difficult or impossible to replicate in real-life (Azwar et al., 2016). Another crucial benefit of VR lies in its ability to provide personalised feedback and training options. This allows individuals to have access to consistent practicing environments, not just to evaluate their stress, but to also help them develop specific skills and to aid in overcoming the issues associated with the activity and mastering the skill (Kassem et al., 2017; Pedro et al., 2016).

For instance, for public speaking, VR can offer multiple interactive and immersive simulations of a classroom, meeting room, auditorium, or assembly, that might be something not always readily available in real-life conditions. In addition, many customised virtual audiences in simulated scenes for practice of speaking abilities can be incorporated in the VR scenarios. Virtual

audiences can be made interactive, according to specific controlled and repeatable scenarios where different training modes can be precisely created. Personalised feedback from a virtual audience can be provided to individuals based on their level of competency, enabling them to practice managing stress in diverse situations such as classroom settings, meeting rooms and auditorium scenes, wherever practice is needed (Soyka et al., 2016).

Research indicates there has been a significant interest in leveraging VR to explore feedback strategies in a virtual environment for public speaking (Brundage et al., 2015; Chollet et al., 2015). Several studies have also suggested that interactions in a VR simulation may be able to elicit both social stress and social support with virtual audiences and their behaviours (Kothgassner et al., 2016; Kothgassner et al., 2019; Kotlyar et al., 2008; Pan et al., 2012). While interactions with audience in real-world settings can provide similar experiences, VR can offer valuable training opportunities tailored to individuals' subjective needs for managing stress levels all at one place by: (i) simulating a wide range of environments, (ii) simulating interactive virtual audiences, (iii) ability to serve as a personalised tool for user by making easier to learn and managing social stressors systematically, and (iv) providing consistent feedback and support strategies. This allows individuals to learn how to overcome stress and develop effective speaking skills.

2.4.2 VR for Assessing, Training and Evaluating Performance to Improving Public Speaking

The following section is dedicated to covering several studies that have utilised VR for assessing skills, evaluating performance, treating speaking issues and improving public speaking. In one of the earliest studies, research (e.g., (Slater et al., 1999)) explored the psychotherapeutic effects of a virtual environment on social phobia for public speaking. Slater et al. (1999) conducted an experimental study using virtual public speaking scenarios with avatars. The avatars exhibited different random behaviours such as nodding, twitching, blinking, yawning, clapping and showing

hostile reactions. Ten graduate students from University College London (UCL) were recruited to be part of the research experiment. The participants either gave their speech to the audience on the monitor or were entirely 'immersed' in a VR setting with a headset. Each participant repeated their speech three times under three different conditions. At the first time, the participants faced either a hostile or a friendly audience. At the second time, participants faced whichever audience they did not face during the first speech. At the third time, participants faced an audience with hostile reactions, which eventually become friendly. Participants using the VR headset reported a strong sense of co-presence with the hostile audience, meaning they felt as though they were physically present with the listeners. However, despite this sense of co-presence, participants did not feel that the audience was genuinely interested in their speech. The results suggested that while the immersive VR environment amplified the perception of the situation, the lack of audience interest negatively impacted the participants' self-ratings and increased their anxiety levels. Even though this study was one of its kind, it had limitations including a very small sample size, lacked diversity in sample data as all the students were graduates from same university and there was limited information given on training of the virtual audience. Only psychological measurements of self-perceived rating of fear were utilised and no measure of any objective data was utilised such as physiological parameters and voice. Furthermore, the avatars' behaviours were random and not tailored to individual speakers, potentially affecting the consistency and accuracy of the induced anxiety. These factors limit the generalisability and applicability of the study's findings (Slater et al., 1999).

A follow-up study by Slater in 2001 was carried out and involved 40 subjects. The results indicated that a hostile, negative audience scenario generated a strong effect in speakers and social anxiety was indeed induced by the virtual audience. The study further indicated that induced

anxiety is directly related to the virtual audience's feedback and the level of anxiety is related to the type of feedback received from the virtual audience. Notably, a hostile and negative audience scenario elicited strong emotional responses in speakers, regardless of their predisposition to fear of public speaking. Findings were coherent with previous research suggesting that the potential efficacy of VR exposure in treating social phobia and highlighting the importance of including different virtual audience interventions (Pertaub et al., 2001).

Hartanto et al. (2014) investigated the use of VR exposure therapy for treating anxiety disorders by controlling social stress among 44 participants. Their study involved one-on-one dialogue sessions between participants and a virtual audience. Participants were exposed to three different virtual scenarios: a neutral scene, a blind date and a job interview. Additionally, the study found that positive dialogue feedback from the virtual character resulted in lower self-reported anxiety, a lower heart rate and longer answers from participants. Conversely, negative dialogue feedback led to higher self-reported anxiety, an increased heart rate and shorter responses. These findings further indicate that VR can effectively simulate different social scenarios and provide valuable feedback, making it a powerful tool for anxiety treatment and social skills training. The results showed that participants' anxiety levels significantly increased as the social scenarios changed with the changing dialogues, demonstrating VR's effectiveness in inducing anxiety and providing a controlled environment for therapeutic interventions. However, there are some drawbacks to the study. Firstly, the sample size of 44 participants may not be large enough to generalise the findings to a broader population specifically in the context of dialogue exchanged in various conditions. The virtual scenarios used in the study might be able to deduce a basic effect of different scenarios and dialogues; however, the study did not fully capture the complexity and variability of real-life social interactions where the larger audience induces negative emotions. The

training of the virtual audience involved programming specific behaviours and responses to simulate realistic social interactions and create varying levels of social stress. This approach allowed the researchers to systematically control the social environment and assess the effectiveness of VR exposure therapy in treating anxiety disorders. However, the study did not provide detailed information on the specific algorithms or programming techniques used to train the virtual audience for their behaviours and responses. Lastly, the aspect of analysing voice to determine the affective states experienced by the speaker participants was also not utilised (Hartanto et al., 2014).

Brundage and Hancock (2015) researched the use of VR exposure therapy for treating people who stutter. In their study, 10 participants first delivered a speech to a real audience and then, on another day, performed in front of two virtual audience members. The results from the study indicated that the participants had an experience in the virtual setting which was similar to their experience with the real audience (Brundage & Hancock, 2015); suggesting that VR can be used as a standardised tool for inducing stress (Jönsson et al., 2010) and fear (Powers et al., 2013) in real-life scenarios. Notably, these studies highlight how people (or virtual audiences) around us, even in a virtual setting simulation, can influence our feelings, behaviour and performance.

In a study, Chollet et al., (2015) explored various strategies to improve public speaking using an interactive virtual audience framework. The framework included three conditions: a noninteractive virtual audience (control), direct visual feedback and nonverbal feedback from an interactive virtual audience. Self-assessment questionnaires, expert assessments and objective measures of eye contact and avoidance of pause fillers were evaluation measures. The results indicated that the interactive virtual audience condition provided the most engagement and challenge, leading to improved public speaking skills as judged by experts (Chollet; et al., 2015).

However, the study did not include voice analysis in its evaluation metrics, which represents a notable gap. While pause fillers are a critical component of assessing speech performance under stress, other vocal attributes, such as tone, pitch, and speech rate, are also important indicators of public speaking efficacy and stress levels. The absence of voice analysis techniques indicates that the study may have missed crucial data points that contribute to a comprehensive understanding of a speaker's performance levels.

Wörtwein and colleagues (2015) evaluated a novel multimodal framework for public speaking training that used automatic assessment of nonverbal behaviour and multimodal modelling with 47 participants. The framework examined various nonverbal behaviours, including audio, visual and physiological cues, to provide a comprehensive assessment based on correlations with expert judges' opinions of key performance aspects. The chosen approach allowed speakers to practice and improve their public speaking skills in a controlled, repeatable environment tailored to individual needs, proving to be a valuable tool for public speaking training. The study results highlighted that pauses in speech and voice variability were correlated with performance improvement, indicating that a regular flow of speech is a key feature of enhanced public speaking performance. However, the recognition of emotions and affective state does not solely rely on pauses during a speech. It also noted that a regular vocal flow does not necessarily mean there is no underlying stress affecting the speaker's voice and overall speech delivery (Wörtwein et al., 2015).

Over the years, research on using VR to improve public speaking has gained momentum. The focus has expanded to include various psychological parameters, objective measures of physiology, the effects of virtual audiences and their different behaviours, assessment of speaking skills with expert feedback, and more comprehensive training environments. Additionally, models

dedicated to improving public speaking skills through voice training, alongside psychological and objective measurements, have also come into play and could lead to significant improvements in skills training.

There were also studies that have made further developments in the use of VR for public speaking, particularly in terms of voice and using it to design or simulate the virtual characters in context of public speaking. One of the studies presented a method for generating 3D virtual characters' performances from the audio signals by analysing acoustic and semantic properties of utterances. This involved prosodic analysis to assess stress and pitch, related to the spoken words to identify the speaker's agitation state. The information then generates corresponding facial expressions and behaviours, such as head movements, eye saccades, gestures, blinks and gazes. The novel method involves voice analysis to gesture animations by considering coarticulation with other closely timed animations. The study indicates that incorporating semantics along with prosody results in more appropriate virtual character performances compared with methods relying solely on prosody, which could serve as a more realistic base for improving the virtual audience's presence in the VR applications, particularly for public speaking (Marsella et al., 2013).

A group of researchers outlined a prototype of VR tool for public speaking training. The innovative system includes a dictation recogniser system, which accepts an audio signal's transcription as the input, and users engage with the training simulator by immersing themselves in a virtual auditorium setting. Based on the audio signal received, the system generates an animated performance as output. However, the speech text conversion might be affected by different accents and speech patterns. The performance metric in the end may also not capture the nonverbal cues and nuances of audience engagement (Pellett & Zaidi, 2019).

A distinct innovative multimodal presentation trainer was developed to support the practice of public speaking skills by giving the user real-time feedback about different aspects of the participant's nonverbal communication. It tracks the participant's voice and body to interpret performance and present an intervention to give feedback to the user. However, while the tool showed promise in enhancing learning, potential drawbacks may include challenges in effectively addressing individual user needs and preferences. While there is a possibility of overreliance on technology and potentially neglecting the development of fundamental speaking skills, still the user experience (UX) evaluation of participants using the presentation trainer for practicing elevator pitches showed it had significant influence of the provided feedback on learning outcomes (Schneider et al., 2015).

A separate group of researchers developed an interface using Google Glass to help people improve public speaking by automatically detecting the speaker's volume and speaking rate in real time. Based on the speaker's volume and speaking rate, the system provides immediate feedback during the speech engagements. It utilises advanced technology to analyse the speaker's performance in real-time, focusing on various aspects such as speech content, delivery and audience engagement. Rhema employs ML algorithms to interpret the speaker's speech patterns, gestures and facial expressions to provide personalised suggestions and guidance. The development of Rhema represents a significant advancement in leveraging intelligent interfaces to support individuals in mastering the art of public speaking (Tanveer et al., 2015).

Similarly, El-Yamri et al. (2019) developed a VR system incorporating a virtual audience character model. This model aimed to elicit emotions, such as stress, in speakers, which were then analysed using a third-party emotion detection application. However, the authors relied solely on their common sense to weigh the detected emotional states, including stress and happiness. This

approach may lack accuracy due to the absence of clear and comprehensive information guiding the basis of their common sense judgements. Furthermore, the model lacks a learning process for the virtual agent, which could incorporate analysis of voice features and physiological parameters. For instance, real-time emotion detection using VR and integration with existing emotion detection models could enhance the accuracy of emotional analysis. This approach would enable the system to provide real-time feedback to speakers based on a comprehensive evaluation, potentially improving the effectiveness of public speaking training (El-Yamri et al., 2019). Table 2.3. shows some of the relevant studies that utilised VR for public speaking.

Table 2.3 *Relevant Studies of Public Speaking in VR*

Article	Main Public Speaking issue focused on	Task	Responses of virtual audience
(Brundage & Hancock, 2015)	Affect generation	Speech	Neutral and challenging virtual audience
(Chollet, Wörtwein, et al., 2015)	Anxiety	Speech	Nonverbal indirect feedback
(El-Yamri et al., 2019)	Fear/stress	Speech	Real-time feedback voice tone, gaze and speech content
(Kothgassner et al., 2016; Kothgassner et al., 2019)	Stress	Speech	Real audience, Virtual audience, Empty virtual lecture hall
(Pertaub et al., 2001)	Fear	Speech	Positive, negative comments and behaviour
(Slater et al., 2006; Slater et al., 1999)	Anxiety	Speech	Positive and negative evaluation feedback
(Hartanto et al., 2014)	Social Anxiety/ Stress	One-to-one conversation session	Neutral and conversing audience

The studies collectively highlight the potential of VR as a valuable training tool for addressing and learning to overcome issues associated with public speaking and enhancing skills. The research studies on VR for public speaking show that virtual environments can effectively simulate various scenarios for giving speeches, inducing anxiety, fear and stress which are common issues faced by individuals in daily life. These VR environments could also be used as a tool to provide controlled settings as an intervention to possibly help people enhance public speaking skills. Furthermore, it could be seen from the various results that generally participants responded to the virtual and fake audiences similarly to real ones, indicating the feasibility offered by VR to be a realistic and impactful training tool.

However, these studies also reveal common limitations, such as small sample sizes, and homogeneous groups of participants (Brundage & Hancock, 2015), which limits the applicability of results for a wider demographic. While several studies relied primarily on the psychological data (Slater et al., 1999) and physiological data, which is of course important, fewer integrated voice analysis into VR environments for a comprehensive understanding of affective and emotional states, especially for real-time stress detection.

A notable gap in the current literature is a lack of clarity in terms of how the affective states of the speakers were determined in real time during public speaking tasks in VR. Changes in the speaker's affective state due to the virtual audience's feedback is or was solely understood based only on either self-reported questionnaires or physiological data. An additional way to determine speaker's affective state is to assess emotions in real time by analysing the voice, as it offers a valuable dimension for affective state detection.

After reviewing the literature on the use of immersive technologies, such as VR within the context of public speaking and the study of various affective states and emotions, the focus now

shifts to the results of Search 2, which are discussed in the next section. Search 2 specifically examines voice-based affective state and emotion recognition algorithmic models, aiming to deepen our understanding of existing approaches and methodologies while identifying nuanced gaps in the research.

2.5 Algorithmic Models for Detecting Affective States

Search 2 aimed to review past research on ML algorithmic methods and approaches used to detect affective states. Among the various ML models, the Gaussian Mixture Model (GMM) is one of the well-known unsupervised ML model for emotion detection (Anagnostopoulos et al., 2015; Busso et al., 2009; Tashev et al., 2017; Truong et al., 2007). GMMs are efficient at modelling multimodal distributions, enabling them to represent diverse sets of input data, by finding probabilistic cluster assignments (Douglas-Cowie et al., 2007). Using the expectation – maximisation [E-M] approach, GMMs start off by setting cluster locations and shapes. It repeats the E-M steps until each cluster is converged with a smooth Gaussian model. These models can vary in form, ranging from being stretched into circular clusters (Figure 2.1) to elongated and oblong clusters (Figure 2.2) (VanderPlas, 2016).

Figure 2.1 *Circular stretched out clusters using GMM*

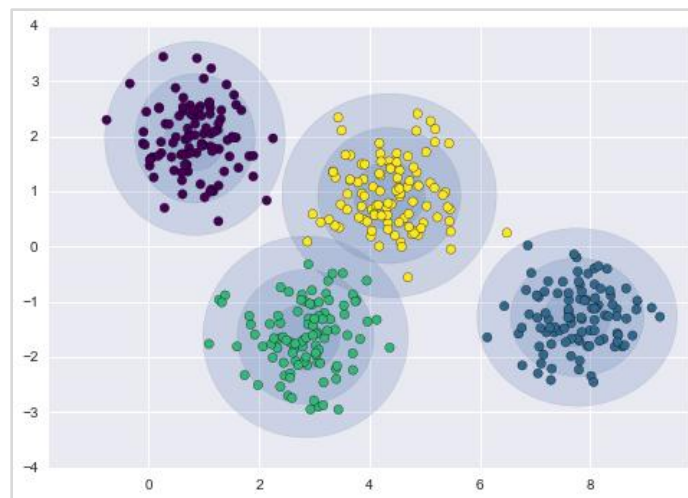
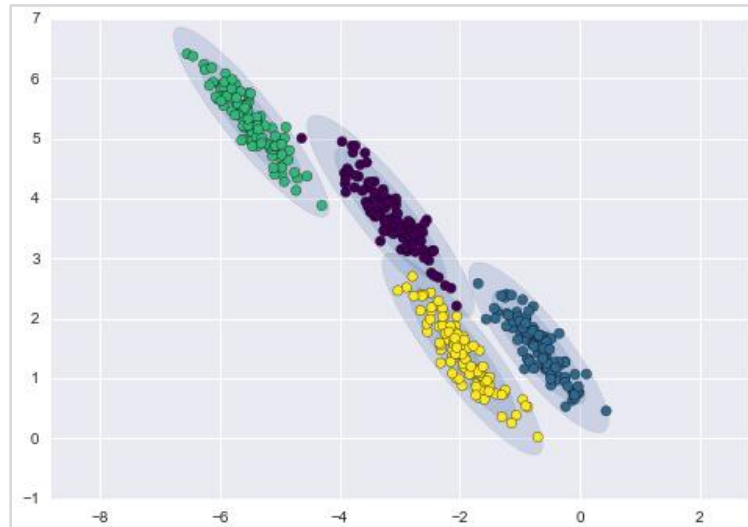


Figure 2.2 *Oblong Stretched Out Clusters using GMM*

Researchers developed and tested the various GMMs with spectral features such as MFCC, MFCC-low and pitch. Neiberg and colleagues (2006) combined three different models for recognising emotions from such features. The developed models were tested on voice controlled telephonic service recordings recorded by Swedish company Voice Provider (VP) and on the meeting recordings from the ISL Meeting corpus. The results showed that the combination of developed GMMs based on the spectral features obtained the best results (Neiberg et al., 2006).

It is important to note that no cluster shape is inherently superior. The effectiveness of GMM clusters and their shapes depends on the specific characteristics of the data and the research objectives. Despite the effectiveness of GMMs with their ability to (i) model the probability density functions and (ii) to be trained relatively quickly (training and testing requirements are less), GMMs are generally considered static models. This implies that once a GMM is trained on a dataset, its parameters (such as the means, variances, and weights of the Gaussian components) remain fixed and do not change over time or with new data. GMMs model the probability distribution of the data using a mixture of Gaussian distributions, and they are typically used for

tasks such as clustering, density estimation, and pattern recognition (Bishop, 2006). Additionally, their performance is said to be reduced in the presence of noise from environmental factors (Chan et al., 1994; Naylor et al., 2006), therefore making it not suitable for model development during real-world scenarios of public speaking.

Some researchers explore and compare combinations of GMMs with other more advanced models such as Hidden Markov Models (HMMs). For example, Schuller et al. (Schuller et al., 2003a) compared HMM with GMM. HMM classified the utterances using low-level statistical features that represent a stationary state of emotions in voice. GMM was used to classify the utterances using global statistics of raw pitch and energy contour. The results indicated that the average accuracy in identifying discrete emotions was more than 86 per cent. Another HMM-based model is shown to achieve a 70 per cent recognition rate for emotions such as happiness, anger, joy, fear, disgust, and sadness by using low-level pitch and energy features along with their contours (Nogueiras et al., 2001). Nwe et al. (2003b) demonstrated that the HMM yields better performance by achieving the best recognition rate of 89 per cent by using spectral features such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) (Nwe et al., 2003a).

Each emotion in a HMM can be modelled by a single-state HMM (Casale et al., 2007). This single-state HMM is trained by increasing the margin between emotions, using a loss function to scale this margin. Unlike GMMs, the HMMs are based on the Markov chain model, which indicates that the future state of any pattern is determined by its current state and not by the past state.

Markov Assumption:

$$P(q_i = a | q_1 \dots q_{i-1}) = P(q_i = a | q_{i-1})$$

Where $q_1 \dots q_i$ are the random i states (Martin., 2021).

In a HMM, various states are hidden from the observer. These hidden states represent the temporal state of the voice data, created as the observation sequence randomly associates with each state (Li et al., 2007). HMMs can have several advantages for (i) capturing temporal dynamics regarding any affective state such as of stress, (ii) capturing the underlying states corresponding to different levels of an affective state (Busso et al., 2008), (iii) robustness to noise in real-world environments (Abe et al., 2015), and (iv) the integration of many features. However, there are several drawbacks to using HMMs. Because they require extensive training data (Young et al., 2002), criteria to determine an optimal number of hidden states remains a challenge (Rabiner, 1989) and generalisation to diverse speakers can be tricky (Deng, 2006).

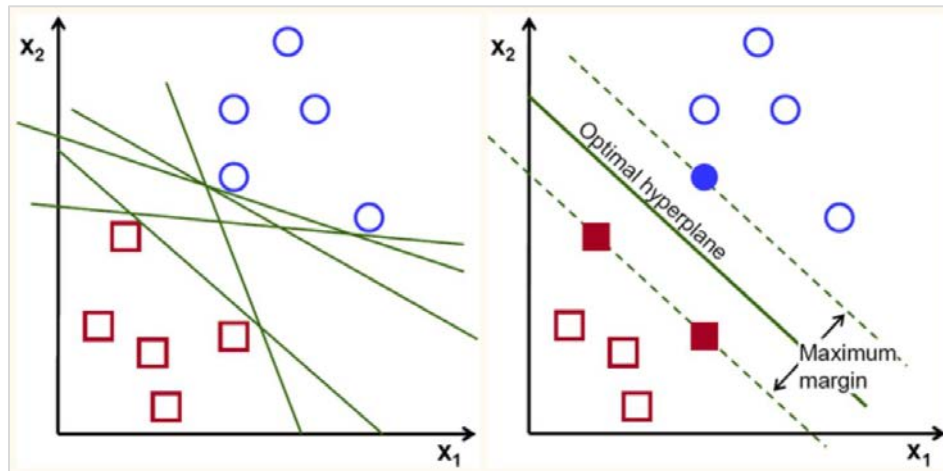
In a separate study, researchers tested a supervised ML algorithm Support Vector Machine (SVM) and statistical model of HMMs to classify five emotions: anger, happiness, sadness, surprise and neutral emotion. The results from the study showed a 99.5 per cent accuracy rate for HMM, and SVM obtained an accuracy rate of 88.9 per cent in recognising the emotions (Lin & Wei, 2005).

The SVM models have generally gained popularity, in contrast to GMM and HMM (Bhavan et al., 2019; Bitouk et al., 2010; Hao et al., 2020; Jain et al., 2020; Semwal et al., 2017; Truong & van Leeuwen, 2007; Wang et al., 2015). SVM shows promise in various research studies for stress classification, distinguishing between stressed and unstressed states (Barreto et al., 2007). Similarly, in the study conducted by Hu et al. (Hu et al., 2007) the classification of emotions using spectral features reported an accuracy rate of 82.5 per cent with SVM, while an accuracy rate of 77.9 per cent was obtained by only using GMM.

Bhavan and colleagues (Bhavan et al., 2019) further utilised a bagged ensemble model consisting of SVM with a Gaussian kernel. The results showed an improved accuracy rate of 92.45 per cent on the Berlin Database of Emotional Speech (EmoDB) (Burkhardt et al., 2005), 75.69 per cent on the Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone et al., 2018), and 84.11 per cent on the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) (Koolagudi et al., 2009) respectively. A combination of spectral features was extracted, further processed and reduced to the required features set including MFCC, spectral centroids and MFCC derivatives. The recognised emotions included happiness, fear, sadness, anger, disgust, boredom, calm, surprise and sarcasm. Similarly, other researchers (Truong & van Leeuwen, 2007) also employed SVM for emotion recognition and discovered that SVM performed better when using acoustic and linguistic features, resulting in an 8 per cent increase in emotion recognition accuracy.

SVM functions by identifying the best hyperplane for linearly separable patterns within an N-dimensional space, where N represents the number of features. This hyperplane acts as a decision boundary, effectively separating data into distinct classes. Figure 2.3 shows some of the possible hyperplanes for given data points (Gandhi, 2018).

Figure 2.3 Possible Hyperplanes

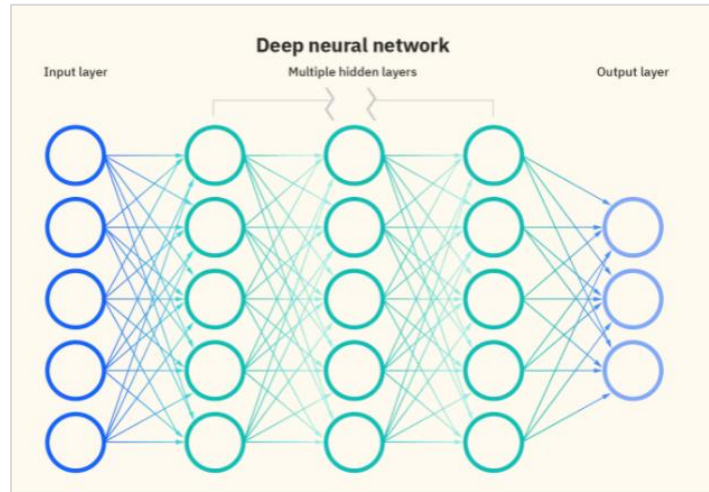


In case it is not possible to separate the patterns linearly, a kernel function is utilised by SVM to map the original data to a new space. However, since there is no standardised method for selection of kernels, ensuring the separation of transferred features is not guaranteed. Despite this, SVM offers some distinct advantages over the statistical models such as HMM, or quite similar model of K-Nearest Neighbour (KNNs), and GMM, including effectiveness in handling high dimensional features (such as in frequency components, cepstral coefficients), robustness to overfitting, versatility in handling linear and nonlinear data, global optimality of training algorithm and memory efficiency, as it utilises a subset of training points in the decision-making function (Scikit Learn, 2022). Therefore, when researching voice analysis in the context of public speaking, considering SVMs along with one of the basic probability density function algorithms like KNN could provide valuable options for training data.

In past two decades, several new computational models are reported in the literature for classification of emotions/affective states from audio involving ML techniques (Palo et al., 2015; Soltani et al., 2007). ML is a broader field that encompasses various algorithms and models for building systems that can learn and improve from experience without being explicitly programmed. One of the subset of ML is Neural Network or Artificial Neural Network (ANN) comprising an input layer, one or more hidden layers and an output layer, as shown in Figure 2.4 (Shubham Kumar, 2021). ANN along with deep learning (another subset of ML), are part of AI, which are both emerging fields for recognising affective states. Although ANN lacks rules for the optimal setting of ANN topology, and learning and training, making the classification procedure difficult, it still offers several advantages with their ability to form intricate patterns from complex and nonlinear data

(Goodfellow et al., 2016), versatility (Schmidhuber, 2015), and scalability (Krizhevsky et al., 2012).

Figure 2.4 *Neural Network or Artificial Neural Network (ANN) with Layers*



ANN, DL consists of nodes and each layer is connected to the next layer. The input layer receives the data, which is then passed onto the hidden layers where features are extracted and final layers produce the classification result (Kumar et al., 2023; Liu et al., 2020). Once the training of the ANN is completed with raw data, the ANN can be used to classify new data. Each node operates based on its own linear regression model, which is composed of input data, weights and a threshold (bias) and an output. ANN encompasses different architectures aimed in respect to data, task and applications. Three main types of ANN are Feedforward Neural Networks (FNNs), Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs) as summarised in Table 2.4.

Table 2.4 *Types of Artificial Neural Network Architecture*

Neural Network Type	Description	Application Areas	Advantages	Challenges	References
Feedforward Neural Networks (FNNs)	Simple architecture, one directional flow of data from input to output layers	Classification, regression, pattern recognition	Simple architecture, fast training	Limitations in capturing complex patterns	(Bengio, 2009; LeCun et al., 2015)
RNNs	Handles sequential data	Natural language processing, time series prediction, sequence classification, video analysis	Captures temporal dependencies, Maintains memory of previous inputs	Prone to gradient explosions, poses challenges in effective training	(Bengio et al., 1994; Pascanu et al., 2013)
DNNs	Multiple hidden layers	Image recognition, speech recognition, natural language processing	Deep architecture, capable of learning intricate patterns	Requires computational resources and data	(Goodfellow et al., 2016; Krizhevsky et al., 2012)

Table continues on next page...

Neural Network Type	Description	Application Areas	Advantages	Challenges	References
Fully Focused Time Delay Neural Network (FFTDNN)	Variation of TDNN with tapped delay lines at input and layer connections	Speech recognition, recognition of isolated words, time series classification	Temporal dependencies effectively handling, shift-invariant classification	Handling long term dependencies, vanishing gradient during training	(Peddinti et al., 2015)
Radial Basis Function Neural Networks (RBFNNs)	Use radial basis functions in hidden layer	Function approximation, pattern recognition, time series prediction, control systems	Intrinsically nonlinear, universal approximators, rapid convergence, geometric interpretation	Selection of optimal network parameters, scalability issues, sensitivity to parameter settings, potential for local minima and limited generalisation	(Montazer et al., 2018)
CNNs	Specialised for processing grid-like data, such as images	Image classification, object detection, segmentation	Automatically learn features, handle spatial relationships	Limitations in processing temporal or sequential data	(Alzubaidi et al., 2021)

Researchers have extensively explored these various architectures of ANNs. Agarwalla and Sarma (Agarwalla et al., 2016) used Feedforward Network (FFN) and DNN approaches to extract relevant data samples from big data space and used them for automatic speech recognition using soft computing techniques, particularly focusing on Assamese speech with dialectal variations. The multilayer perceptron (MLP) (which is a Feedforward ANN), was configured with inputs to learn emotion classification information using clustering and manual labelling. The extracted features comprised spectral and prosodic features. The obtained features were applied to RNN and FFTDNN for the evaluation of their performance to recognise mood, dialect, speaker and gender variations in Assamese speech. Agarwalla and Sarma tested these models by considering the recognition rates which were obtained by using confusion matrices and manually calculating time. The results showed that their proposed ML based sentence extraction techniques and the composite feature set when used with RNN as a classifier outperformed all other approaches.

The RNNs are commonly used deep learning algorithms specialised for sequential data processing with short-term internal memory. Unlike other ANNs, RNNs learn not only from training data but also from prior inputs to improve current outputs. Additionally, RNNs share parameters across each network layer and use the same weight within each layer (IBM, 2020). Despite their unique features, RNNs are prone to gradient explosion and effective training issues. Conversely, FFTDNNs are multilayer ANNs that classify patterns based on shift invariance without explicit segmentation before classification. Aside from detecting various emotions, FFTDNNs are useful for detecting stress due to their ability to handle temporal dependencies and focus on relevant parts of the input (Li et al., 2020) and their preprocessing, feature extraction and

training processes can be computationally intensive. Furthermore, real-time processing in a given context such as public speaking may require optimisation for efficient deployment.

Various deep learning methods contribute to emotion recognition from speech or song. Deep learning involves the creation of algorithms with multiple layers, working similarly to the human brain process of sensing (information travels within many neurons before making a decision). Deep Neural Networks (DNNs) are the hidden networks of algorithms, as seen in the ANN's middle layers in Figure 2.4. The input data pass through different layers of DNNs. DNNs then form high-level invariant appropriate features called identifiers from raw input data and proceed to classify audio and voice data for detecting emotions (Bengio et al., 2013). Therefore, DNNs perform very well in ML tasks such as speech recognition (Dahl et al., 2011). In recent years, DNNs have outperformed traditional ML algorithms, with features automatically selected using deep learning algorithms.

Ooi et al. (2014) used prosodic and spectral features to recognise six different emotions: anger, happiness, sadness, disgust, fear and surprise. They evaluated the performance of the proposed architecture, including two main paths, on the eNTER-FACE'05 (Martin et al., 2006) and RML database (Ryerson Multimedia Research Lab., 2017). Their findings revealed that the Radial Basis Functions Neural Network (RBFNN) was more accurate in recognising emotions compared to RNNs. RBFNN transforms the input signal into a different form that can be fed into the single hidden layer (referred to here as the feature vector) to achieve linear separability (Ooi et al., 2014). RBFNNs offer several advantages for voice-based emotion recognition, such as universal approximation capability, simple training processes, good performance with small datasets, and robustness to noise. However, they also come with disadvantages including scalability issues, sensitivity to parameter settings, potential for local minima and limited

generalisation. These factors must be carefully considered when choosing RBFNNs for voice-based emotion recognition tasks.

Hao et al. (2020) used an ensemble visual-audio emotion recognition algorithm to identify different emotions. They built a multitask model (comprising four submodels) based on CNN networks, another type of DNN. The authors conducted two experiments using the eNTERFACE database, demonstrating that multitask CNNs recognised emotions 3 per cent better than the CNN model in speaker-independent experiments and an average of 2 per cent better in speaker-dependent experiments. The emotion recognition accuracy rate also increased to 81.36 per cent in speaker-independent and 78.42 per cent in speaker-dependent experiments (Hao et al., 2020).

Issa et al. (Issa et al., 2020) used one-dimensional deep CNN. The input features included MFCC, Mel-scale spectrogram and spectral contrast features. The features were extracted using samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018), Berlin (EMO-DB) (Burkhardt et al., 2005), and Interactive emotional dyadic motion capture (IEMOCAP) datasets (Busso et al., 2008). They utilised an incremental process to modify initial models to improve classification accuracy. The proposed framework obtained an accuracy rate of 71.61 per cent for RAVDESS, 86.1 per cent for EMO-DB, 95.71 per cent for EMO-DB and 64.3 per cent for IE-MOCAP in speaker-independent audio classification tasks.

Despite the challenges associated with each of the architectures employed for ML, they remain promising for voice analysis. We summarise the most common algorithms used by researchers for detecting emotions in Table 2.5. Much of the research related to emotion recognition and affective state detection has emphasised the detection of emotions: disgust, sadness, happiness, fear, anger and surprise. It is evident from the Table 2.5 summary that a vast

range of corpora, such as Speech Under Simulated and Actual Stress (SUSAS) (Hansen et al., 1997), Emo-DB (Burkhardt et al., 2005) and Interactive emotional dyadic motion capture database (IEMOCAP) (Busso et al., 2008) and Simulated Telugu speech Corpus (IITKGP-SEHSC) (Koolagudi et al., 2009) are commonly utilised databases for research work.

Table 2.5 *Various Models for Affective State Detection from Human Voice*

Algorithms	Features Used	Affective States and Emotions Detected*	Performance	Reference
GMM	Fundamental frequency contours, pitch contours	An, Hp, Sd, Bo, Ds, Fr, Pc, El, Sh, Pr, De, In, Co	Over 77% accuracy on three different databases collected at various universities	(Busso et al., 2009)
	Spectral features	An, Fr, Ne, Hp, Sd	82.5% accuracy on the dataset collected by hiring participants	(Hu et al., 2007)
HMM	MFCC, Pitch, TEO-CB-AUTO-ENV, 16-GA feature, 48-GA feature	Ne, An, Lm, Lo	Increased performance of emotion detection using the 16-GA features on SUSAS database	(Casale et al., 2007)
	Pitch, log energy, formant, MFCCs,	Ne, An, St, Lm, Lo	SUSAS using GSVM 90% and 92% for neutral and stress speech, 96.3% of recognition rate obtained by using HMM while 70% was obtained for 4-class style classification	(Kwon et al., 2003)
	Prosodic and spectral features	Sr, Jo, An, Fr, Ds, Sd, Ne	70% accuracy on Spanish corpus of INTERFACE	(Nogueiras et al., 2006)

Emotional speech Synthesis Database (Barra Chicote et al. 2008)			
	Log frequency power coefficients (LFPC)	Lo, and Lu, St Ne	For Burmese:78.5% accuracy For the Mandarin utterance: 75.7% accuracy on the constructed database (Nwe et al., 2003a)
SVM	MFCC, LPCC, TEO-AutoCor	An, Ex, Fu, Hp, Ne, Sd	SVM 90.12% on EMO-DB, 83.2% on IEMOCAP (Bandela et al., 2019)
			k-NN 89.3% on EMO-DB, 78% on IEMOCAP
	Energy, pitch, MFCC coefficients, LPCC coefficients and speaker rate	An, Fr, Hp, Sd	90.08% accuracy with Linguistic data consortium (LDC) datasets and of 65.97% with University of Georgia (UGA) datasets (Jain et al., 2020)
			MFCCs features gave an accuracy of 85.085% in comparison to LPCC features accuracy of 73.125 %
	Fundamental frequency, formant, short-term energy, loudness and (MFCCs) Line Spectral Pairs (LPSs)	Sr, An, Fr, Ds, Hp, Sd, Ne	Accuracy reaches 81.36% and 78.42% in speaker-independent and speaker-dependent experiments conducted on eNTERFACE database [Martin, Kotsia, Macq and Pitas 2006] (Hao et al., 2020)

	MFCCs, Spectral centroids, and MFCC derivatives	Sr, An, Fr, Ds, Hp, Sd, Ne, Sr, Bo, Ne, Ca	An accuracy of: 92.45% on EmoDB 75.69% on RAVDESS and 84.11% on IITKGP-SEHSC	(Bhavan et al., 2019)
	Spectral features Prosodic features	An, Fr, Ds, Hp, Sd, Ne	46.1% on LDC 81.3% for Berlin database of German emotional speech	(Bitouk et al., 2010)
	Global and local Prosodic features	An, Fr, Ds, Hp, Sd, Sc, Sr	Local prosodic features performed better compared to the global prosodic features on IITKGP-SEHSC	(Rao et al., 2013)
	MFCC	St	66.4% on manually collected data	(Han et al., 2018)
	Prosodic and spectral features	An, Hp, Sd, Ds, Fr, Su	68.57 by using the RBF neural network	(Ooi et al., 2014)
DNN	MFCC feature, pitch related features,	Ex, Hp, Ne, Sd	54.3% average recognition rate	(Han et al., 2014)
k-NN	Pitch, spectral features, zero-crossing rate (ZCR), Intensity, MFCCs	Hp, An, Sd, Fr and Ne	66.24% average recognition rate on Chinese emotional dataset	(Rong et al., 2009)

	Spectrograms	An, Hp, Sd, Bo, Ds, Fr, Ne	80.79 % on Emo-DB and Korean speech Database	(Badshah et al., 2019)
CNN rectangular kernels	Chroma gram, spectral contrast features, Mel-scale spectrogram, Tonnetz representation	An, Ca, Ds, Fr, Hp, Ne, Sd, Su	71.61% for RAVDESS 86.1% on EMO-DB 95.71% for EMO-DB and 64.3% for IEMOCAP with four classes	(Issa et al., 2020)
Hierarchical classifiers	Spectral, prosodic features, mean of the log spectrum	Jo, An, Fr, Ds, Bo, Ne, Sd	An average 71.5% recognition rate on Berlin Emo-DB	(Albornoz et al., 2011)

Note: *Anger=An, Happiness=Hp, Sadness=Sd, Boredom=Bo, Disgust=Ds, Fear=Fr, Panic=Pc, Elation=El, Shame=Sh, Pride=Pr, Despair=De, Interest=In, Contempt=Co, Neutral = Ne, Excitement = Ex, Frustration = Fu, Lombard=Lo, Loud= Lu, Surprise =Sr, Joy= Jo, Sarcastic=Sc, Here Lombard (Lo) may refer to the Lombard effect in consideration with the database used, the environment in which it was created, and how the researchers applied it in their research.

Several studies have detected stress in different contexts using various algorithms such as independent component analysis (Palacios et al., 2020), neural networks (Hansen et al., 1996), decision trees (Tavi, 2019). Nonlinear features were also employed for classifying stress (Zhou et al., 2001), using available stress voice database such as SUSAS (Hansen & Womack, 1996) or emergency call recordings (Demenko, 2008). However, few studies have extended the research to detect stress during specific stressful events, such as public speaking. These studies utilised either nonverbal features (Han et al., 2018; Han et al., 2014; Pfister & Robinson, 2011) or physiological features (Barreto et al., 2007; S. Gillespie et al., 2017).

Barreto et al. (2007) was one of the limited studies that detected stress by monitoring physiological signals, such as galvanic skin response, temperature of skin, blood pulse volume and left eye pupil diameter of the subjects, during a determined computer-based, paced Stroop test. The Stroop incongruent segment was related to the stressed state in participants, while congruent Stroop was linked to a nonstressed state in participants. Twelve out of 32 samples were used to train the classifiers while the remaining 20 samples were used to test the classifiers. All the extracted features of the physiological signals, such as the mean amplitude of the individual blood pulse volume beat and the mean value of pupil diameter, which were expected to increase under stress, were given as the input to various ML algorithms, such as Naïve Bayes, Decision Tree and SVM classifiers, to classify the relaxed versus stressed state in participants (Barreto et al., 2007).

Another study conducted by Han et al. (2018) proposed an algorithm to determine the state of stress (stressed or unstressed) via voice features using a binary decision criterion involving two layers of Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) and a classifier. The authors evaluated the proposed algorithm by obtaining the speech, video, and bio-signal data under stressful (interview of the Korean speaking subjects in English by a foreigner) and nonstressful (comfortable conditions such as watching videos) conditions and

surveyed participants about the stress they felt. Based on the survey results and cortisol levels, the speech signals from 25 subjects whose salivary cortisol level changed more than 10 per cent were labelled as stress or nonstress in the database. Additionally, 15 subjects were used to train the model, while five stressed and five unstressed speech samples from the database were used to test the data. The authors preprocessed the data before extracting the MFCC, which were used to determine stressed and unstressed state from the testing data. The proposed LSTM-RNN with SVM classifier achieved 66.4 per cent accuracy in stress detection (Han et al., 2018).

Pfister and Robinson (2010) presented a new classification algorithm to assess public speaking skills. They trained the classifier using the Mind Reading corpus (Junek, 2007) for the detection of both simple and complex emotions. For the classifier to detect the emotions, the authors chose a wide variety of emotions such as joy, interestedness, unfriendliness, excitement, unsureness and sureness from the Mind Reading corpus. They preprocessed the corpora, extracted the nonverbal features of speech and computed the SVM model. Then, to assess public speaking skills, Pfister and Robinson first retrained the classifier using six labels which are commonly used by professional experts. For this purpose, they asked an experienced speech coach to label 124 one-minute-long speech samples obtained from 31 speakers attending speech coaching. Finally, Pfister and Robinson segmented the live audio, extracted the audio features and ran the classifiers. For pairwise machine, the results yielded an average cross-validation accuracy of 89 per cent while for the fused machine, 86 per cent of accuracy was obtained. For assessing public speaking skills, the novel application of the classifier achieved 81 per cent cross-validation accuracy and 61 per cent accuracy when performed using a leave-one-speaker-out method (Pfister & Robinson, 2011).

The review in this section of various algorithmic models developed for affective state and emotion recognition using voice analysis highlights the extensive exploration of traditional

methods, such as GMM, HMM, SVM, KNN, and Random Forest, and newer architectures like ANN, including feedforward networks (FFN). These models have achieved accuracies ranging from 54.3-96.3 per cent across various datasets using diverse voice features. While all these traditional and new algorithms have shown high performance in general affect and emotion recognition across several contexts, the application of these algorithms for SDM development, particularly in the context of public speaking, remains significantly underexplored. Furthermore, most studies have relied on generic emotion datasets, such as Emo-DB and IEMOCAP, with limited focus on stress-specific voice databases. The lack of a contextual public speaking stress voice database further constrains research on SDM development. This highlights a need for further exploration of different voice features and algorithmic models and their greater integration into VR-based public speaking training systems. These gaps accentuate the need for contextual datasets to develop specialised stress detection models and then manage stress in public speaking.

2.6 Discussion of Findings from Search 1 and Search 2

Communication through speech is a powerful medium of human expression and connection, with the voice serving as a reliable channel for conveying emotions and affective states. Mastering effective public speaking skills is important in educational, professional and social settings (Docan-Morgan et al., 2015). However, affective states such as psychosocial stress often hinders the quality of the speech by impacting a speaker's voice (Koroleva et al., 2014; Ruscio et al., 2008). This discussion synthesises findings from Search 1, which focused on affective states, voice analysis, the use of VR for public speaking, and Search 2, which explored models for detecting emotions and affective states such as stress.

Search 1 revealed a strong relationship between human affective states, emotions and voice. Many research studies have demonstrated the effectiveness of voice in conveying

emotions via analysis of voice features such as pitch, intensity, formants and MFCCs, which play a crucial role in emotion detection (Cowen et al., 2019; Juslin & Laukka, 2003; Mitchell & Ross, 2013). Temporal, frequency domain, spectral and prosodic features have been extensively analysed to study and identify six common emotions; i.e., disgust, sadness, happiness, fear, anger and surprise (Bhavan et al., 2019; Busso et al., 2009; Hao et al., 2020; Hu et al., 2007; Jain et al., 2020; Nogueiras et al., 2001; Nwe et al., 2003a; Ooi et al., 2014; Schuller et al., 2003b; Semwal et al., 2017). For stress detection specifically, research has also found that TEO based features, and especially the TEO CB AutoEnv, are effective in identifying stress from voice in comparison to all other types of features.

Despite significant advancements in affective computing and HCI, voice features indicative of affective states like stress still remains highly complex due to their subjective and context-dependent nature. There is insufficient research specifically on voice features for real-time stress during a potentially stressful event such as public speaking. Importantly, voice features often overlap between different emotions and affective states, therefore, require careful feature extraction within a given context. In the context of public speaking, overlapping emotions such as fear and anxiety, which share heightened physiological responses with stress, may provide valuable insights for developing stress detection models, highlighting the need for a broader approach to recognise affective states like stress, as traditional emotion recognition methods may not be sufficient. The development of algorithmic models for stress detection further highlights the importance of selecting appropriate voice features for targeted affective state detection from selective context-specific voice database, as they significantly influence the accuracy and reliability of stress detection models for a given context such as of public speaking.

Search 1 focusing on use of VR for public speaking further highlights VR as a promising tool for public speaking training, offering controlled and immersive environments

for simulating realistic scenarios and allowing researchers to study emotions like fear, anxiety and stress. Studies make it evident that conducting public speaking tasks in a laboratory setting can help in the evaluation of real-life stress (Jezova et al., 2016). Previous studies have shown that participants in VR environments respond to virtual audiences similar to real-life audiences, suggesting that interactions with virtual entities could be used to elicit social stress and be used to provide virtual social support (VSS) during stressful activities (Kothgassner et al., 2019). In this regard, research has been carried out using VR to examine the effects of VSS for public speaking (Chollet et al., 2015; Hartanto et al., 2014; Pertaub et al., 2001; Slater et al., 2006; Slater et al., 1999). However, despite its potential, most of the VR studies (e.g., Barreto et al., 2007; Slater et al., 2006) in context of public speaking focused on inducing, studying the effects of virtual audience behaviour on a speaker's anxiety and fear of public speaking by measuring the speaker's physiological parameters such as heart rate or blood pressure, with limited integration of voice analysis techniques for real-time stress detection. In an example, El-Yamri et al. (El-Yamri et al., 2019) created a VR system that included a virtual audience character model. Stress elicited in a speaker was identified and analysed using a third-party affective state detection application and the speaker's state was weighted using the authors' common sense only, which can be culturally dependent and hence could potentially bias the recognition of certain emotions over others. Moreover, this model lacks a learning process for the virtual agent based on the analysis of voice features and other factors such as physiological features. Also, knowing how the virtual audience were made to work can help in improving the experiment and real-time feedback to the speaker during public speaking. The research could have been improved by detecting emotions or affective states from speakers' voices, which could have made the detection of stress more accurate since the human voice is a powerful channel to convey emotions, and this has not been done using VR before as accurately as it

could be done. Similarly, many studies rely on subjective self-reports or third-party tools with unclear criteria, limiting reliability.

Search 2 was focused on reviewing and exploring various ML models for affective state detection and emotion recognition systems through voice analysis. Traditional algorithms such as GMM (Li et al., 2007), HMM (Nogueiras et al., 2001; Nwe et al., 2003a), SVM (Barreto et al., 2007; Truong et al., 2012) with modern computational algorithms such as ANN, RNN, CNN, DNN have been widely used for detection of affective state and emotion. For instance, HMM outperformed SVM, achieving 99.5 per cent accuracy compared with SVM's 88.9 per cent (Lin & Wei, 2005). Modern computational approaches such as ANN, RNN, CNN and DNN have demonstrated superior performance in emotion detection, leveraging voice features for more nuanced affective state recognition (Dahl et al., 2011).

Despite these advancements, SDMs, particularly in the context of public speaking, remain underexplored. Most studies rely on generic emotion datasets and lack contextual public speaking stress voice databases, limiting their applicability to real-time public speaking scenario SDM development. While most studies still utilised physiological features (Barreto et al., 2007; S. Gillespie et al., 2017) or self-reported indices (Bandela & Kumar, 2019) in several different contexts to identify stress, some studies have attempted stress detection during stressful events using nonverbal features (Han et al., 2018; Pfister & Robinson, 2011), these approaches often neglect contextual correlations with physiological and voice parameters. For example, in one of the few studies, Pfister and Robinson (2011) used nonverbal features to assess public speaking and the underlying affective state, leaving out their correlation with physiological parameters. The accuracy of this method of detecting stress may not be high because of the restriction to just nonverbal features, leaving out the contextual correlation with physiological parameters in the context of public speaking. Search 2 highlights that real-time voice-based stress detection in the areas of affective computing and HCI has not received much

attention, specifically within a public speaking context. These findings highlight the complexities of creating a suitable testing environment specific to public speaking, the need for specialised voice datasets, features and stress detection.

The findings from both searches underscore the critical role of voice features in affective state detection and the potential of VR as a tool for public speaking training. The studies reveal that various voice features, datasets and algorithms play a significant role in emotion recognition and affective state identification. Several features such as pitch, intensity and spectral properties have been successfully used to identify common emotions and affective states, but stress-specific voice features remain underexplored particularly in the context of public speaking. Although VR has proven effective for public speaking training by providing immersive and controlled environments, the lack of integrated stress detection mechanisms limits its applicability. This highlights the need for developing and embedding an affective state detection model into VR systems to improve their effectiveness for stress management during public speaking. To develop an accurate SDM, it is essential to have a context-specific training dataset, carefully extracted voice features and an appropriate classification algorithm. These components are crucial for creating reliable models that can identify stress in real time and enhance the utility of VR-based public speaking training systems.

2.7 Identification of Key Gaps and Research Opportunities

The review shows that there are opportunities for integrating real-time audio and voice analysis within a VR set-up, which can help in creation of a valid training tool to improve public speaking, contributing to voice analysis based affective computing and human-computer interaction research. However, the synthesis of studies from this review highlights that the subtle yet profound impact of stress on voice within the context of public speaking needs investigation and exploration. Specifically, there remains a gap in understanding of the real-

time dynamics of vocal stress indicative features, especially during live public speaking scenarios. Additionally, the lack of a context-specific stress voice training datasets limits precise voice feature extraction and development of building accurate and reliable SDMs, specific to the context of public speaking. The integration of affective state detection models into VR systems could significantly improve their usefulness by providing personalised real-time feedback to the speaker based on real-time stress levels. Hence, this gap presents an opportunity for exploring and advancing research into voice-based stress detection in the given context of public speaking.

2.8 Outline of the Present Research and Addressing Anticipated Barriers

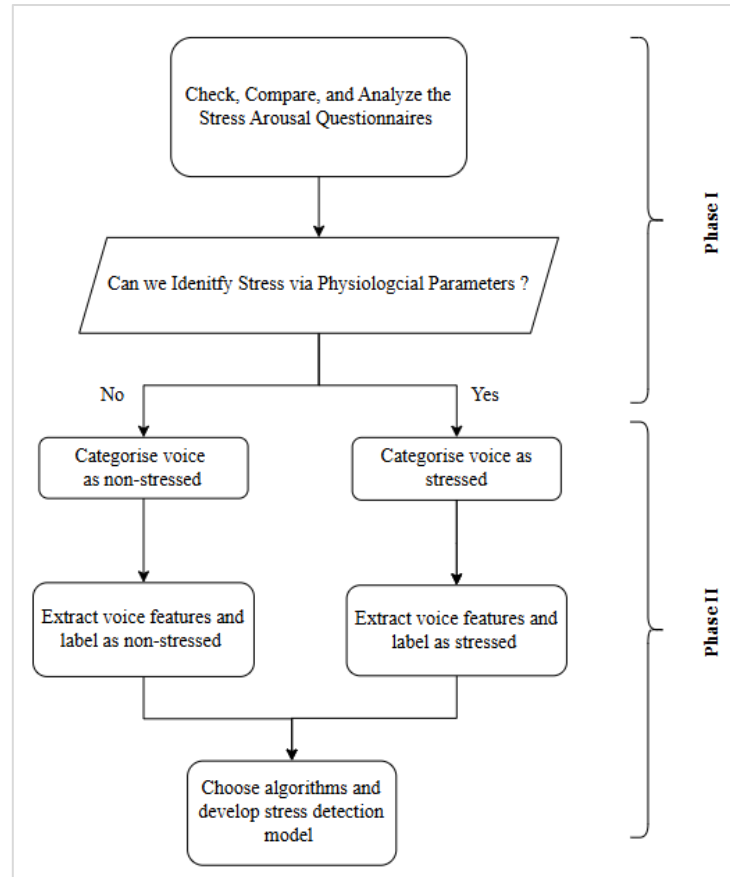
This section aims to provide an overview of the present research, including objectives, methods, phases and key steps. Additionally, it addresses anticipated barriers that could arise during the research process and discusses strategies implemented to mitigate these challenges.

The purpose of this research was to develop a stress detection model and integrate it into a VR application, providing realistic feedback based on stress levels detected through voice analysis to improve public speaking skills. This research employed a quantitative approach (Creswell et al., 2017) and follows an experimental design structured into two phases: Phase I and Phase II.

Phase I (**Chapter 3**) of the research was dedicated to collecting relevant data consisting of stressed and nonstressed voice samples and determining the stress induction capabilities of VR scenarios via Experiment 1. A primary objective and concern was to acquire contextual public speaking voice data necessary for developing an effective SDM. To achieve this, VR prototypes were designed to induce stress by simulating different behaviours from a virtual audience. Two distinct VR scenarios were developed: one featuring positive, supportive audience behaviour and the other featuring negative, nonsupportive audience behaviour. Experiment 1 sought to determine whether VR scenarios with nonsupportive virtual audiences

induce greater stress compared with those with supportive audiences. This approach aimed to assess the stress-inducing properties of VR scenarios across individuals and to collect contextually relevant voice data. The Experiment 1 data collection of voice samples began following the steps mentioned in Figure 2.5.

Figure 2.5 Steps in Development of Stress Detection Model to be Integrated in VR application



To ensure the reliability of stress voice sample categorisation, stress levels were measured using physiological data (heart rate, blood pressure) collected through manual devices and wearable heart monitoring belts. Closed-ended questionnaires and pretest/posttest assessments were used to record psychological stress levels. Statistical tests on the physiological data and questionnaire responses collected before and after the VR scenarios were conducted to explore differences in stress levels induced by the varied VR conditions. Interplay between physiological parameters and psychological stress levels were established. The statistical analysis revealed significant differences between the two conditions in their

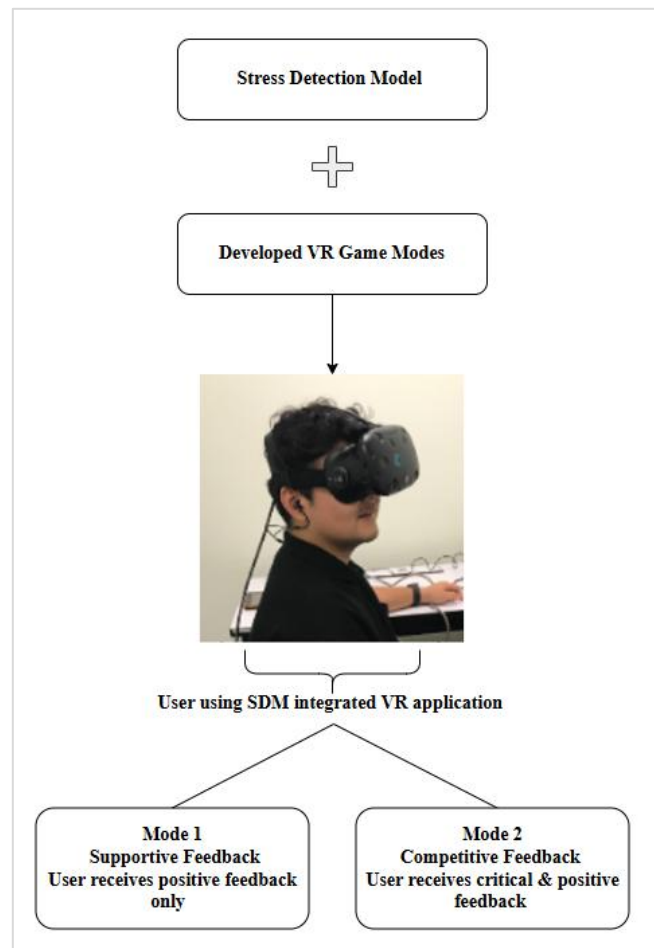
ability to induce stress, thereby enabling accurate categorisation of voice samples into stressed and nonstressed voice samples needed for Phase II of the study.

Based on the results from Phase I, Phase II (**Chapter 4 and Chapter 5**) of the study involved creating a comprehensive database of stressed voice recordings during public speaking in VR and developing a machine learning-based SDM.

Another anticipated barrier was the development of a SDM or software for extracting and analysing the necessary voice features. To overcome this challenge, preliminary analysis using the publicly available RAVDESS (Livingstone & Russo, 2018) was conducted. Python libraries such as My Voice Analysis (Shabahi, 2020), LibROSA (McFee et al., 2015) or pyAudioAnalysis (Giannakopoulos, 2015) were employed to extract temporal, frequency domain, spectral features. The short preliminary work provided a robust foundation for feature extraction process. A similar approach was applied to the collected voice data, focusing on extracting features such as Mel-frequency cepstral coefficients (MFCCs). These features were crucial to building the SDM. The SDM was trained using the previously collected voice data to accurately detect stress levels in real time.

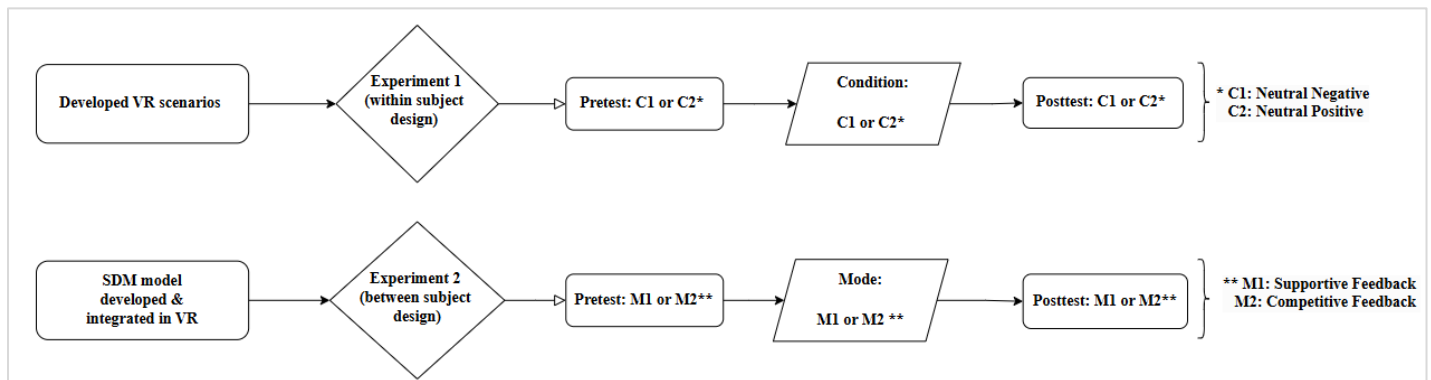
The integration of the SDM into the VR application required overcoming technical challenges. A script was written to enable the virtual agents to provide real-time feedback to the speaker every 10 seconds based on the speaker's voice data. By integrating the virtual audience with the SDM, the system was designed to analyse stress elicited in speakers and deliver personalised feedback. The real-time implementation of the SDM within the virtual audience feedback mechanisms allowed the VR application to serve as both a stress analysis tool and a training platform for managing public speaking stress.

The second part of Phase II tested the VR game's integrated SDM's capacity to provide real-time feedback and mitigate stress during public speaking, addressing the final research question in Phase II via Experiment 2 with two VR game modes (Figure 2.6).

Figure 2.6 *Developed VR application for Public Speaking*

Statistical analysis of physiological data and closed-ended responses collected before and after the VR game modes in Phase II evaluated the stress-buffering effects of real-time feedback on public speaking improvement. The findings indicated whether the integrated SDM could effectively reduce stress and enhance public speaking performance by providing real-time feedback based on detected stress levels.

The below diagram (Figure 2.7) illustrates the overall research methods for answering the identified research questions as mentioned in Chapter 1 [Research Objectives and Questions](#), developing the VR game, and testing the various specifically derived hypotheses in Phase I: Experiment 1 (E1) (refer to Chapter 3) and Phase II: Experiment 2 (E2) (refer to Chapter 5).

Figure 2.7 *The Overall Process of the Research Design*

Ethical concerns remained a top priority throughout the research. The ethics approval follows Australia’s National Health and Medical Research Council guidelines as stated in the National Statement on Ethical Conduct in Human Research, 2007 (updated 2018), specifically referring to Statement Chapter 2.1. The project’s overall risk category was, more than low risk “research with the potential to cause mild psychological distress or physical stress. Minor deviations from the frank disclosure of the true nature of the research may be involved”. The ethics approval was sought for both studies involving human participants (relevant in-depth details of approvals are discussed in Chapter 3 and Chapter 5).

The goal of Chapter 3 is to provide in-depth information about the development of the VR prototypes, methods, evaluation and results of Experiment 1.

Chapter 3: Evaluating and Validating Stress Arousal Capability of Virtual Reality for Public Speaking via Experiment 1

This chapter describes the research methods used in Experiment 1 (E1), which involved the development of two simple interactive and noninteractive virtual environments for creating a database of stressed voice samples, specific to context of public speaking. The chapter then continues to report the findings of E1 and discuss their implications.

3.1 Rationale and Aim of Experiment 1

Mastering public speaking skills is a vital skill for success. Various affective states such as stress during public speaking impairs the effective communication of the message to the audience and therefore hinders individuals' progress in professional and social settings. Managing stress during public speaking is required for effective message delivery, connecting with audience and creating greater impact (Arnold, 2018). For details, please refer to [Chapter 1](#).

Stress is a subtle, yet complicated psychological response characterised by its valence (positive or negative experience) and arousal levels (ranging from high alertness to low alertness), which is a subjective feeling of the individual, perceiving the assigned demands as overwhelming (Russell, 1980). Initially arising in the mind, stress later manifests physiologically, impacting various bodily systems such as the cardiorespiratory functions. The affective state of stress triggers associated emotions such as anxiety and fear and therefore makes the stress detection process a difficult and challenging problem to be solved (Jackson et al., 2006) as previously discussed in [Chapter 2](#).

Stress, being subjective in nature, is expressed in different ways. Stress is recognised by usually capturing the individual's stress levels using various psychological tools (e.g., self-reported questionnaires), or measuring physiological parameters (e.g., heart rate, blood

pressure, salivary cortisol, and galvanic skin response) in laboratory experiments and simulated stressful situations during public speaking (Feldman et al., 2004; Lipp et al., 1999). These assessments and changes also highlight the effect on cardio respiratory functions in the bodily systems, and therefore manifest in specific variations of voice, such as breathing patterns, voice modulation, and speaking rate, indicating an individual's stress during public speaking (Pisanski et al., 2018; Pisanski et al., 2016; Pisanski & Sorokowski, 2021).

However, voice-based stress detection in the context of public speaking is notably underexplored within the fields of affective computing and HCI (Pfister & Robinson, 2011). There are few dedicated studies on voice analysis for detecting stress specifically in the context of public speaking. Research gaps exist in identifying stress indicators from changes in breathing patterns or nonverbal voice features. Accurate voice features indicative of stress during public speaking remain not known in the context of public speaking.

Context is one of the important factors to determine affective state and emotional state. Research shows that studies often employed available databases, such as SUSAS, which comprise stressed voice samples from professional actors or pilots in stressful situations to analyse voice samples and identify stress voice features (Hansen et al., 1997). Research also points towards a lack of proper database containing stressed voice samples recorded in real-time specific to the context of public speaking. The lack of contextual voice database limits the development of algorithmic models specifically designed for detecting stress during public speaking and for training virtual audiences to enhance feedback mechanisms within VR simulations.

There exists a notable gap in research concerning stress experienced during public speaking in VR. Studies utilising VR and various types of virtual audience to explore public speaking and detect emotional states often shows a neglect to incorporate voice-based stress analysis for identifying stress (El-Yamri et al., 2019). This omission limits the potential for

providing real-time feedback to speakers, which could enhance their management of affective state of stress and consequently improve their public speaking abilities.

To effectively handle stress and develop public speaking skills, it's important to explore and incorporate advanced technologies like voice analysis within VR simulations for training of such skills. VR offers consistent and controlled training scenarios for public speaking and when combined with ML algorithms to detect stress, it can provide real-time feedback on stress levels through automated analysis. This integration enhances user experience and improves learning outcomes significantly. These integrated technologies serve as invaluable training tools through the implementation of advanced techniques like real-time voice analysis for detecting affective states and providing immediate feedback during public speaking exercises. Developing new machine learning-based detection models for real-time stress analysis requires the identification of features that can reliably distinguish and detect stress, which often impedes effective communication. The development of such systems requires a curated database containing relevant voice recordings related to the specific context for ensuring their relevance and effectiveness in training scenarios.

Therefore, Phase I of the current study, comprising Experiment 1, aims to create a new database of stress-related voice samples for further research by validating the stress-arousal capabilities of the developed immersive VR public speaking scenarios. This is achieved through the investigation of the following research question and corresponding hypotheses as mentioned in Section 3.2.

There are several potential implications of this study. Firstly, the findings can enhance the existing literature on the stress arousal capabilities of VR. If the VR prototype environments, featuring various virtual audience behaviours, successfully induce stress in speakers – measured by changes in heart rate, systolic blood pressure, and diastolic blood pressure – this could lead to the creation of a voice database containing stressed voice samples.

This database would be valuable for further research on stressed versus nonstressed voice samples during public speaking. By analysing these samples, researchers could identify precise stress-related voice features, which could then be used to develop ML models for stress detection from voice. Furthermore, integrating these ML models with the VR environments could be used to provide real-time feedback from virtual audiences to the speakers, thereby enhancing the interactivity and effectiveness of virtual public speaking training tools.

3.2 Research Questions and Hypotheses of Experiment 1

The primary focus of E1 was to create a contextual voice database of stress vocal samples linked to public speaking, and answer RQ₁ via trying to find answer to following derived hypotheses (H1a) and (H1b):

- **Research Question 1 (RQ₁):** Could the different behaviours demonstrated by the virtual audience in both supportive and nonsupportive scripted VR scenarios as provided in our specifically developed simulation impact a speaker's stress level irrespective of their performance, potentially leading to stress induction in the users?
 - **Hypothesis 1a (H_{1a}):** Applause [A] in comparison to booing [B] may reduce stress (perceived stress level) and may decrease blood pressure and heart rate because participants may feel motivated in Neutral-Positive (NP). Please refer to Table 3.1 for details on the NP condition.
 - **Hypothesis 1b (H_{1b}):** Booing [B], in comparison to applause [A] may increase stress (perceived stress levels) and may increase heart rate and blood pressure because participants may feel stressed in Neutral-Negative (NN). Please refer to Table 3.1 for details on the NN condition.

3.3 Design of Experiment 1

The design of this experiment focuses on manipulating two distinct conditions using VR to assess the impact of virtual audience behaviour on speaker stress levels and collect voice samples of speakers giving speech in both conditions. These conditions are detailed in Table 3.1.

Table 3.1 *Virtual Reality Scenario Conditions*

Condition	Description
Neutral-Positive (NP)	This condition involves a virtual audience providing neutral responses initially, followed by positive behavioural feedback. The NP condition's initial phase of the first 30 seconds includes the virtual audience that exhibits neutral response such as showing no applause or attentive listening behaviour, or feedback to establish a baseline. In the subsequent 2 minutes feedback phase, the virtual audience transitions to positive feedback behaviours, such as attentive listening and applauding (denoted as [A]).
Neutral-Negative (NN)	The virtual audience in this condition involves providing neutral responses, followed by negative behavioural feedback. Similarly, the NN condition starts with the first 30 seconds involving neutral audience behaviour. In the following 2-minute feedback phase, the virtual audience displays progressively negative behaviours, including inattentive listening, looking at phone screens, talking on the phone, and eventually booing (denoted as [B]).

The NP condition is designed to observe how positive reinforcement from a virtual audience affects speaker stress levels. Positive feedback, such as attentive listening and

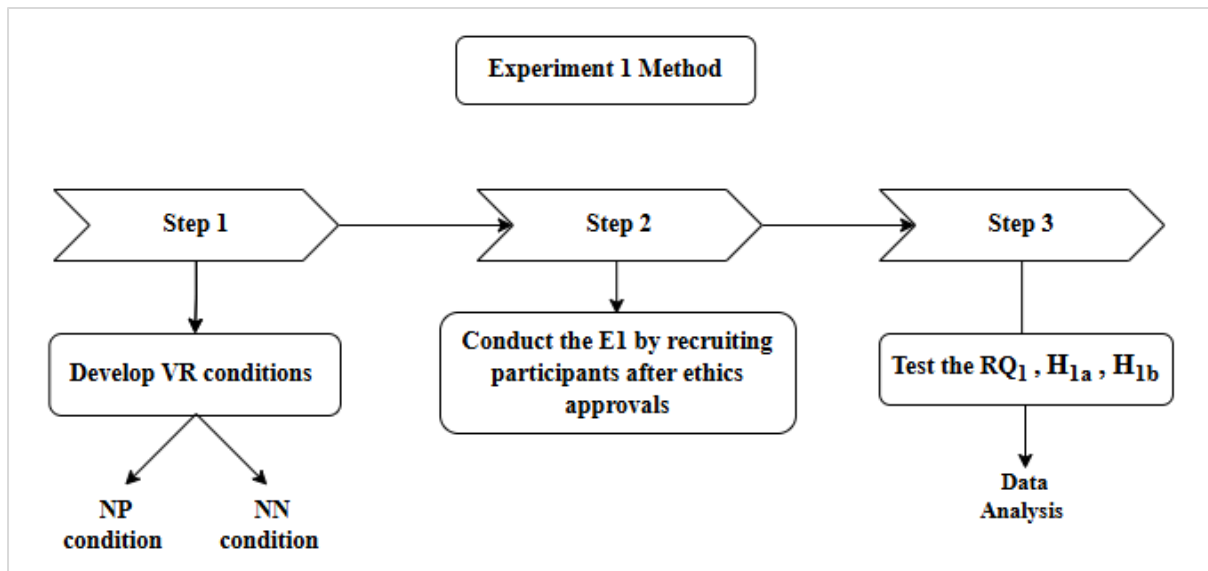
applause, is hypothesised to reduce stress. The NN condition aims to investigate the impact of negative feedback on speaker stress levels. Behaviours like inattentive listening and booing are expected to increase stress.

It was anticipated that participants will exhibit lower stress levels due to the positive feedback from the virtual audience in the NP condition. In a more critical condition, it was expected that participants will show higher stress levels in response to the negative feedback behaviours. By manipulating these two conditions, the experiment seeks to validate the hypothesis that virtual audience behaviour significantly induces speaker stress levels and influences the stress in speakers. This allows for voice data collection and creating a database of stress voice samples during public speaking in VR environments.

Collecting voice samples in an authentic non-VR environment might reflect real-world public speaking scenarios but the use of VR offered several advantages. VR allowed for controlled, repeatable conditions, ensuring consistency in stress-inducing factors including virtual audience behaviour and environmental settings. This consistency was essential for systematically studying the effects of stress on vocal characteristics and for generating a reliable dataset. Additionally, VR enabled the creation of tailored scenarios for inducing varying levels of stress, which would have been challenging to achieve in a non-VR environment due to logistical issues in being able to have large audiences and respective additional ethical considerations.

3.4 Methods of Experiment 1

The methods for E1 (Figure 3.1) involve developing various prototypes of VR audiences exhibiting different behaviours, as defined by the previously established conditions. These prototypes are then used to test hypotheses and collect voice data from speakers in different states via E1 for next Phase II of research.

Figure 3.1 *Overall Methods of Experiment 1*

Note. VR = Virtual Reality, NP = Neutral-Positive, NN = Neutral-Negative, E1 = Experiment 1, RQ₁ = Research Question 1, H_{1a} = Hypothesis 1a, H_{1b} = Hypothesis 1b.

Keeping in mind the main aim of this Phase I consisting of E1, the first step was to create the required VR scenarios, featuring two distinct virtual audience behaviours. The second step involves conducting the E1, by recruiting participants and checking their engagement with the VR scenarios. Participants were asked to give speeches in both VR scenarios. During speeches, physiological parameters (systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (HR)), self-reported perceived stress, self-reported perceived arousal and voice samples were recorded simultaneously.

The third step involved data analysis of physiological parameters and self-reported questionnaires to validate the stress arousal capability of the VR scenarios. After completing Experiment 1 (E1) and analysing the data, the voice samples collected are used in Phase II of the research. This phase focuses on detecting various features indicative of stress.

Thus, experiment 1 serves to validate the stress arousal capability of the developed VR game scenarios and concurrently collects voice data for subsequent research Phase II.

3.5 Development of VR Prototypes for E1

3.5.1 The Virtual Environments for Public Speaking

The scenes in the VR game were defined based on the conditions with respective behaviours defined earlier. Two different scenarios with NP (Figure 3.2) and NN virtual audiences (Figure 3.3) were created in the VR.

Figure 3.2 *Neutral-Positive (NP) Virtual Audience (Positive Behaviour)*



Figure 3.3 *Neutral-Negative (NN) Virtual Audience (Negative Behaviour)*



The public speaking scenarios were created to simulate an auditorium setting, thus providing a realistic environment for the experiment. The virtual auditorium was designed to be spacious, large and open. The ample space allowed the speaker to move around freely, much

like on a real stage mimicking the dynamics of an actual stage. This design allowed participants to experience the physical aspects of public speaking environments, such as moving around while addressing the audience (although participants remained seated due to physiological recording requirements; detailed later in this chapter on page 127). The design was aimed to enhance the immersive experience for the participants, making the VR scenarios as lifelike as possible. This could be seen in the following Figure 3.4, Figure 3.5, and Figure 3.6.

Figure 3.4 *The stage speaker can walk around and move to right*



Figure 3.5 *The stage speaker can walk and move to front*



Figure 3.6 *The stage speaker can walk around and move to left*



The ceiling of the auditorium was kept high to add to the sense of realism and immersion as shown in Figure 3.7 and Figure 3.8. The stage itself was designed to allow the speaker to look around and see the entire audience, as well as the back walls of the auditorium, providing a comprehensive view of the environment. The full auditorium can be visualised as follows in Figure 3.9.

Figure 3.7 *The ceiling of the VR auditorium*

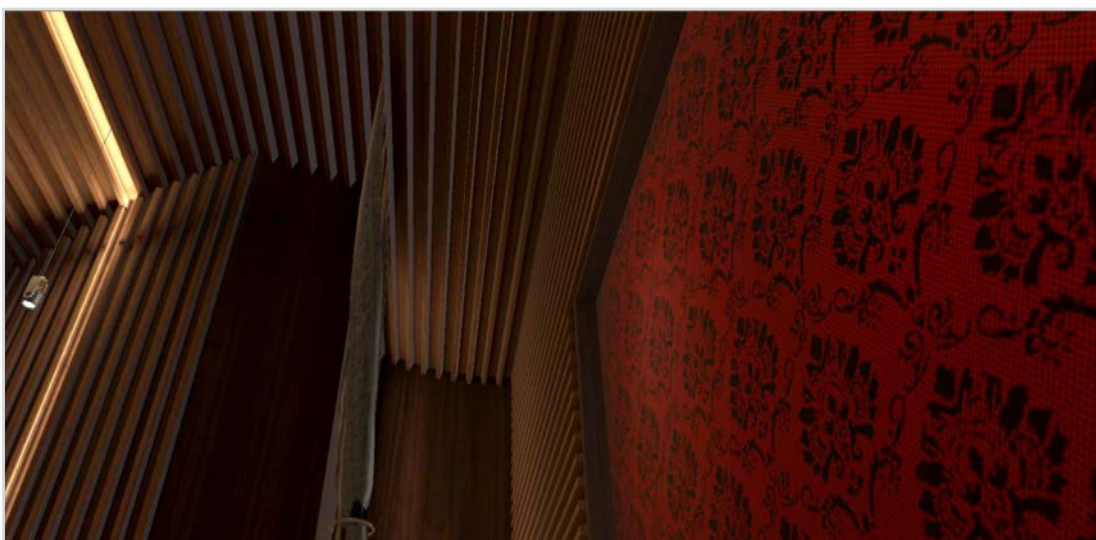


Figure 3.8 *The ceiling of the VR auditorium kept high*



Figure 3.9 *Full auditorium view*



Furthermore, the auditorium package was sourced from the Unity online asset store, specifically purchased from 3D artist and freelancer SatishParab (<https://assetstore.unity.com/packages/3d/environments/auditorium-141127>). The Unity online asset store is a comprehensive marketplace offering a wide range of 3D assets, tools, audio files, and templates for game or prototype creation (<https://assetstore.unity.com/>).

The selected auditorium model is a low poly design, ensuring efficient performance while maintaining visual quality. It features proper UV unwrapping, ensuring that textures are applied seamlessly without any overlapping issues. This level of detail is crucial for creating a realistic and immersive environment. Additionally, the auditorium model used in the VR environment included pre-baked lighting and textures, created in external 3D software by the asset artist, eliminating the need for further baking in Unity. Choosing these models streamlined the development process allowing for a more efficient integration into the VR prototypes.

The low poly nature of the model ensures that it runs smoothly on a variety of hardware configurations (in our case, HP Omen laptop with an Intel Core i7-10750H CPU (2.60 GHz) and 32 GB RAM), which is essential for maintaining consistent performance during the experiment. The prebaked textures and lighting contribute to the overall realism of the environment, enhancing the immersive experience for participants. By utilising high-quality assets from the Unity asset store, I ensured that the visual and functional aspects of the VR auditorium met the necessary standards for conducting the research effectively.

3.5.2 Virtual audience

A total of 70 virtual audience members were incorporated into each scene. These virtual audience members were represented by male and female human avatars, designed with a range of appearances to reflect both professional and casual attire. This diversity in the audience's appearance aimed to mirror real-world public speaking environments accurately. Not all the 70 virtual audience members were scripted to react identically or simultaneously during the VR public speak task. Approximately 25 to 30 avatars were scripted to display explicit and emotionally salient behaviours (such as clapping, booing, talking or using phones) depending on the NP or NN condition. The remaining 30 to 40 audience members remained seated and exhibited specific behaviours, again depending on the condition, such as maintaining eye contact or appearing disengaged. Reactions were coordinated but not randomized. Further

implementation details are provided on pages 115–116. The virtual audience was strategically placed throughout the auditorium to ensure that participants could see and interact with them from various angles, enhancing the sense of being in a real public speaking setting

The virtual audiences were also sourced from the Unity online asset store. Specifically, the ‘Toon People’ used in the VR scenario creation were purchased from the modeller and animator JBGarraza

(<https://assetstore.unity.com/packages/3d/characters/humanoids/humans/toon-people-111042>). This package included a comprehensive set of 3D models, prefabs, scenes, and animations of characters, providing a diverse range of options for creating realistic virtual audiences.

In designing the virtual audience for this study, the focus was on achieving behavioural realism – such as appropriate gestures and reactions – rather than photorealistic visual fidelity. Choosing 3D Toon People for the development of two VR scenarios was aligned with research which shows that realism such as gestures, applause, and booing has been shown to play a more critical role in eliciting social evaluative pressure in VR environments, which was central to the study’s objectives. Research in VR suggests that the sense of presence and engagement is more influenced by interactive and behaviourally appropriate entities than by their photorealism (Herrera et al., 2020). Although integrating highly realistic virtual characters into the interactive VR environment would have been possible, it could have significantly increased computational demands in terms of rendering and animation resources. This would have affected the realism and fluidity of avatar behaviours, requiring high-end hardware to maintain immersive performance. In addition to these graphical demands, implementing a real-time stress detection and adaptive affective feedback system within a single pipeline also would require careful system optimization to ensure that responsiveness does not disrupt the user experience. The current system was developed and tested on an HP Omen laptop with an Intel

Core i7-10750H CPU (2.60 GHz) and 32 GB RAM, which was sufficient for developing, testing, and running immersive scenarios with scripted audience behaviours. However, the integration of more graphically detailed virtual characters was intentionally avoided to ensure stable performance across sessions, especially when combined with the integrated affective feedback system and dynamic audience behaviours.

The package's 3D toon models were flexible enough to be customised for our scenario's development in the Unity game design software. The purchased package contained eight low poly animated characters and eight VR avatars. These 3D models provided a solid foundation for creating a diverse and realistic set of virtual audiences. The package included a comprehensive set of 139 animations which covered a wide range of actions and behaviours. Animations were provided for male, female and elder characters ensuring that the audience could represent different age groups and genders. A total of 93 static animations were available, depicting various postures and reactions such as attentive listening, clapping, looking at phones and more. These animations included both dynamic and static poses, allowing for realistic audience reactions and movements (example shown in Figure 3.10, Figure 3.11, Figure 3.12, Figure 3.13).

Figure 3.10 *An animation model of woman telling off*



Figure 3.11 *An animation model of woman telling off*

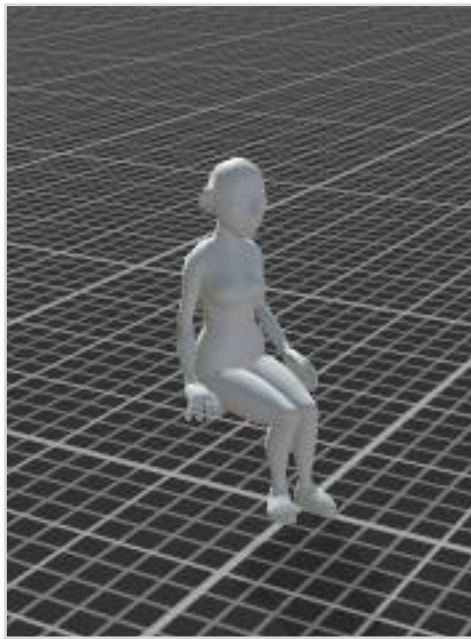
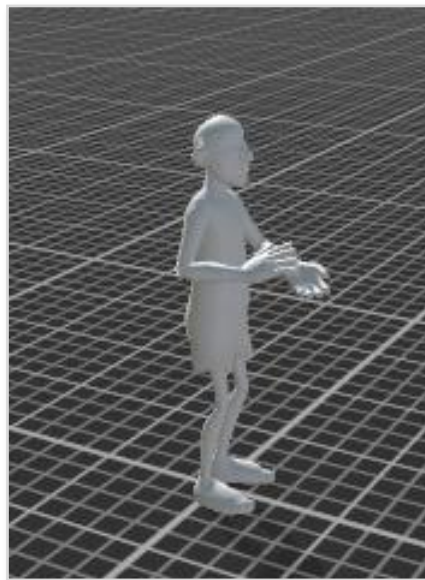


Figure 3.12 *A static animation model of man using his mobile*



Figure 3.13 *An animation model of man clapping*



The Unity Game Engine was utilised to develop the virtual audience characters for the scenarios. Prefabs which are specialised assets representing game objects or collections of game objects with predefined components, were employed from the available package which

included a total of 110 prefabs. These prefabs were modified and customised to suit the use case scenario of a public speaking audience. Modifications and customisations involved altering clothes, accessories (e.g., hat, glasses) and features (e.g., beard, hairstyle) for some prefabs. The characters were customised for the use case scenario to represent an audience in a public speaking setting. The character's hair and beard styles were altered to reflect a variety of professional and casual looks. Characters were dressed in a mix of professional attire, such as suits and ties, as well as casual clothing, to create a realistic and varied audience. Additional accessories like hats, glasses and different types of clothing were customised to enhance the diversity of the virtual audience. Characters' skin tones were adjusted to represent a diverse audience, ensuring a realistic mix of avatars. An example of these modifications is illustrated in Figure 3.14.

Figure 3.14 *An initial female character prefab (left) and changed female character prefab (right)*



The customised prefabricated models of virtual audiences were then integrated into two distinct VR auditorium scenarios, each designed to exhibit either NP or NN behaviours. After incorporating all the virtual audience models into the scenes, animations (Section 3.5.3) and audio elements (Section 3.5.4) were added to enhance the realism of the auditorium setting.

These elements were designed to be triggered based on specific conditions and were integrated with scripted behaviours (Section 3.5.5) to simulate various audience reactions.

The addition of animations, audio and scripted behaviours enabled the creation of realistic, lifelike scenarios where the virtual audience could exhibit diverse behaviours. These dynamic interactions were intended to reflect real-world situations and had the potential to influence the affective states of the speaker.

3.5.3 Animation

Several diverse animations from the 3D ‘Toon People’ package were utilised to create simulated scenarios for both NP and NN conditions in Experiment 1 (E1). The extensive range of animations enabled the virtual audience to have realistic interactions and responses, significantly enhancing the immersive experience for participants.

The NP condition is designed to provide positive and supportive feedback to the speaker. Accordingly, the animations included in the NP condition are chosen to show and maintain the appearance of neutral and positive only. During the initial 30 seconds of the NP condition, the virtual audience are portrayed visually to appear neutral with animations of observing around the auditorium, looking at speaker, turning their heads left and right and staying quiet. The remaining 2 minutes of the NP condition include the virtual audience providing positive feedback through animations such as attentive listening, maintaining eye contact with the speaker, looking directly at the speaker while listening, nodding and applauding.

In contrast to the NP condition, the NN condition is structured to provide negative and critical feedback to the speaker. Accordingly, the animations in NN condition are characterised by a neutral and negatively critical nature of the virtual audiences. The initial 30 seconds of the NN condition shows the virtual audience to visually appear neutral, along with animations of observing around the auditorium, looking at speaker, turning their heads left and right and

staying quiet. The remaining 2 minutes of the NN condition include virtual audience giving negative feedback through animations such as showing inattentiveness, checking their phones, engaging and talking on phone in front of the speaker loudly, to an extreme critical behaviour of booing. Participants in E1 were randomly assigned to experience either the NP or NN condition first. Following their initial experience, they were then exposed to the opposite condition (NP or NN) for their second session, ensuring that each participant encountered both types of virtual audience behaviours.

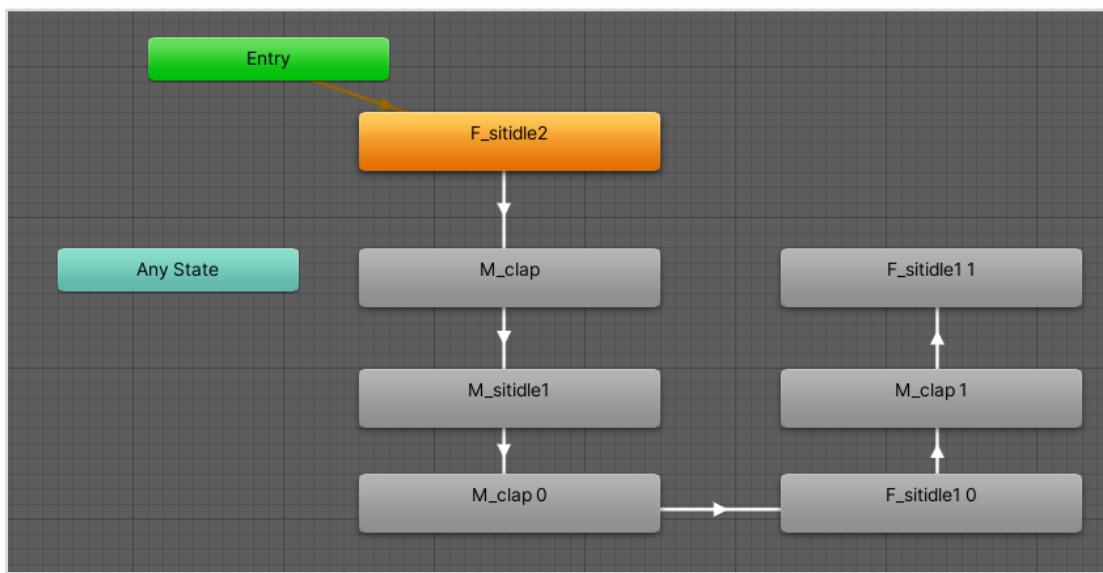
The animations in both conditions were precisely timed and set to run for only 2 minutes and 30 seconds, totalling 5 minutes together. The positive or negative feedback was shown by setting different sets of animations in motion. In both the NP and NN scenarios, different sets of animations were activated to convey either positive or negative feedback from the virtual audience. The scripted behaviours of the virtual audience in both NP and NN scenarios were given the same amount of time to play, to ensure consistency in duration and timing. This allowed the participants in both conditions to have the same amount of time for their reactions to appear and be measured during the experiment.

To execute the animated behaviours in motion for each condition, we used the controller property of the Unity game development engine and software tool. The controller property of the Unity enabled us to control and sequence the virtual audience's movements and reactions according to the specific behaviours defined for each condition. We created two different animator controllers for both the conditions. Animations in an animator controller are contained in different animation states.

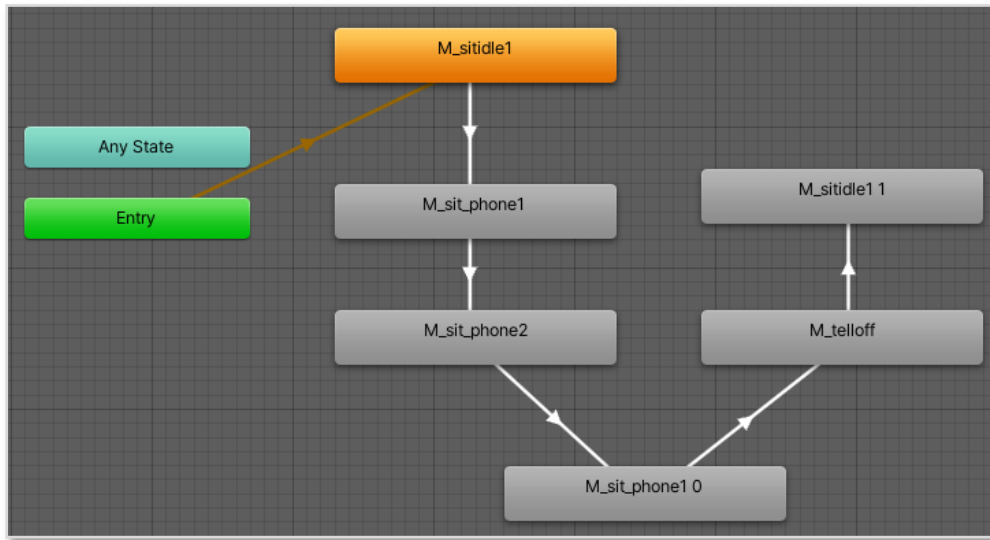
The animator controller for NP condition is shown in Figure 3.15. The NP state machine has seven states. The animator controller contained states of quiet sitting (attentive) and applause states. Furthermore, to allow the animator controller machine to make the decisions, animator parameters were added. The default state is displayed in orange. In the NP controller,

the default state is set to a sitting position (fsit_idle2). The animator controller includes another animation state of clapping for 12 seconds (m_clap). The third animation state included moving to an attentive listening, sitting position (m_sitidle1) and going back to clapping for another 12 seconds (m_clap0), followed by listening attentively with sitting down pose (f_sitidle1 1) and a final round of applause to clap (m_clap1) to sitting down (f_sitidle1 0). Each animation state was made to transition to next one with an arrowhead.

Figure 3.15 *Neutral positive (NP) animation controller*



In NN animator controller (shown in Figure 3.16), the default state is set to a sitting position (msit_idle1). The animator controller includes other animation state of picking up phone (m_sit_phone1). The third animation state included virtual audience talking on phone (m_sit_phone2), bringing the phone back (m_sit_phone1 0), followed by extreme booing (m_telloff) and a finally getting back to default sitting position (m_sitidle1 1). The parameters for the NN animation controller were added as well and their values were set and triggered by a script which will be discussed under scripts section.

Figure 3.16 *Neutral negative (NN) animation controller*

3.5.4 Audios

In a typical auditorium setting, various audience behaviours are accompanied by distinct sounds, and ambient noises are consistently present in the background. Drawing inspiration from real-life public speaking environments, diverse audio elements were integrated into the NP and NN scenarios within the VR simulations. The NP condition audios included an initial neutral and quiet auditorium background sound (shown in Figure 3.17). The second audio for NP condition included an applause sound effect. The initial neutral and quiet auditorium background sound was played each time the audience in NP condition were sitting quietly and attentively listening to the speaker. The applause audio always played out whenever the applause animation was set in motion.

Figure 3.17 *NP audios triggered with animations*

The NN condition audios (shown in Figure 3.18) included an initial neutral and quiet auditorium background sound, which was used for the default sitting positions. The second audio for NP condition included chatting and murmuring audio. The third audio in NN condition is people talking on phones in loud voices. The final audio always played out at the end when the extreme booing was set.

Figure 3.18 *NN audios triggered with animations*



3.5.5 Scripts

Different scripts were created as per need of the conditions in both VR scenarios development in C# language utilising the Visual Studio for this purpose, as it is linked to Unity Scripts automatically. Detailed scripts were developed to trigger specific animations along with their respective audios at predetermined times, accurately reflecting the NP and NN conditions. The audience reactions/behaviours were carefully scripted to provide feedback at specific intervals with sufficient buffer time between reactions. Each VR scene was designed to last between 2 minutes and 30 seconds for a total duration of 5 minutes. The scripting animation set-up ensured that the virtual audiences behaved in a consistent and controlled manner allowing for precise testing of the hypotheses.

In both scripts, different Boolean parameters were both passed to the respective coroutine method. The defined scripts were used to trigger the animation states based on its respective animation controller; respective audios followed by respective time to stop playing the animations. The scripts included utilising various audio files and parameters to activate the animator controller. Different game objects were created and attached together to form an

object to be used in the game. The scripts also included an end time for each of the animations that needed to be play.

For example, the NP condition script code is shown in Appendix A. The NP script is written to control animations and audio for this specific condition. First, a class called ‘ThreeMinPositiveS’ is defined, inherited from the ‘MonoBehaviour’. The defined class for this NP condition uses several variables and methods to coordinate the playback of audio clips and activate the specific animation states over a time of 2 min. 30 seconds. The scripts include various variables: ‘Animator anim’, ‘AudioSource audienceClip’, ‘public AudioClip [] clip’ and several other variables of type bool such as M_clap, Msitidle, et cetera.

The ‘Animator anim’ is an animator component that is a reference to control animations. The ‘AudioSource audienceClip’ is an audioSource component reference for playing audio clips along with the animation. The public variable ‘public AudioClip [] clip’ is for an array of audio clips to be used during the animations are played in NP condition. The bool parameters are used as parameters for the animator to control different animation states. The void start method is called before the first frame update in the Unity game. It initialises the first two variables and starts the coroutine called ‘posi’. The IEnumerator pos handles the timing and sequence of the animation and the corresponding audios played. The ‘waitforsecondsrealtime’ is used by the coroutine to wait for specific times before changing animation states and playing audio clips.

In the NP condition the coroutine posi (), begins by printing a start message and waiting for 2 seconds. It sets the first audio clip which is ambience sound clip for an auditorium setting, plays it and waits for 58 seconds. After this, the coroutine function triggers the first clap animation, plays the second audio clip, and waits for 12 seconds. The method coroutine then stops the first clap animation and triggers the first sitting idle animation, waiting for 25 seconds.

The method repeats the process for additional clap and sitting idle animations, interspersed with waiting periods and audio clip changes.

Similarly, the NN condition script code can be found in the same Appendix A. The NN script is written to control negative condition's animations and audio. In this script, a class called 'NegativesScript' is defined, inherited from the 'MonoBehaviour'. The defined class for this NN condition also uses several variables and methods (functions) to coordinate the playback of relevant audio clips and also activate the specific animation states over a time of 2 min. 30 seconds for this NN simulation. The script also includes variables: 'Animator anim', 'AudioSource audienceClip', 'public AudioClip [] clip' and several other variables of type bool such as (M_sitphone1, M_stidle2, M_tell off, M_sitidle1_0 etc).

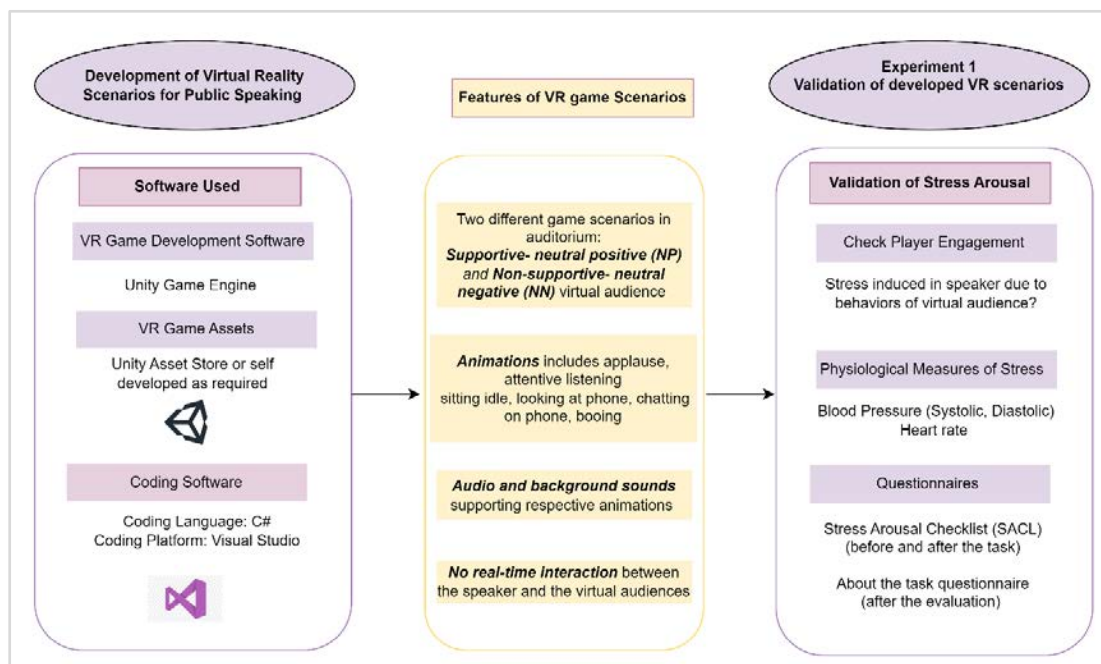
The 'Animator anim' is an animator component reference to control animations of NN condition in this script. The 'AudioSource audienceClip' is an audioSource component reference for playing audio clips along with the animations of negative and critical behaviours. The public variable 'public AudioClip [] clip' is for an array of negative audio clips to be used during the animations are played in NN condition. The Boolean variables are used to set parameters on the animator, triggering specific animations from the NN controller. The void start method (function) is called before the first frame update in the Unity game. It initialises the first two variables and starts the coroutine called 'nega()'. The IEnumerator nega() handles the timing and sequence of the animations and the corresponding audios played for the current negative simulation. Again, the 'waitforsecondsrealtime' is used by the coroutine to wait for specific times before changing animation states and playing audio clips.

In the NN condition the coroutine nega (), begins by printing a start message and waiting for 2 seconds. It sets and plays the first audio clip, with a volume of 0.45, and waits for 57 seconds. After this, the coroutine then plays the second audio clip on loop with a lower volume and waits for 20 seconds. It proceeds to activate and deactivate various animations

(M_stidle2, M_telloff, M_sitidle1_0), with corresponding audio clips and waiting periods. Each animation sequence involves setting an Animator Boolean to true (in the animator controller defined for this condition in previous section of animation), playing an audio clip, adjusting the volume, waiting for a specified duration, and then stopping the audio and resetting the Animator Boolean. The scrip then manages a sequence of negative animations and audio. The coroutine handles the timing of these events efficiently, so that animations and audio clips play in a coordinated manner. The script for a total duration of 170 seconds approximately.

The development work of the Experiment 1 prototype for VR public speaking scenarios involved creating virtual audiences and an auditorium using the Unity Game Engine, Visual Studio, and C# programming language (Figure 3.19). Two game prototype conditions were designed: NP and NN. The NP scenario features a virtual audience that claps and listens attentively, while the NN scenario includes audience behaviours like phone usage, chatting and booing. These scenarios incorporate animations and audios to enhance realism.

Figure 3.19 Overall Development for Conducting E1



The validation of these VR scenarios via E1 ([in next section 3.5.6](#)) involved assessing player engagement, measuring physiological stress responses such as systolic and diastolic

blood pressure and heart rate, and administering questionnaires like the stress arousal checklist (SACL) before and after the task, along with posttask feedback. The comprehensive approach ensures that the VR scenarios effectively simulate immersive public speaking environments and elicit the intended stress responses for research purposes in both conditions. Additionally, this process allowed to develop a voice database with various voice samples exhibiting different affective states, which will be valuable for further research in Phase II.

3.5.6 Method

3.5.6.1 Participants

Past studies utilising VR for public speaking involved 12-20 participants (Slater et al., 1999) and up to 36-40 participants (Hartanto et al., 2014; Pertaub et al., 2001; Slater et al., 2006). In the field of IT, studies that analysed voice used the predefined databases voice samples (Busso et al., 2008; Douglas-Cowie et al., 2007; Hansen et al., 1997), instead of real-life participant's voice samples to evaluate the accuracy of affective state detection models. Mostly, the databases require 10-30 professional actors or speakers to record voice samples with different tone and speak in different emotions. In the present research, a G*Power analysis for repeated measures multivariate analysis of variance (MANOVA), with power of .80, recommended a sample size of 30. The target sample size was then set to 60 to ensure a greater statistical power and collect a more extensive and diverse dataset of voice samples. The stressed voice data was collected when participants felt stressed giving speech in developed VR scenarios (detailed information in previous [Section 3.5.5. Development of VR prototypes for E1 \(Step1\)](#)) for our further analysis and SDM development in Phase II of research. Each participant spoke for a total of 5 minutes. In E1, participants between the ages of 18-40 and able to speak and understand English were recruited. Furthermore, the age range of participants was kept to 18-40 because individuals in this age range are usually most scared of speaking publicly and are required to be highly active and contribute to professional contexts and societal

causes (Marinho et al., 2017; Selvan et al., 2022). The participants showed no signs of motion sickness or epileptic seizures during the study.

Participants were recruited via seeking assistance from course lecturers by requesting them to showcase research posters in their respective online classes or face-to-face classes. Some participants were also recruited by word of mouth, contact connections and social media such as LinkedIn, JCUS Telegram Psychology group by sharing the social media post along with study poster. The study poster included a QR code to scan linked to a Google form. The Google form required interested individuals to fill in their email address and their availability of date and time to participate in experiment (*Appendix B Poster, and Email to JCU staff for recruiting participants and Social Media Post for Experiment 1*). Taking part in the study was completely voluntary and participants could withdraw at any point of time without prejudice (see consent form and information sheet in Appendix C).

In E1, 60 participants were recruited. From this sample, one participant did not speak anything in their speech during the experiment in one condition and thus was removed from the dataset. This resulted in final sample size of 59 (52.54 per cent female) from James Cook University, Singapore (JCUS). Participants' mean age was 24.02 ($SD = 5.18$), ranged between 19-38. Most participants were from Singapore (38.98 per cent), the rest were from China (11.8 per cent), India (10.16 per cent), Malaysia (10.16 per cent), Germany (6.77 per cent), Myanmar (5.08 per cent), Philippines (5.08 per cent), Vietnam (3.38 per cent), Indonesia (3.38 per cent), Taiwan (3.38 per cent), and USA (1.69 per cent). Half (50 per cent) of participants' mother tongue was English. 83.05 per cent of the participants who took part in the experiment were students, and the rest were full-time employed or self-employed.

3.5.6.2 Measures

Stress Arousal Check List (SACL): The SACL is a 20-item checklist measuring the feelings experienced (King et al., 1983). The SACL includes a stress subscale and arousal

subscale with 10 items for each subscale (Appendix C). Participants rated each item on a scale from 0-3; where (0) = Not at all, (1) = A little, (2) = Quite a bit, and (3) = Very much. Higher scores indicate a greater intensity of feelings. Sample items in the checklist include ‘calm’, ‘comfortable’, ‘uneasy’, ‘tense’ and ‘aroused’. The SACL has been shown to have a high internal consistency of 0.86 for the stress subscale (King et al., 1983). In the present study, the stress subscale had Cronbach’s alpha values that ranged between .71 and .84 and the arousal subscale between .68 and .71. The Cronbach’s alpha for the 9-item stress subscales of SACL was greater than .70. The Cronbach’s alpha calculated for: pretest stress subscale was .746, postpositive condition was .877, postnegative condition was .714. The Cronbach’s alpha for the 10-item arousal subscales of SACL was in the acceptable range and was between .68 and .71. The Cronbach’s alpha calculated for: pretest arousal subscale was .687, postpositive condition was .713, postnegative condition was .687.

Heart Rate (HR) & Blood Pressure (BP): HR and BP are physiological parameters that are reported to be raised under conditions of stress (Ursin et al., 2004). They can easily be measured with little inconvenience. HR and BP were measured using the OMRON Automatic Blood Pressure Monitor. The OMRON Automatic Blood pressure monitors allowed evaluation of the effect of virtual audience’s response in the VR game on the speaker’s HR and BP and manually record the readings. A separate Polar H10 HR monitor was also used to record the HR for the complete 30-minute experiment duration. This separate Polar H10 device allowed us to investigate variation in HR during the whole experiment session, and the readings could be used, if needed for further analysis. Both OMRON HEM-907 automatic BP monitor (El-Assaad et al., 2002) and Polar H10 (Speer et al., 2020) are clinically validated and for that reason were included in experiment as suitable devices for getting the respective average and continuous readings (See Devices in Appendix D).

Posttask questionnaire (a general survey): The posttask questionnaire is a survey to assess participants' personal opinion about public speaking and VR. It is a 21-item questionnaire. Items are scored in a scale of 1 to 7, where (1) = Strongly disagree, (2) = Disagree, (3) = Somewhat disagree (4) = Neutral, (5) = Somewhat agree, (6) = Agree, (7) = Strongly agree. This questionnaire included items such as 'During the speech, I was not expecting the sudden responses from the audiences', 'During the speech, I felt that I was evaluated by the audience', 'I felt pleasant about the speech I gave', 'During the speech, I had sudden feelings of uncertainty'. Ten questions related to perceived feedback from the virtual audience. The questionnaire also included 11 questions about the general experience of speaker using VR and public speaking such as 'I have tried virtual reality (VR) before', 'I usually have difficulty in giving a public speech' and it was given at the end of the speaking task in each condition to the participants. This questionnaire was used in the experiment to gather information about participants' perceived feelings regarding the use of VR for public speaking and the feedback received within the given condition (See Appendix C *Scales and Demographic Questionnaires*).

Demographic Questionnaires: The demographic questionnaire as shown in Appendix C was also included in E1. The items include age, occupation, country of birth, mother tongue and number of months remaining to completion of their studies and graduation. These questions allowed to check for the effects of age, occupation, country and first language on systolic blood pressure (SBP), diastolic blood pressure (DBP), HR, stress arousal checklist responses and posttest questionnaires of the participants.

Devices: HP OMEN gaming laptop and HTC Vive were used to play the VR scenarios in the experiment session. An iPhone 7 was used to record the participants' voice during their speeches. In practical public speaking practice, individuals typically do not rely on professional studio equipment. Using an iPhone 7 ensures the set-up is more relatable and realistic, where

individuals may use readily available devices for practice. The chosen approach also allows us to collect voice samples that closely resemble everyday public speaking practice sessions. As a result, the dataset becomes more realistic and representative of genuine experiences, rather than imitated performances (*Appendix D Devices Used*).

3.5.7 Procedure

An approval (HREC Approval Number: H8298) to conduct the experiment was obtained from Human Research Ethics Committee of the James Cook University prior to commencement (Approval letter from HREC in Appendix E). The experiment was conducted using a within-subject design. Each participant went through both the NP and NN conditions. The order of both conditions was randomised for each participant.

First, after the arrival of participants, 5 minutes were allocated to introduce the project, task, and devices (Polar H10 HR monitor, OMRON Automatic BP monitor HEM – 907, iPhone 7, HP OMEN gaming laptop, and HTC Vive) to be used and how the devices will be used in experiment.

After an initial introduction to the experiment, the participant read the information sheet and provided consent for voice to be recorded, to give a public speech, filling in of the questionnaires, recording of the physiological parameters of HR and BP.

Next, participants were requested to wear the Polar H10 HR monitor around their chest as per instructions and its operability was checked. After this, the participant was asked to be seated in the right posture for the experiment on a chair by keeping their left arm straight on the table, thus allowing to put the OMRON Automatic Blood pressure monitor's cuff on the participant's left arm. After waiting for 5 minutes, the first baseline reading of HR and BP (systolic and diastolic) were taken using the OMRON Automatic BP monitor. Next, the SACL was used to measure the participant's stress and arousal.

Following this, the Polar H10 was set to on. After recording the pretest readings for baseline SBP, baseline DBP, baseline HR and recording the baseline responses in the SACL, the participant was then randomly assigned to any one of the given scenarios of NP or NN condition. The experimenter then highlighted that the public speaking task the tester was taking part in was designed to assess certain psychological capabilities. The experimenter highlighted that research has shown that public speaking including verbal intelligence, tone of voice, breathing pattern, poise, and overall performance on these tasks predicts how successful a person will be in their relationships or occupation, whatever that may be. The participant was then randomly assigned to give a speech on the topic of either ‘online vs offline classes’ or ‘why an employer should hire you’. The experimenter emphasised that the participant can choose to say anything about the given topic, and it is completely up to the participant. However, the participant needs to speak for the entire duration of 2 minutes and 30 seconds clearly. The participants were then given 2 minutes and 30 seconds of time to prepare the points they wanted to say about the chosen topic. Following this, the experimenter took a back seat in the same room. No paper and pencils were given unless asked to write their points down during speech preparation time. This allowed participants to clear their minds about specific points they were going to make in their speech and ensured them of speaking for the full duration of 2 minutes and 30 seconds. However, participants didn’t have access to any of their points in writing to look at while speaking as they were wearing the VR HTC Vive headset.

Once the speech preparation timer went off, and the participant was ready, the experimenter provided the HTC Vive VR headset and helped in fixing it properly with the participant. Next, the experimenter then started the VR simulation condition in the HP OMEN gaming laptop and asked the participant whether the VR scenario could be seen in the HTC Vive VR headset view and audio hearing were tested. After everything was set in motion and

participant agreed to being fully ready to speak, the VR simulation scenario was on. The experimenter set the voice recording in the iPhone 7 to on.

The participant spoke either firstly in NP condition) followed by secondly speaking in NN condition) or vice versa or the participant spoke in NN condition) first and then in NP condition) second.

In a scenario where the participant went through the NP condition first, the virtual audience remained neutral for 30 seconds (no smiling, with only sitting quietly, looking around and at speaker). This was then followed by applauding (in around 80 seconds, 120 seconds and 180 seconds into the simulation).

The second reading of the participant's SBP, DBP and HR reading are taken at the first 30 seconds in the condition speech in the allotted VR simulation scenario (when the virtual audience appeared visually neutral [N]), followed by third reading of same parameters at around 80 seconds (when the virtual audience appeared visually positive by applauding [P]), and fourth reading at 130 seconds (again when applause was played [P]).

After the participant finished speaking in the first condition, the iPhone recording session was stopped, and the recording was saved. The HTC Vive VR headset was then removed from the participant's head. Next, the participant was given the SACL and a questionnaire to complete as a posttest. Each participant took approximately 4-7 minutes to complete this sequence.

Once the participant has completed speaking in the first condition (NN or NP depending on the case), the participant is then told about the second speech task in the NP condition if the first condition was the NN condition). Again, the topic of speech is then given to the participant to speak on ('why an employer should hire you' or 'online vs offline classes'), followed by preparation time of 2 minutes and 30 seconds. Once the participant was ready, the same procedure was followed with the alternative VR simulation. The participant then delivers a

speech in this alternative VR simulation for the next 2 minutes and 30 seconds while their voice was recorded. In the NN condition), the virtual audiences start with neutral feedback for initial 30 seconds into the NN condition (no smiling, with only sitting quietly, attentively looking around, and a subtle nodding at speaker), slowly becoming inattentive (i) by picking up and looking at their phone (80 seconds into the NN condition), (ii) chatting on their phone (at 120 seconds into the NN scenario), and then finally providing negative feedback by booing (seconds into the NN scenario) to express approval or disapproval of the speaker. Different subsets of the virtual audience provide different feedback.

The fifth and sixth SBP, DBP and HR readings were taken at around 72 seconds and 130 seconds (both readings are taken when the participant is speaking in the NN condition when booing and other critical feedback responses are provided in the scenario [B]). There was a gap of around 30 seconds between each behaviour by the virtual audiences in each condition. Once the participant completed their speech in this final condition, the experimenter ended the voice recording on the iPhone, and removed the VR headset, followed by participants completing the SACL and another posttest questionnaire for this condition.

During the time when a participant gave speech, SBP, DBP, and HR readings were measured five times, for each instance of applause, booing and neutral behaviour from the audience in both conditions. Combined with one baseline reading, a total of six readings were taken for HR and BP. The voice of each participant giving their speech was recorded in both conditions (as a third dependent variable to be used for further analysis). Participants filled in SACL before taking part in each condition to measure the perceived stress and arousal and responded to SACL and about the task questionnaire at the end of their speech in each condition to check perceived stress and experience speaking in different conditions with varied set of virtual audience's behaviour. For filling in of the SACL, a posttest and demographic questionnaires, Qualtrics (2018) was used. Next, a set of demographic questionnaires was

administered towards the end of the experiment. After the completion of the study, participants received SGD25 Grab e-voucher via email or physical SGD 25 NTUC FairPrice voucher on the spot.

3.6 Results

3.6.1 Data Preparation

The baseline readings (Baseline_SBP, Baseline_DBP, Baseline_HR) of the systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) were computed by calculating the average scores. These average baseline scores were obtained by carrying out the average of the first baseline reading when the participant first came to the experiment and second reading taken during the neutral response [N] of the first given condition (NP or NN depends) in which participant spoke.

The responses to positive feedback (P_SBP, P_DBP, P_HR) for the same physiological parameters were computed by calculating the reactivity scores. The reactivity scores were calculated by computing an average of the two readings recorded when the person gave the speech in the NP condition when the participant was applauded at 80 seconds and 130 seconds (the third and fourth readings) and then subtracting it from average of the baseline score.

Similarly, the responses to negative feedback (N_SBP, N_DBP, N_HR) for the same parameters were obtained by calculating the reactivity score. This was done by taking an average of the physiological parameter's readings recorded during the speaker giving speech in the VR scenario with the virtual audience showing NN feedback at 72 seconds and 130 seconds (the fifth and sixth readings) and subtracting it from the average of the baseline scores.

The scores for the stress and arousal checklist items for the baseline normal readings (Pre_stress, Pre_arousal), as well as during both the NP (P_stress, N_arousal) and NN condition (N_stress, N_arousal) were computed by taking the average of the stress subscale items and

arousal subscale items. The stress subscale contained nine items, and the anxiety subscale contained 10 items.

For the stress and arousal readings, the scores for the subscales were reversed as per the need. Then an overall average was computed for both the stress and arousal subscale for both the conditions.

3.6.2 Assumption Testing

All results were analysed using SPSS version 26, with the alpha level set at 0.05. Before analysing the data, several assumptions were addressed during the experiments, and the remaining assumptions were evaluated before conducting the final analysis using a one-way repeated measures MANOVA. To meet the assumption of independence of errors, each participant took part only once in the experiment and was instructed not to share details with others.

The assumption of normality was examined using the Shapiro Wilk statistic and visual inspection of histograms. Results showed that the normality for physiological parameters (Baseline_SBP, Baseline_DBP, Baseline_Heartrate, P_SBP, P_DBP, P_HR, N_SBP, N_DBP, N_HR) and computed average scores of stresses (Baseline_stress, P_stress, N_stress) and arousal (Baseline_arousal, P_arousal, N_arousal) from the SACL subscale items was met except for the pretest computed average stress (Baseline_Stress), baseline systolic BP (Baseline_SBP), baseline diastolic BP (Baseline_DBP), positive baseline diastolic BP (P_SBP) and positive computed average stress (P_Stress).

The extent of these violations was minor, as visual inspections of histograms showed approximately normal distributions, and no severe outliers were observed. Additionally, the degrees of freedom for the error term exceeded 20, providing some robustness against normality violations. Despite these minor deviations, MANOVA analysis is known to be robust

against violation of the normality assumption when sample sizes are sufficient (Fidell et al., 2003). Therefore, no data transformation was performed.

Second, the absence of univariate outlier assumption was checked using boxplots. The boxplots results indicated no severe outliers for dependent variables. Third, the Mahalanobis Distance coefficient was 16.80. This value was smaller than the critical chi-square value, $\chi^2(5) = 20.51$, at an alpha level of .001, implying that the assumption for absence of multivariate outliers was met. Fourth, the inspection of the scatterplot between each of the dependent variables indicated the assumption of linearity was met. Multicollinearity between dependent variables was not of concern by assessing the correlation table obtained from the regression analysis. The sphericity assumption was assumed for SBP, DBP and HR.

3.6.3 Preliminary Analysis

Preliminary analyses were carried out to check any effect of the variables such as age, occupation, gender, first language, country of birth, no. of months to graduate and the posttest questionnaire have on dependent variables (DVs) obtained across both conditions. The DVs are Baseline_SBP, Baseline_DBP, Baseline_HR, Baseline_stress, Baseline_arousal, P_SBP, P_DBP, P_HR, P_stress, P_arousal, N_SBP, N_DBP, N_HR, N_stress, N_arousal.

To assess the effect of age on DVs, a bivariate Pearson's product moment correlation coefficient (r) was calculated. The bivariate correlations between the age and all the DVs were not significant (Table 3.2 Correlation Results).

An independent samples t -test was used to compare the effect of gender on the DVs across both NP and NN conditions. All results were not significant. However, the effect of gender was significant on baseline systolic blood pressure (Baseline_SBP), where male had higher baseline_SBP in comparison to female participants.

Furthermore, multiple one way between-subjects analyses of variance (ANOVAs) (see Table 3.3 t -Test and ANOVA Results) were carried out. The first ANOVA was used to

investigate the effect of country in which participants were born on the same five DVs. The ANOVAs were statistically not significant, indicating that all the DVs were not influenced by the country of the participants, ($F_s < 2$, $ps > .05$), except the baseline_SBP ($F_s = 2.14$, $ps < .05$) and N_SBP ($F_s = 2.13$, $ps < .05$).

The second set of one way between-subjects ANOVAs was used to investigate the effect of first language spoken by participants and stress, arousal, HR and systolic BP and diastolic BP in positive and negative conditions. The ANOVA was statistically nonsignificant, indicating no effect of first language spoken by participants on all the DVs, except for Baseline_DBP ($F_s = 2.21$, $ps < .05$).

To investigate the effect of participant's occupation on the DVs, final and last set of one-way ANOVAs was carried out. The results of ANOVA were nonsignificant ($F_s < 2.50$, $ps < .05$).

Lastly, two separate correlations tests were conducted with 'about the task questionnaire' (posttest questionnaire) items across the responses obtained from both NP (Results in Table 3.4) and NN conditions (Results in Table 3.5), in order to identify covariates. Six covariates were identified from the posttest questionnaires but they were not included in the final hypothesis testing of MANOVA as they had very minor correlations with only very few DVs.

Table 3.2 *Correlation Results*

Pearson's correlation	Baseline SBP	Baseline DBP	Baseline HR	Baseline Stress	Baseline Arousal	P_SBP	P_DBP	P_HR	P_Stress	P__Arousal	N_SBP	N_DBP	N_HR	N_Stress	N_Arousal
Age	.25	.08	-.12	-.15	.06	-.03	.09	.02	-.15	.09	-.01	-.02	-.08	-.25	.17
No. of Months to Course Completion	.243	-.08	-.06	.03	-.13	-.12	.04	.29*	-.03	.10	.08	.19	.27	-.15	.06

Note. This table demonstrates Pearson's correlation coefficients of Age and Number of Months to Course Completion with various physiological and psychological measures at baseline (Baseline), during the positive condition (P), and during the negative condition (N). Measures include SBP, DBP, HR, stress and arousal levels.

Correlation coefficients marked with an asterisk (*) are statistically significant at $p < .05$.

* $p < .05$.

Table 3.3 *T-Test and ANOVA Results*

<i>F</i> -statistic	Baseline SBP	Baseline DBP	Baseline HR	Baseline Stress	Baseline Arousal	P_SBP	P_DBP	P_HR	P_Stress	P_Arousal	N_SBP	N_DBP	N_HR	N_Stress	N_Arousal
Country of Birth	2.14*	1.58	1.27	.76	.47	.75	.63	.89	.64	1.17	2.13	1.02	1.14	.87	1.66
First Language	1.50	2.21*	.98	.72	.84	.54	.61	1.69	.94	.72	1.05	.98	1.63	1.49	1.54
Occupation	.22	.69	.82	.90	.16	2.49	1.17	.69	.44	.90	.58	.90	.18	.33	.95
Gender	6.11*	1.18	.34	.14	2.25	.65	.36	2.32	4.47	.37	.50	1.04	2.46	1.12	1.25

Note. This table shows *F*-statistics from ANOVA tests comparing physiological and psychological measures (SBP, DBP, HR, stress, and arousal) at baseline (baseline), during the positive condition (P), and during the negative condition (N) across different demographic factors: Country of Birth, First Language, Occupation and gender.

F-statistics marked with an asterisk (*) are statistically significant at $p < .05$.

* $p < .05$

Table 3.4 *Correlation Results for Posttest Questionnaire Items for NP condition*

Item/Pearson's Correlation	P_SBP	P_DBP	P_HR	P_Stress	P__Arousal	N_SBP	N_HR	N_Stress
P_Q1	.01	.08	.15	-.41**	.30*	.158	.17	-.17
P_Q2	-.07	.01	-.01	.14	.14	-.09	-.02	.28*
P_Q3	-.02	.08	.18	-.03	-.08	.10	.15	-.07
P_Q4	.05	.15	.17	-.34**	.25	.33*	.18	-.34**
P_Q5	-.06	-.02	-.08	.00	.02	-.13	.06	.08
P_Q6	-.10	.03	.01	.04	.14	-.04	.21	-.07
P_Q7	-.10	-.28*	-.05	-.05	-.02	-.26*	-.02	.12
P_Q8	.18	.10	-.13	.05	.13	.01	-.12	.21
P_Q9	-.06	-.07	-.17	.21	-.06	-.29*	-.13	.21
P_Q10	-.10	-.03	-.00	.09	-.20	-.06	.05	-.10
P_Q11	-.23	.00	-.11	-.08	.01	-.09	-.01	-.03
P_Q12	-.03	.01	-.07	.12	.12	-.06	-.00	.22
P_Q13	.04	.27*	.32*	-.36**	.40**	.15	.28*	-.07
P_Q14	-.17	-.18	-.15	.43**	-.16	-.22	.00	.44**
P_Q15	-.01	-.05	-.17	.35**	-.03	-.28*	-.12	.35**
P_Q16	.04	.02	.17	-.44**	.26*	.01	.09	-.16
P_Q17	-.37**	-.17	-.23	.37**	.22	-.19	-.04	.94
P_Q18	-.18	-.21	-.25*	.31*	-.24	-.38**	-.25	.30*
P_Q19	-.07	-.02	-.07	.19	-.02	-.20	-.06	.34**
P_Q20	-.17	.45	-.03	.38*	-.18	-.20	-.03	.17
P_Q21	-.25	-.01	-.09	-.20	.13	-.13	-.06	.14

Table 3.5 *Correlation Results for Posttest Questionnaire Items for NN Condition*

Item/Pearson's Correlation	Baseline DBP	Baseline HR	Baseline Stress	P_SBP	P_HR	P_Stress	P__Arousal	N_SBP	N_DBP	N_HR	N_Stress
N_Q1	-.10	-.19	-.40*	.04	.19	-.40**	.24	.33**	.17	.19	-.50**
N_Q2	.11	.04	-.13	.05	-.06	.05	.18	-.04	-.01	.02	.16
N_Q3	-.11	-.20	-.09	-.11	.08	-.13	.09	.01	-.06	.14	-.14
N_Q4	-.10	-.18	-.31*	.05	.15	-.36*	.34**	.35**	.34**	.25	-.31*
N_Q5	.14	.04	-.04	-.21	-.25	.77	.32	-.03	.05	.05	.10
N_Q6	-.04	-.12	-.16	.05	.11	.12	.05	.15	.02	.10	.07
N_Q7	.06	-.10	-.06	.10	.18	-.02	.17	.00	-.10	.11	.04
N_Q8	.32*	.26**	-.02	.09	-.29*	.03	.01	-.10	-.12	-.27*	.16
N_Q9	-.09	.22	.10	-.03	-.17	.24	-.11	-.28*	-.22	-.17	.34**
N_Q10	-.29*	-.07	-.11	.00	.85	.10	-.19	.05	.05	.07	-.81
N_Q11	-.14	-.07	-.06	.18	.00	.14	.14	.18	.14	-.11	.00
N_Q12	-.03	.00	-.13	-.28*	-.34**	-.06	-.05	-.15	-.13	.02	-.01
N_Q13	-.16	-.14	-.42**	.06	.12	.40**	.16	.15	.04	.05	-.56**
N_Q14	-.06	.14	.22	-.04	-.54	.38*	-.10	-.19	-.17	-.08	.51**
N_Q15	-.21	-.11	.10	-.09	-.04	.16	.83	-.04	.03	.02	.26*
N_Q16	-.11	-.13	-.25	.03	.14	-.33**	.11	.11	.03	.09	-.41**

N_Q17	-.12	.02	-.04	-.28	-.35	.19	.22	-.21	-.09	-.15	.33
N_Q18	.22	.19	.28*	-.27*	-.32*	.33**	.01	-.25*	-.15	-.15	.55**
N_Q19	.04	-.10	.05	.03	-.11	.23	.23	.10	-.03	.01	.33*
N_Q20	.05	.00	.07	.03	.03	.23	-.19	-.11	.06	-.09	.47*
N_Q21	-.13	.07	.00	-.16	-.02	-.01	.09	-.12	-.11	.00	.00

Note. This tables presents Pearson's correlation analyses between various psychological and physiological responses during the positive (NP) and negative (NN) VR conditions and responses to specific questions (Q1-Q21) in the posttest questionnaires. The correlations indicate relationships between participants' self-reported experiences and their physiological and psychological states during the task in the given condition.

The correlations related to difficulty giving speech (Q9), feeling threatened (Q14), felt challenged (Q15), control over situation (Q16), feeling being evaluated by audience (Q17), feeling uncertainty (Q18), trying to avoid embarrassment (Q19), trying to complete task as soon as possible (Q20) were consistently higher in the negative VR condition compared to the positive condition. The stronger correlation in the NN condition indicates greater influence of these negative psychological factors on stress responses when participants face unsupportive and adverse audience behaviours. These findings reinforce the conclusion that negative VR environments significantly heighten stress.

* $p < .05$, ** $p < .01$

3.7 Main Results: Hypothesis Testing

Two separate repeated measures mixed MANOVA analyses were carried to determine the effects of two distinct VR conditions on speaker's stress levels measured via physiological parameters (reactivity scores of (i) SBP, (ii) DBP, (iii) HR and self-reported questionnaire responses (average scores of (i) stress and (ii) arousal)). The two distinct VR scenarios were with NP and NN behaviour of the virtual audience.

The MANOVA analyses were only carried out after assumption and preliminary analyses requirements were met.

3.7.1 Physiological Measures

The first MANOVA was used to examine the effect of both the NP and NN VR conditions on the reactivity scores of physiological parameters SBP, DBP, HR. The analysis was done without including any covariates.

Findings showed that there was a significant effect of the VR conditions (giving speech in NP and NN conditions) on all DVs combined, $F(3, 57) = 4.49, p = .007$, partial $\eta^2 = .194$. Analysis of the univariate tests revealed significant differences between the two VR scenarios in SBP reactivity ($p = .009$) and HR reactivity ($p = .001$). However, there was no significant difference in DBP reactivity ($p = .262$) between the two conditions.

To further compare the individual differences in DVs, a final look at the descriptive statistics table revealed that the SBP reactive score in the NN condition ($M = 11.60, SD = 11.17$) was significantly higher than in comparison to the SBP reactive score in the NP condition ($M = 7.82, SD = 10.57$). The DBP reactive score in the NP condition ($M = 11.78, SD = 8.93$) was higher than the DBP reactive score in the NN condition ($M = 10.49, SD = 8.60$). Lastly, HR reactivity score in the NN condition ($M = 9.20, SD = 9.88$) was higher in comparison to the HR

reactivity score in the NP condition ($M = 5.54$, $SD = 7.79$). Effect sizes for these comparisons were 0.34, 0.14, and 0.41. (Summarised in Table 3.6).

Table 3.6 *Mean and Standard Deviations: Physiological Measures*

DV's	Independent Variables conditions				Effect Size
	NP		NN		
	Mean	SD	Mean	SD	
SBP	7.82	10.57	11.60	11.17	0.34
DBP	11.78	8.93	10.49	8.60	0.14
HR	5.54	7.79	9.19	9.88	0.41

3.7.2 Self-Reported Questionnaire Responses

The second one-way repeated measures analysis (MANOVA) was conducted to examine the effect of two different VR conditions [(NP and NN)] on the stress and arousal items answered by participants in the SACL questionnaire.

Sphericity was assumed for stress and arousal, and a further look at the univariate tests table indicated there was significant difference in the participant's ratings of the stress average score ($p = .00$) and arousal average score ($p = .00$) in the two VR scenarios.

Findings showed that there was a significant effect of the VR conditions on speakers giving speech in NP and NN conditions on combined each dependent variable, $F(2, 57) = 380.23$, $p = 0.00$, partial $\eta^2 = .93$.

Furthermore, pairwise comparisons indicated that the stress average score in NN condition ($M = 1.52$, $SD = .45$) was significantly higher than in comparison to the stress average

score in NP condition ($M = 1.12$, $SD = .59$). The arousal average score in the NN condition ($M = .85$, $SD = .18$) was significantly lower than in comparison to the arousal average score in the NP condition ($M = 1.85$, $SD = .397$). (Results summarised in Table 3.7) The effect sizes for these comparisons were -0.75 and -3.20.

Table 3.7 *Mean and Standard Deviations: Self-Reported Questionnaire Responses*

DV's	Independent Variables conditions				
	NP		NN (NN)		Effect size
	Mean	SD	Mean	SD	
Stress	1.12	.59	1.52	.45	-0.75
Arousal	1.85	.397	.85	.18	-3.20

3.8 Discussion

The main aim of this experiment was to collect voice samples in both VR conditions and examine the stress arousal capabilities of the developed supportive and nonsupportive VR game scenarios. The hypothesis tested was that the audience's scripted negative behaviour (booing, talking on the phone, and inattentive listening) would induce higher levels of stress (measured via physiological parameters and self-reported questionnaire responses) compared with positive behaviour (applause, attentive listening, and nodding) from the virtual audience.

The results supported the hypothesis that the negative behaviour of the virtual audience in the VR game scenario would induce higher stress levels in the participants compared to the positive behaviour. Specifically, the statistical analysis of physiological parameters revealed that HR and SBP were significantly higher in the NN condition than in the NP condition. Additionally, only DBP measurements were lower in the NN condition compared with the NP

condition. These results are consistent with previous research (Droppleman & McNair, 1971; Lynch et al., 1982; Tardy et al., 1989; Thomas et al., 1984).

The analysis carried out on the self-reported questionnaire responses of stress arousal check list's items (SACL) indicated significant difference between the stress and arousal levels recorded in both the NP and NN conditions. Specifically, stress levels in the NN condition were higher in comparison to stress felt by participants in the NP condition. The SACL questionnaire results thus supports the result analysis regarding stress detected physiologically. However, the arousal levels in the NN condition were lower in comparison to arousal in the NP condition.

While stress and arousal are often positively correlated, the relationship is context-dependent and not always linear. The relationship between stress and arousal suggests that it is inversely proportional within a given condition. In this study, participants in the negative condition experienced high stress and lower arousal. Intuitively, this could be explained in terms of when a speaker feels less active and energetic during a public speaking task, the experience of stress becomes higher. It could be due to the nature of the feedback provided in negative condition where participants might have experienced stress without positive engagement or excitement, due to negative feedback provided to the speaker, leading to a lower arousal state.

Conversely, when a person feels more energetic and more active to deliver their speech in a positive way, they tend to experience less stress. Participants in the positive condition showed lower stress levels and higher arousal. This might be due to experience of less stress because of positive feedback provided to the speaker, resulting in high arousal showing positive engagement or excitement. Finding highlights that stress and arousal, while related, can manifest differently depending on the emotional and motivational context in an environment.

The implications of these findings are important for understanding the dynamics of stress and arousal in VR settings. By demonstrating that negative virtual audience behaviour

can significantly increase stress levels consistent with other research (Byron et al., 2023) and positive virtual audience behaviour reduces the stress levels, this study highlights the potential for VR environments to be used in stress induction and management research. The development of a stress voice database, as discussed in the implications section, provides a valuable resource for further investigation into stress detection and analysis through voice (which will be covered in the next two chapters).

In addition to contributing to academic research, these findings have practical applications in fields such as public speaking training, VR therapy and stress management. VR environments can be designed to simulate high-stress scenarios, allowing individuals to practice and develop coping strategies in a controlled setting. The integration of real-time stress detection and feedback systems, as proposed in subsequent chapters, could enhance the effectiveness of these VR-based interventions.

This study advances our understanding of how VR can be utilised to study and manage stress, paving the way for innovative applications and further research in this area. The evidence that VR can effectively induce stress through audience behaviour underscores its potential as a powerful tool for both research and practical applications in stress management.

3.9 Theoretical and Practical Implications

The present experiment has provided new insights into the stress arousal capabilities within the specific VR conditions developed for this study. While previous research focused on inducing and studying fear and anxiety in VR simulations related to public speaking, this study specifically demonstrated that stress can be effectively induced in speakers delivering speeches in VR settings. Notably, stress levels were found to be higher in negative conditions compared with positive conditions. These stress levels were measured using HR, SBP, DBP and self-reported stress arousal values.

The resulting voice samples were analysed to identify stress-indicative features, leading to the creation of a stress voice database. This database represents a significant advancement, as it provides a valuable resource for future studies aiming to investigate and understand stress through voice analysis.

Chapter 4 will focus on identifying stress voice samples to refine and enhance the stress voice database. This refined database will serve as the foundation for developing a stress detection ML algorithm. The goal is to integrate this algorithm into the VR system, enabling real-time feedback on the speaker's stress levels. Such integration holds promising implications for various applications, including improving public speaking training, enhancing virtual therapy sessions and providing real-time stress management support in VR environments. The evaluation of this integrated system will be thoroughly discussed in Chapter 5, assessing its effectiveness and potential for broader implementation.

3.10 Summary

The goal of Chapter 3 was to provide an in-depth overview of the research methods used in Phase I. The primary aim of Phase I of the research included conducting Experiment 1 which was designed to create this stress voice database and address Research Question 1 (RQ₁) by testing the derived hypotheses (H1a and H1b). The chapter details the development of the VR game prototypes used in Experiment 1, emphasising their role in validating the scenarios in forming the stress voice database. It covers the procedures and methods of the study, including participant recruitment, data collection and ethical considerations regarding the experiment. The results were analysed through assumption testing and hypothesis testing, offering insights into the stress arousal capabilities of the VR scenarios.

In Chapter 4, the focus will shift to the development of a SDM using the collected voice samples from this Phase I. The goal is to build on the findings of Experiment 1 and further

explore the potential of implementing VR environments for stress detection via voice analysis-based techniques and utilising ML algorithms.

Chapter 4: Development of Voice Analysis Based Stress Detection Model

Chapter 4 focuses on developing a ML model for stress detection based on voice parameters, with the primary aim to identify the best voice features indicative of stress levels in speakers. To achieve this, a specialised voice database was created in the context of public speaking in order to answer Research Question 2 (RQ₂): What reliable set of voice features (e.g., prosodic, spectral) can be analysed in real-time (established in correlation to the physiological parameters e.g., BP, HR) from recorded voice to detect stress level in a speaker? (as outlined in Chapter 1 under [Research Objectives and Questions](#)).

The chapter is divided into two main sections. The first section involves preliminary work with the publicly available Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). An ethics application was submitted to include this database in the research. Preliminary work involved testing the initial database, familiarising with the feature extraction process and building initial models. Using the RAVDESS database, the first section aims to develop a model for affective state detection by extracting voice features that distinguish between calm and stress in the labelled samples. Various supervised and unsupervised ML algorithms were employed to build these models.

The second section describes the process for utilising voice samples from Experiment 1 (E1) to create a stress voice database. Relevant voice features indicative of stress were extracted from these samples. Once the stress indicative voice features were identified, ML algorithms were used to build an affective state model for integration into a VR application. The chapter discusses the findings and results in relation to the defined aims of identifying accurate voice features and development of SDM.

4.1 Preliminary Work

4.1.1 Database

The preliminary work was dedicated to exploring the voice feature identification procedure, in order to develop an initial model for the detection of stress in real-time during public speaking. An informed decision was made to utilise and work with the RAVDESS database (Livingstone & Russo, 2018). The database contains 24 professional actors (12 males and 12 females) speaking two distinct statements that differ in lexical content, each spoken in seven different emotions (calm, happy, sad, angry, fearful, surprised and disgusted). Each speech expression was produced with a normal, strong and neutral level of emotional intensity. A total of 1440 available speech files were downloaded, consisting of 60 trials per actor. Although the primary aim was not to classify all the original discrete emotional expressions (calm, happy, sad, angry, fearful, surprise, and disgust) in the database, the focus was placed on the neutral state (i.e., calm) and an altered emotional state (i.e., fear). Files corresponding to these two emotions across two levels of intensity (neutral and strong) were selected for all 24 actors' voice samples.

The voice files selection from the RAVDESS database was adopted considering the context of public speaking, where speakers can experience a range of affective states and emotions such as anxiety, fear, enthusiasm, confidence, calmness and stress. Two primary emotional states were selected for all actors: calm, representing a relaxed and confident state, and fear which are closely associated with stress and anxiety in public speaking contexts (Juslin & Laukka, 2003). Two voice clips (each around 3 seconds) per actor were selected representing a relaxed and confident state (calm) and the other associated with stress and anxiety (fear) in public speaking contexts. This resulted in a balanced binary classification setup across 24 actors. The rationale for this selection was based on aim and interest in distinguishing between a neutral (i.e., calm) state and an altered emotional state (i.e., fearful). For this purpose, the

normal emotional intensity of fear was considered similar to stress, while the calm state was indicative of confidence. This approach allowed to effectively create a classification for stress detection in preliminary work with the RAVDESS database. Neutral intensity files were chosen because they represent the natural or baseline state of the emotion, reflecting the affective state experienced by the speaker. The files with strong intensity were selected to capture the highest energy level of the speaker while expressing the emotion, thus indicating a significant emotional affective response.

4.1.2 Feature Extraction

For the preliminary work, the LibRosa Python package (McFee et al., 2015) was used to analyse the features of the respective confident and stressed states for all 24 actors.

Informed by the findings from the literature review detailed in Chapter 2, particularly the section titled ‘Voice Features and Emotion Detection’, the research began with the extraction of both time-domain and frequency-domain features, which have been identified as effective for identifying affective states from voice samples. From the time domain, amplitude envelope (AE), root mean square (RMS), and zero crossing rate (ZCR), were extracted. These features were selected based on their ability to capture variations in loudness, energy and frequency content that are indicative of stress and high arousal, as supported by previous research. Next, frequency-domain features such as spectral centroid (SC), spectral bandwidth (SB) and Spectral Rolloff (SR) were extracted. These features help in understanding the distribution of energy across different frequencies in the voice signal, which can vary significantly under stress. Lastly, spectral features, specifically Mel-frequency cepstral coefficients (MFCCs, all 40 coefficients), which are widely recognised for their effectiveness in capturing the nuances of speech characteristics and have been frequently used in detection models were extracted as well (refer to the literature review and Table 1 of Chapter 2). All features were extracted using a frame size of 1024 samples and a hop length of 512 samples

corresponding to ~23 ms frames with 50% overlap. These parameters (frame size and hop length) were applied uniformly across all audio files in the RAVDESS dataset for feature extraction process which included both confident and stressed voice samples for 24 actors. All features were computed per frame over the entire signal. Each frame was labelled according to the emotional class of the clip it was extracted from and these frame-wise values were used directly without aggregation (i.e., not summarized into mean or standard deviation per clip) for model training. Furthermore, the decision to use the framewise feature extraction and classification was guided by the short duration (approximately 3 to 4 seconds) of the audio clips. Aggregating features at the clip level would have resulted in a small dataset which would have been insufficient for the training of even simple models. The audio segmentation with overlapping frames (1024 samples per frame with a 512-sample hop, ~23 ms per frame) allowed each frame to serve as an independent training instance and increased the number of training samples. Given the emotional state consistency within these short utterances, labelling each frame with the overall emotion of the clip was considered valid. This approach provided a robust foundation for carrying out the preliminary analysis and gain understanding with a public database for developing an affective state detection model.

4.1.3 Hypothesis

Following the extraction of features from the 24 actors' voice samples in the RAVDESS database, a t-test assuming unequal variance for two samples was conducted for each extracted voice feature. The analysis was guided by the below formulated hypothesis and specified rejection criteria.

H₂: There is significant difference between respective actor's normal calm (confident) and normal fearful (stressed) voices for the extracted feature.

Rejection Criteria: Reject H₂ if p-value < 0.05.

4.1.4 Preliminary Results

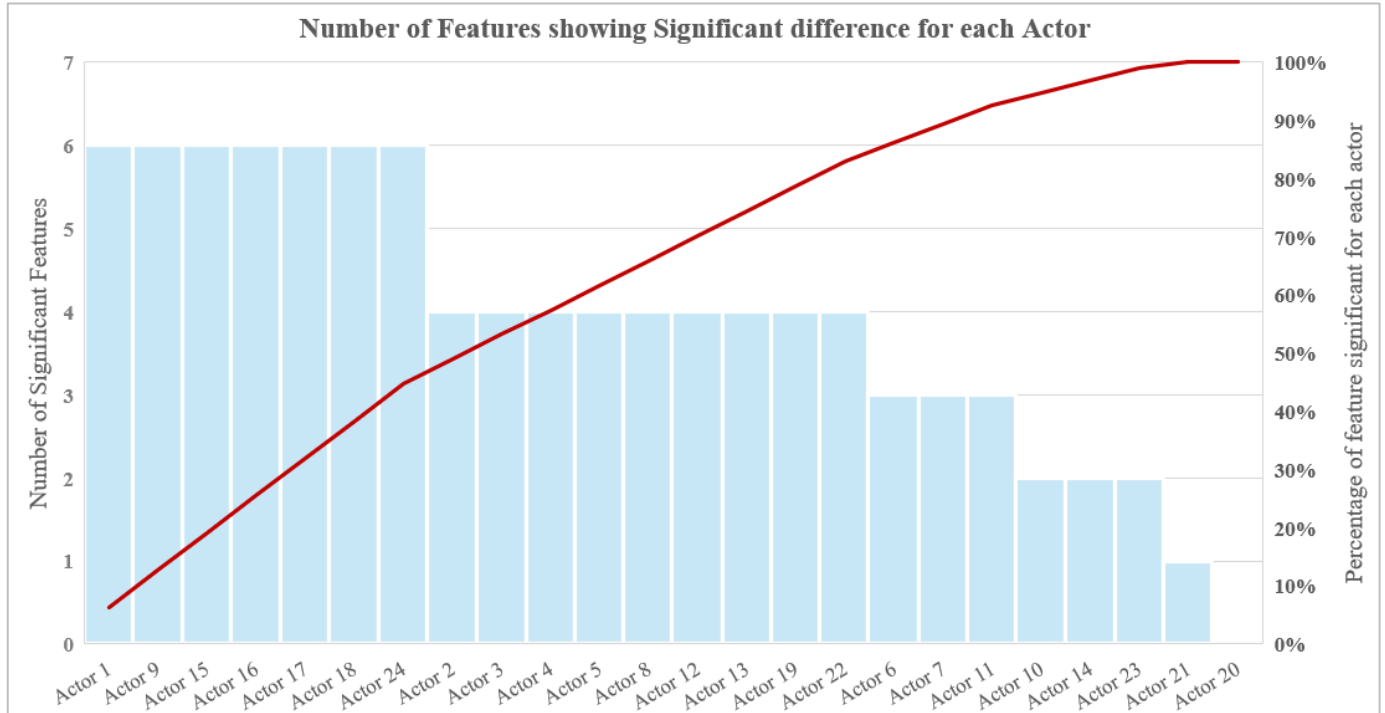
Preliminary work shown in Figure 4.1 uses tick mark (✓) to indicate significant difference ($p < .05$) between respective actor's normal calm (confident) and normal fearful (stressed) voice's extracted features and wrong mark (✗) show nonsignificant difference ($p > .05$) for various extracted voice features.

Figure 4.1 *Overall significant and nonsignificant values with Preliminary dataset from RAVDESS*

where AE = Amplitude Envelope, RMS = Root-Mean Square, ZCR = Zero-Crossing Rate, SC = Spectral Centroid, SB = Spectral bandwidth, SR = Spectral Rolloff and MFCCs = Mel-Frequency Cepstral Coefficients (all 40 coefficients).

	AE	RMS	ZCR	SC	SB	SR	MFCCs
Actor 1	✓	✓	✓	✗	✓	✓	✓
Actor 2	✓	✓	✗	✗	✗	✓	✓
Actor 3	✗	✗	✗	✓	✓	✓	✓
Actor 4	✓	✓	✗	✗	✓	✗	✓
Actor 5	✓	✓	✓	✗	✗	✗	✓
Actor 6	✗	✓	✗	✗	✗	✗	✓
Actor 7	✗	✗	✓	✓	✗	✓	✗
Actor 8	✓	✓	✗	✗	✓	✗	✓
Actor 9	✓	✓	✓	✓	✗	✓	✓
Actor 10	✗	✗	✗	✗	✓	✗	✗
Actor 11	✓	✓	✗	✗	✗	✗	✓
Actor 12	✓	✓	✓	✗	✗	✗	✓
Actor 13	✓	✓	✗	✗	✗	✓	✓
Actor 14	✗	✗	✓	✗	✗	✗	✓
Actor 15	✗	✓	✓	✓	✓	✓	✓
Actor 16	✓	✓	✓	✓	✓	✗	✓
Actor 17	✓	✓	✓	✗	✓	✓	✓
Actor 18	✓	✓	✓	✗	✓	✓	✓
Actor 19	✓	✓	✗	✓	✗	✗	✓
Actor 20	✗	✗	✗	✗	✗	✗	✗
Actor 21	✗	✗	✗	✗	✗	✗	✓
Actor 22	✓	✓	✓	✗	✗	✗	✓
Actor 23	✗	✗	✗	✗	✓	✗	✓
Actor 24	✓	✓	✗	✓	✓	✓	✓

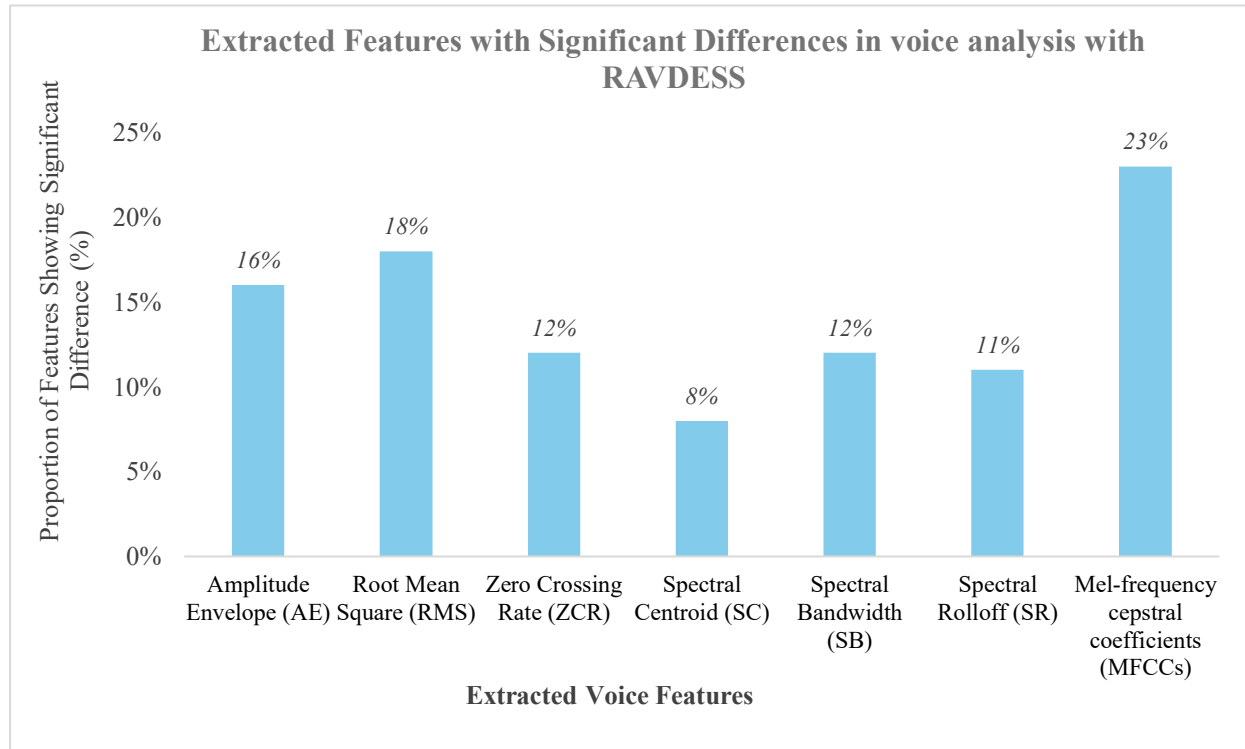
Results (Figure 4.2) indicate that Actor 1, Actor 9, Actor 15, Actor 16, Actor 17, Actor 18 and Actor 24 are a few of the actors in the dataset who showed significant differences between their normal calm (confident) and normal fearful (stressed) voice's extracted features. These significant differences were common for specific voice features including AE, RMS, MFCCs (all 40 coefficients), ZCR and SB. Four or less extracted voice features consistently show significant results for most of the actors.

Figure 4.2 Total number of features showing a significant difference for each actor

4.1.5 Feature Selection

After extracting voice features for all 24 actors, t -tests assuming unequal variance were conducted. The results indicate that different features showed significant differences for each actor's vocal expression. Figure 4.3 shows that MFCCs (all 40 coefficients), RMS and AE were the most significant features to detect relevant differences. Furthermore, these three features were also significant for Actor 1, Actor 15, Actor 16, Actor 17 and Actor 18, who showed the most significant difference between their normal calm (confident) and normal fearful (stressed) vocal expressions. Therefore, MFCCs (all 40 coefficients) RMS and AE were selected to train and test various algorithms.

Figure 4.3 MFCCs (overall for all 40 coefficients), RMS, AE respectively showed 23%, 18%, 16% significant difference between their normal calm (confident) and normal fearful (stressed) vocal expressions



4.1.6 Training and Testing of Model Development

Various ML algorithms are available for classifying emotions and affective states from voice. Chapter 2, in the section titled ‘Models for Detecting Affective States’, highlights several algorithms for analysing voice and audio for emotion detection, including GMM, HMM, KNN, SVM, Decision Trees, Random Forest, ANN, DNN and CNN. For the conceptualisation and preliminary work with RAVDESS, the focus was placed on Random Forest classification algorithms, K-nearest neighbor (KNN) algorithm and Support Vector Machines (SVM). KNN was included due to its straightforward implementation and effectiveness with small datasets (Rong et al., 2009). Random Forests, based on decision trees, were selected for their robustness and ability to handle datasets with higher dimensionality (Breiman, 2001). Additionally, SVMs were chosen for their widespread use and demonstrated effectiveness in stress detection models, as highlighted in the literature review. In contrast,

algorithms like GMM and HMM are typically more suited for temporal sequence modelling and may not provide the same level of classification accuracy for static voice features in our application. Meanwhile, ANN, DNN and CNN require extensive computational resources and larger datasets, which may not be feasible within the scope of initial preliminary work. Therefore, the selection was guided by the balance between effectiveness, computational efficiency, and the specific requirements of the application in detecting stress from voice during public speaking (Kufel et al., 2023). The included three algorithms (Random Forest, KNN and SVM) were implemented using the scikit-learn library. The Random Forest algorithm was configured with 10 estimators ($n_estimators = 10$) and a fixed random seed ($random_state = 30$), while all other parameters were kept at their default values for other models. The KNN model used the default setting of 5 neighbors ($n_neighbors = 5$) with Euclidean distance (Minkowski distance with $p = 2$) and uniform weighting. The SVM classifier was implemented using the default radial basis function (RBF) kernel, with regularization parameter $C = 1.0$ and $gamma = 'scale'$.

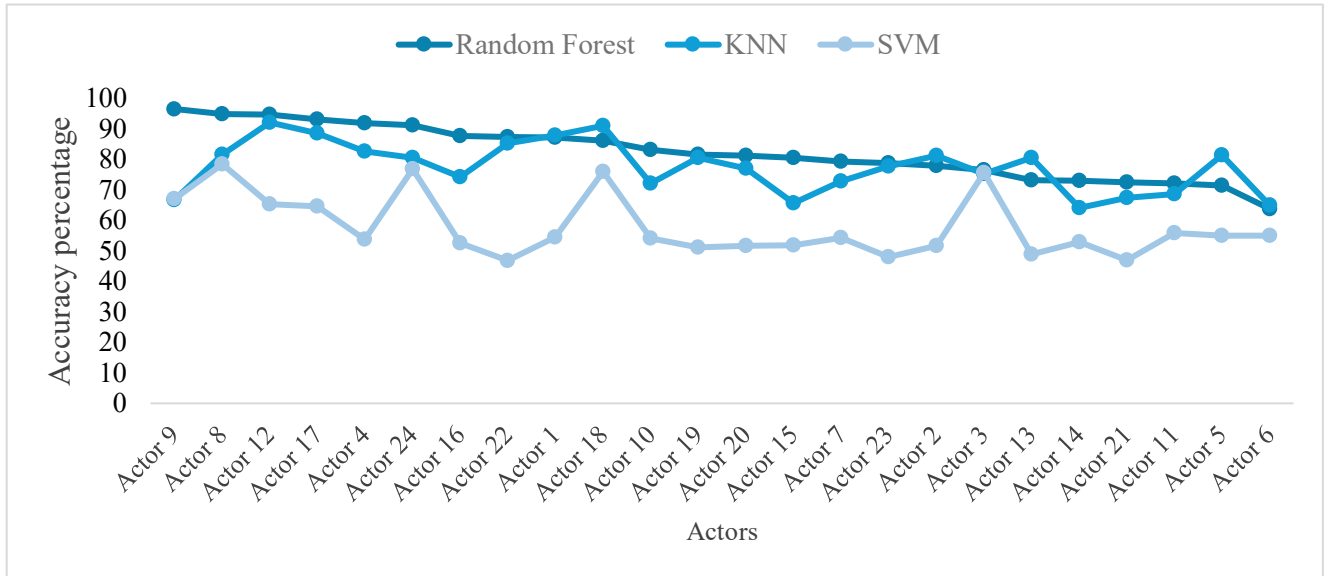
We allocated 70 per cent of the data (16 participants) for training and 30 per cent (8 participants) for testing, using each actor's voice data sample for both confident and stressed voice samples across the selected algorithms.

4.1.7 Final Results from the RAVDESS Database Model Development

The results shown in Figure 4.4 demonstrated that the Random Forest algorithm, with an average accuracy of 82 per cent, outperformed both the KNN algorithm, which had an average accuracy of 77.52 per cent, and the SVM, which had an average accuracy of 57.98 per cent. Specifically, Random Forest showed a 5 per cent increase in accuracy compared with KNN and a 24 per cent increase compared with SVM in detecting stress. These results were achieved using features: RMS energy, AE, and MFCCs (all 40 coefficients). The superior performance of Random Forest can be attributed to its ability to handle complex interactions

between features and its robustness against overfitting, making it a more reliable choice for stress detection in voice analysis.

Figure 4.4 Comparison of Random Forest, KNN and SVM for detection of stress levels for each of the actor.



The successful preliminary work with the RAVDESS database for detecting stress in speakers' voices provided important insights for developing an affective state detection model. However, as context plays a critical role in designing and developing ML algorithms for affective state detection, the decision was made to create a virtual reality public speaking (VRPS) database, as described in the following sections of this chapter. This new database includes stress-related voice samples collected within a specific public speaking context, intended for the development of SDMs.

4.2 Development of Stress Detection Model

4.2.1 Virtual Reality Public Speaking (VRPS) Database

As discussed in Chapters 2 and 3, there is a notable lack of stress voice databases collected in real-time within the context of public speaking, therefore limiting research on

stress voice analysis in the field of affective computing and the integration of such technologies in VR to improve speaking skills. The objective was to address this gap by developing a stress voice database using the voice samples collected from our first experiment (E1).

To develop the ML model for detection of real-time stress during public speaking, the voice samples collected from E1 served as training and testing dataset. The initial dataset contained 120 voice recordings collected from 60 participants. However, only 110 voice recordings were included in the final dataset, excluding samples from participants P1, P7, P20, P45, and P48 due to either no speaking at all or a lack of speech recording from one of the conditions.

In creating the final dataset for feature extraction and model development, 55 voice samples from a NP condition, and the remaining 55 from a NN condition were identified. Each voice sample involved the speakers giving a speech for 2 minutes and 30 seconds on the topics of ‘why an employer should hire you’ and ‘online vs offline courses’. The speakers had complete freedom regarding their speech content and delivery.

Voice recordings were categorised into stressed and nonstressed voice samples and further confirmed using Polar HR readings and MANOVA analysis results of speakers in the NN and NP conditions. While it would have been possible to categorise recordings based solely on the assigned conditions, the inclusion of HR readings and statistical analysis provided an additional confirmation criterion. This approach allowed us to correlate physiological stress with the psychological stress experienced by the speakers, hence providing a more robust understanding of the impact on voice. The additional confirmation was important because much of the existing research relies on voice recordings performed by professional actors, which are imitations of stress rather than genuine experiences. Uniquely, the recordings in this study reflect the authentic experiences of individuals in a public speaking context, which is

critical for validating subjective experiences and understanding the nuanced effects of stress in realistic scenarios.

Polar HR monitor readings tracked changes in HR every second throughout the experiment. These readings helped us pinpoint specific moments when stress levels peaked for the speaker, such as when the speech topic was given, during speech preparation, at the beginning of the speech, and after the speech concluded. The Polar HR readings indicated that HR readings were consistently higher and reached their peaks when the speaker gave their speech in both conditions in comparison to their baseline HR readings and when the person completed their speeches. Additionally, the overall MANOVA analysis on reactivity scores of HR and BP (calculated from average readings during pretest, during, and posttest phases of the NN and NP conditions) confirmed these findings. Results on physiological data (HR, SBP, DBP) corresponding to NN and NP conditions and stress arousal check lists (SACL) further confirmed that the participants in the NN condition were more stressed in comparison to giving speech in the NP condition (detailed discussion can be found in Chapter 3 [Main Results: Hypothesis Testing](#)).

Given the established link between increase in HR and BP readings and elevated perceived psychological stress levels of the individual during a stressful task (Hjortskov et al., 2004; Kim et al., 2018) or event such as of public speaking (Bassett et al., 1987; Feldman et al., 2004), in the current research, it was predicted that there would be a significant difference between the voice samples in negative neutral (NN) and NP condition.

Following the analysis of the NN and NP conditions, the voice samples were categorised into stressed and nonstressed groups resulting in the creation of a stressed voice sample database. This categorisation was necessary to ensure that the dataset accurately represented the varying stress levels experienced by participants. The development of the VRPS database, therefore, included voice samples from both stressed and nonstressed

participants, enabling the extraction and comparison of specific voice characteristics across these conditions. The foundational categorisation was critical for the subsequent development of a stress detection model, which will be explored in the next phases of this research. The steps taken to create the VRPS database were instrumental in producing a dataset tailored specifically for the context of public speaking. In the second section of this chapter, the VRPS database was utilised to extract voice features indicative of stress and to develop a model capable of detecting stress in real-time public speaking scenarios using VR.

4.2.2 Feature Extraction Using VRPS Database

LibRosa Python library was used to extract all the voice features from the voice recordings in the VRPS database. Normalisation was not applied to the features extracted from these recordings. Although the process of normalisation is usually done to standardise data and improve model robustness, its application still depends on the nature of the data and the specific goals of the analysis. In our case, first, global normalization was unnecessary because the conditions (equipment, recording environment, supportive & non-supportive VR scenario) for voice recording in the VRPS database were consistent across all participants (see Chapter 3), which minimized variability due to external factors. This ensured that no additional variability needed to be normalized across sessions. Second, due to the within-subject experimental design, normalization could have obliterated emotional variation between each participant's calm and stressed voice samples in VRPS database. Specifically, normalisation may have flattened or distorted features like breathing patterns, vocal tension and absolute differences in intensity and energy, which are crucial for capturing emotional changes and are often reflected in acoustic features such as Root Mean Square (RMS), Amplitude Envelope (AE) and low-order MFCCs. Third, inappropriate normalization can reduce the discriminative power of emotion-related acoustic features and negatively affect classification performance (Busso et al., 2013). Since the goal was to explore and identify stress-indicative voice features, preserving

the raw voice files was essential to achieving the study's objective to carry out the necessary investigation of within-subject emotional differences. For these reasons, normalization was intentionally not carried out to retain the emotion-relevant acoustic information critical to both analysis and model development.

We began by extracting time domain features such as AE, ZCR and RMS. We further extracted frequency related voice features such as spectral centroid (SC), spectral bandwidth (SB), spectral roll off (SR) and spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs). All features were extracted using a frame size of 24,576 samples and a hop length of 8,192 samples. Each of these features helps to build a unique profile of the sound under analysis.

Following extraction of features, we conducted a cosine distance score calculation test (specifically for AE) and *t*-tests assuming unequal variance for two samples obtained for all other extracted voice features. We formed the following hypothesis and rejection criteria.

H₃: There is significant difference between each respective participant's voice in NP condition (nonstressed) and NN condition (stressed) for the extracted feature.

Rejection Criteria: Reject H₃ if cosine similarity score was above 0.90 (for AE) and $p\text{-value} < 0.05$ (for remaining voice features).

4.2.3 Results from the Feature Extraction

Table 4.1 shows the statistical differences obtained on various extracted voice features. It includes cosine similarity metrics (for the AE) and *t*-tests (for ZCR, RMS, SC, SR, SB, and MFCCs) where AE = Amplitude Envelope, RMS = Root-Mean Square, ZCR = Zero-Crossing Rate, SC = Spectral Centroid, SB = Spectral bandwidth, SR = Spectral rolloff and MFCCs = Mel-Frequency Cepstral Coefficients in both NP and NN conditions).

Specifically, the statistical values highlighted in orange indicate a significant difference ($p\text{-value} < 0.05$) between the respective participant's voice in NP condition (nonstressed) and

NN condition (stressed) for the extracted features. On the other hand, the statistical values in black colour indicate no significant difference ($p\text{-value} > 0.05$) between the respective participant's voice in NP condition (nonstressed) and NN condition (stressed) for the extracted features.

The amplitude envelope (AE*) for each participant's stressed and non-stressed voice samples in both NP and NN conditions were extracted. While a visual inspection of the AE plots for each participant speaking in their respective condition of NN and NP did not show any major differences, therefore an additional measure of a cosine similarity score was calculated to quantify the similarity between the two vector AEs on voice samples of participants speaking in NP and NN conditions. Cosine similarity is one of the commonly used metrics that is used to measure how similar (directional alignment) two data objects are irrespective of their size or magnitude.

$$\text{Similarity (x.y)} = \text{Cos}(\theta) = \text{x.y} / ||\text{x}|| * ||\text{y}||$$

The cosine similarity was used as an exploratory tool to assess whether the vocal energy patterns (as captured by AE) differed between the stressed (NN) and non-stressed (NP) voice samples of different affective states. It provided an exploratory view of how amplitude patterns aligned or diverged between the two emotional conditions, with lower similarity potentially reflecting more dynamic or irregular energy shifts in the stressed condition. As the voice recordings in VRPS were matched in duration across NN and NP conditions and were collected under identical recording conditions, this comparison provided insight into within-subject signal-level variation. The results (Table 4.1, AE* column 1) shows the cosine similarity scores for each of the participants' AE comparisons based on the voice samples obtained in NP and NN conditions. For all the 55 participants, the cosine similarity score was closer to 1 and thus indicated a high similarity between the amplitude envelope in both (NN and NP) for respective participants' voice samples, suggesting that the temporal energy pattern of speech remained

largely consistent for each individual, despite the differing emotional contexts. This analysis was not used for model training but served to support qualitative observations of signal-level emotional expression.

The t -tests were conducted on each participant's ZCR, RMS, SC, SR, SB and MFCCs obtained from the voice samples in NN and NP conditions. The results for $n = 55$ participants indicate (Table 4.1) that 12 participants (14, 19, 45, 44, 17, 38, 46, 41, 32, 13, 57 and 58) showed significant difference in six voice features. Twenty-six participants showed significant difference in five voice features. Nine participants showed significant difference in four voice features. The remaining eight participants showed significant difference in three or one voice features.

Table 4.1 *t-Test Analysis Results on Extracted Voice Features*

	AE*	RMS**	ZCR**	SR**	SC**	SB**	MFCCs**
P3	0.9874	<.05	<.05	<.05	<.05	<.05	<.05
P4	0.9871	<.05	<.05	<.05	<.05	<.05	<.05
P8	0.9883	<.05	<.05	<.05	<.05	<.05	<.05
P13	0.9805	<.05	<.05	<.05	<.05	<.05	<.05
P14	0.9844	<.05	<.05	<.05	<.05	<.05	<.05
P17	0.9799	<.05	<.05	<.05	<.05	<.05	<.05
P19	0.9791	<.05	<.05	<.05	<.05	<.05	<.05
P32	0.9713	<.05	<.05	<.05	<.05	<.05	<.05
P38	0.9820	<.05	<.05	<.05	<.05	<.05	<.05
P41	0.9818	<.05	<.05	<.05	<.05	<.05	<.05
P44	0.9715	<.05	<.05	<.05	<.05	<.05	<.05
P45	0.9707	<.05	<.05	<.05	<.05	<.05	<.05
P46	0.9758	<.05	<.05	<.05	<.05	<.05	<.05
P47	0.9765	<.05	<.05	<.05	<.05	<.05	<.05
P53	0.9805	<.05	<.05	<.05	<.05	<.05	<.05
P57	0.9778	<.05	<.05	<.05	<.05	<.05	<.05
P58	0.9778	<.05	<.05	<.05	<.05	<.05	<.05
P2	0.9873	<.05	<.05	<.05	<.05	<.05	.122
P6	0.9880	<.05	<.05	<.05	0	<.05	<.05
P9	0.9815	<.05	<.05	<.05	<.05	<.05	.77
P11	0.9754	<.05	<.05	<.05	.11	<.05	<.05
P12	0.9775	<.05	<.05	<.05	<.05	0	<.05
P21	0.9862	<.05	<.05	<.05	.62	<.05	<.05

Chapter 4: Phase II, Stress Detection Model Development

P22	0.9825	<.05	<.05	.40	<.05	<.05	<.05
P23	0.9838	<.05	<.05	<.05	<.05	<.05	.29
P24	0.9831	<.05	<.05	<.05	<.05	<.05	.25
P25	0.9835	<.05	<.05	<.05	.42	<.05	<.05
P26	0.9770	0	<.05	<.05	<.05	<.05	<.05
P27	0.9834	<.05	<.05	<.05	.17	<.05	<.05
P28	0.9819	<.05	<.05	<.05	<.05	.005	<.05
P29	0.9821	<.05	.06	<.05	<.05	<.05	<.05
P31	0.9901	<.05	<.05	<.05	.78	<.05	<.05
P34	0.9757	<.05	<.05	<.05	.18	<.05	<.05
P36	0.9781	<.05	<.05	<.05	<.05	<.05	<.98
P39	0.9710	<.05	<.05	.84	<.05	<.05	<.05
P43	0.9818	<.05	0.188	<.05	<.05	<.05	<.05
P49	0.9832	<.05	<.05	<.05	<.05	<.05	.39
P50	0.9772	<.05	.014	<.05	<.05	<.05	<.05
P51	0.9785	<.05	<.05	<.05	<.05	<.05	.25
P54	0.9819	0	<.05	<.05	<.05	<.05	<.05
P16	0.9636	<.05	0.27	<.05	<.05	<.05	.86
P30	0.9855	.42	<.05	<.05	.78	<.05	<.05
P33	0.9843	<.05	<.05	.11	.25	<.05	<.05
P37	0.9875	.24	.62	<.05	<.05	<.05	<.05
P40	0.9835	<.05	<.05	<.05	<.05	.16	.80
P42	0.9809	0	<.05	<.05	<.05	0	<.05
P52	0.9755	.05	<.05	<.05	<.05	<.05	.85
P56	0.9819	0	<.05	.91	<.05	<.05	<.05

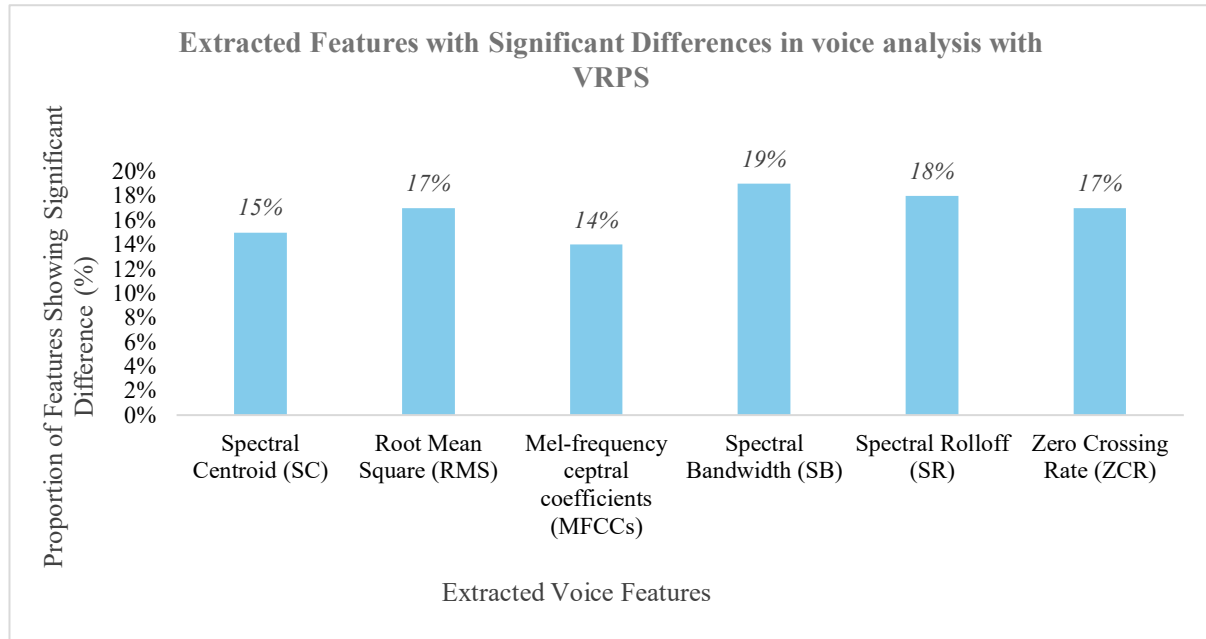
P60	0.9753	<.05	.005	<.05	<.05	<.05	.12
P10	0.9809	<.05	0.182	<.05	.17	<.05	.35
P15	0.9844	<.05	0.67	<.05	.70	<.05	.72
P59	0.9734	.10	<.05	.11	<.05	<.05	.71
P5	0.9891	0	0	0	0	<.05	1
P35	0.9789	<.05	.26	.92	.96	.73	1

Note. The table presents the p-values for various voice features across different participants (P3 to P60) when comparing confident (nonstressed) and stressed voice samples. Significant differences ($p < .05$) are highlighted in orange for the features: AE (Amplitude Envelope), RMS (Root-Mean Square), ZCR (Zero-Crossing Rate), SR (Spectral Rolloff), SC (Spectral Centroid), SB (Spectral Bandwidth), and MFCCs (Mel-Frequency Cepstral Coefficients). Nonsignificant p-values ($p > .05$) indicate no significant difference between the voice samples for the respective features. The values that are less than .05 indicate statistically significant differences between the confident and stressed voice samples, whereas values greater than .05 indicate nonsignificant differences.

* Cosine Similarity Test Results, **: t -test Results, Orange highlights: significant difference, Black highlights: non significance difference

4.2.4 Feature Selection

Results indicate that among all the extracted features, SB, SR, RMS, ZCR, SC and MFCCs showed significant difference in participants voice in NP (nonstressed) and NP (stress) conditions (refer to Figure 4.5 below for significant percentages of features).

Figure 4.5 *Significance Percentage of each Voice Feature*

As discussed in Section 4.2.4 on Feature Selection, voice features such as MFCCs (23 per cent), RMS (18 per cent), AE (16 per cent), ZCR (12 per cent), SR (12 per cent), and SC (8 per cent) exhibited significant differences between the actors' calm (confident) and fearful (stressed) voice samples when using the RAVDESS database. Although the order of significance percentages for these features differed from previous findings, it's important to note that all features, except AE, consistently showed significant differences across both databases. This suggests that, despite some variation in the significance levels, the majority of voice features reliably indicate stress in different contexts.

The databases (RAVDESS database and VRPS database) and their respective emotional content of the voice recordings is unique, and there could still be inferences and insights drawn upon. Hence, the process of finding out significant voice features also included comparing and analysing significant features of voice obtained with the RAVDESS database and significant voice features obtained using the VRPS database from experiment 1 (E1).

Firstly, there is a noticeable contrast in the significance values of the voice features, which may be attributed to the differences in the emotional content of the voice recordings in the two databases. The RAVDESS database features recordings of actors portraying fear (assumed to represent stressed samples) and calmness (assumed to represent confident samples), while the VRPS database includes recordings of university students in stressed (NP condition) and nonstressed (NN condition) states. Despite these differences, the common thread is the presence of various emotional states and the respective different characters of the voice recordings of both databases. Secondly, the seemingly incongruent results can be explained by recognising the disparity between the speakers in each database. The RAVDESS database engaged professional actors acting out certain suggested emotions, whereas the VRPS database consist of audio recordings from active university students who are relatively inexperienced and speaking in different conditions. The vocal qualities of the speakers in these two databases are inherently different, with students likely having less control over their vocal expressions compared with the trained actors. Lastly, the settings in which the voice samples are recorded needs to be noted as well. The RAVDESS voice files were captured in professional settings, while the VRPS voice files were collected in an experimental setting using a virtual reality set-up that simulated a real-life auditorium environment

Therefore, due to different emotional states of the speakers in both databases (professional actors as speakers vs university students as speakers), and the speeches given in different conditions and settings (real-life acting vs lab-based VR experimental scenarios), and unique selected speakers of different background (proficient vs inexperienced speakers), the voice features extracted were expected to show some variable results.

In Chapter 2, Table 2.1, we highlighted that temporal feature (such as RMS, ZCR, AE) denote intensity of signal, therefore signifying energy of the signal with respect to time. The literature regarding spectral features suggested that pitch, fundamental frequency (F_0),

formants and coefficients such as MFCCs, etcetera, are classified to be frequency-based features. Specifically, formants provide the location of where frequency-based resonance occurs, or concentration of energy happens in any signal. These formants provide the information about amplification or attenuation of any power spectrum. The formants are an extremely important voice feature that effects the voice quality. Lastly prosodic features (intensity, pitch, duration, shimmer, jitter, etc.) are a combination of both temporal and spectral features, as also discussed in Chapter 2.

Synthesising all the previously provided information and examining our specific case in detail, we can attribute the findings to the complex and multifaceted nature of the voice samples produced by different participants in the VRPS database under two varying VR scenarios: NP (nonstressed) and NN (stressed) conditions. Although these scenarios showed significant differences in physiological parameters and stress arousal checklist scales, they did not result in substantial differences in the amplification or attenuation of the participants' voice samples. This is evident in the significance values of MFCCs features obtained with VRPS compared with those from the RAVDESS database. This phenomenon could be also ascribed to an intuitive understanding of the fact that resonances in the formants could have been dispersed out in the participant's voice due to high arousal (Echternach et al., 2016) together with high levels of stress in the voice recordings of NN condition (stressed) in comparison to NP condition (nonstressed), instead of the frequencies resonating and coming together to form high significant differences in MFCCs.

Therefore, we still selected MFCCs for development of our SDM. In our preliminary analysis with RAVDESS (Section 4.1.5) (Arushi et al., 2021) and current work done with VRPS database voice samples (Section 4.2.2), the MFCCs were the only consistently significant voice features for detecting any emotional changes in the voice pattern, tone and rhythm of any speaking voice and the affective state of the speaker, despite the MFCCs being less consistently

significant in comparison to other voice features with the VRPS database of E1. Recent studies (Dolka et al., 2021; Jain et al., 2020; Kandali et al., 2008; Mishra et al., 2023; Nalini et al., 2016) aimed at detecting general emotional states of speakers have reported MFCCs to be one of the most reliable and consistent sets of voice features for detecting any deviation in normal voice such as stress (Hilmy et al., 2021; Li et al.; Mohapatra et al.; Nwe et al., 2003b; Tomba et al., 2018) in comparison to other conventional voice features.

MFCCs are a specific set of voice features that are computed using a Mel filter bank and provide information about the power spectral envelope of any voice signal. These features are specifically designed to mimic the human perception of speech due to use of logarithm spectra and are more closely related to sound quality (or timbre) of the produced vocals (Heidemann, 2016; Woollen, 1995). MFCCs comprise 39 sub bands and could provide information about the short-term power spectrum of any voice sample, removing the micro level structure of the voice and typically focusing on part of the signal that contains most information. MFCCs, being a vector feature with 39 sub features, offer an efficient and straightforward procedure to compactly represent the spectrum and, therefore, develop a stress detection algorithm. Conversely, while other temporal and prosodic features could provide information pitch, duration and pace, they need to be combined to perform well with any ML algorithms such as KNN, SVM and GMMs (Besbes et al., 2019; Lech et al., 2014).

Despite little information about MFCCs' different sub bands and specifically their utility and their usability to detect stress, MFCCs are reported to show the reduced airflow articulation and breathy or harsh voice quality in clinical studies (Godino-Llorente et al., 2006). Therefore, we also conducted a *t*-test with different MFCCs sub bands and results are summarised in Table 4.2. The result from the table indicates specifically the different 39 features of MFCCs which were extracted for all the voice samples and checked for significant difference between the voice samples obtained from NP and NN condition. For the training and

testing and real-time implementation of the stress detection model, we selected 32 MFCCs excluding MFCCs from 1-8, keeping in mind the fact that it is difficult to find MFCCs spectra which is unbiased and congruent with representation, therefore including most of the MFCCs in our model development (Tom Bäckström, 2022).

4.2.5 MFCCs Extraction Procedure

We used LibRosa, a Python language library (McFee et al., 2015) dedicated to extracting various voice features. We extracted the MFCCs as well as other voice features for all the 55 voice samples described in [Section 4.1.5](#) using the same library, except for AE. The code for all the extracted voice features is shown in Appendix F of this chapter. Along with LibRosa, we used NumPy, SciPy, SciKit, Pandas and other Python libraries required for other purposes such as for plotting graphs.

Table 4.2 MFCC features bands Significance Value

MFCC																																									
band	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
P27	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	
P16	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	
P21	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	1	1	
P24	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	
P30	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P57	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	
P11	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0
P25	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	
P45	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	
P47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	1	1	
P51	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
P3	1	1	0	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
P13	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	

Chapter 4: Phase II, Stress Detection Model Development

P6	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	
P9	1	0	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	
P10	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	
P14	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1	
P17	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	0	1	1	
P35	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	0	1	1
P2	1	1	0	1	1	0	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	
P8	0	1	1	1	1	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	0	
P22	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	0	0	
P29	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	
P31	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	1	1	0	1	1	1	1	1	1	
P43	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	1	0
P44	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
P46	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
P50	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	1	0		
P52	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	0	
P53	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	

Chapter 4: Phase II, Stress Detection Model Development

P60	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0	1	1	
P12	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1	0	0
P33	0	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	
P36	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	1	
P37	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	
P56	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	0	1	1	
P19	0	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1	1	
P40	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	
P54	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	
P58	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	1	1	1	
P4	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	0	1	1	1	0	1	0	1	0	1	0	1	
P32	0	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	0	1	1	1	1	0	1	1	
P49	1	0	0	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1	
P23	1	1	0	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	0	0	1	
P38	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	1	0	0	1	
P39	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	0	1	1	0	0	0	
P59	1	1	0	1	1	1	1	1	1	0	1	1	1	0	1	0	0	1	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	1	1	0	0	

[illegible]

MFCC 1-12: Capture the broad spectral envelope and are the most significant features (represents the general shape of the spectrum)

MFCC 26-39: Denotes delta (difference) and delta-delta (acceleration) coefficients, (indicates the dynamic changes in the spectrum over time).

(Rabiner et al., 1993)

0 indicates nonsignificant difference in the extracted MFCCs band for the voice recordings of NN and NP condition

4.2.6 Development of Stress Detection Model (SDM) for VR Integration

Theoretical Framework

We developed a feedforward neural network (FNN) [also commonly known as a multilayer perceptron (MLP) based stress detection model (SDM)]. The FNN can play a crucial role in stress identification from the identified voice feature MFCCs due to its ability to learn complex patterns and analyse the relationships within the different bands of MFCCs.

Generally, a FNN constitutes a fundamental network architecture with three or more different layers. The input layers take in the data and the hidden layers are responsible for undertaking computational processes on the data received while the output layer is responsible for synthesising the hidden layer's computation outcomes and generating the network's resulting output. When the input received by a node from the previous layer is above the threshold value, the node gets activated and the data or information is then sent to the next layer for further processing. In our case, the input layer can take in the extracted MFCC voice features and analyse the dynamic nature of stress. Hidden layers can hierarchically represent the input data and automatically extract the relevant stress information, and an output layer can generate the network's results accordingly. All the nodes across the layers are interconnected and consist of a weight and threshold value.

FNNs include two main phases: namely a feedforward phase and a backpropagation phase. The feedforward phase involves feeding of features data (in our case, extracted MFCC feature data) in the input layer. As the propagation of MFCCs data continues to the next hidden layer, a weighted sum is calculated and passed through an activation function. This process is continued until a prediction is made and the results are transferred to the output layer. FNNs can be trained therefore using our labelled dataset (stressed and nonstressed corresponding to the MFCCs) obtained from VRPS database voice samples obtained from E1. The second phase of backpropagation works based on a gradient descent optimisation algorithm and it involves

calculating errors and adjusting the weights to minimise the error for the prediction to be made (in our case, stressed or nonstressed MFCCs).

FNNs also involve different activation functions that play a crucial role. These activation functions are responsible for introducing the nonlinear properties in the network, allowing models to learn complex patterns to form the decisions. Common activation functions include sigmoid and ReLU (Rectified Linear Unit). In our case we are using a sigmoid function as an activation function (details on next page). FNNs do not involve circulation or looping of information within the network infrastructure in comparison to a Recurrent Neural Network (RNN). FNNs are effective in capturing contextual dependencies within a voice audio sample and will be effective in considering how stress varies based on context of spoken word with a certain tone. A trained FNN can help process the speech voice samples in real time making it suitable for us to integrate it with the developed virtual reality-based application for improving public speaking skills. FNNs allow for implementation of models in various environments, for example, in our case of continuous stress detection from voice in our VR application. FNNs are also considered to be adaptive to learning different speakers and diverse stress patterns, making it versatile to deploy in real-life scenarios across a variable range of individuals.

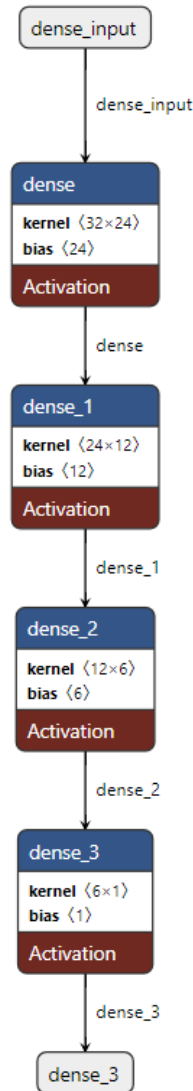
Lastly, FNNs can be integrated with different modalities and can improve the accuracy over time with fine-tuning, making efficient tools for understanding and identification of stress from human voice communication.

FNNs are the most simplistic form of neural networks, although they are flexible regarding the specific number of hidden layers and nodes, which ultimately are responsible for the performance of the network model. Overfitting is one of the possible issues that must be addressed. However, FNN provides a straightforward approach to developing and making predictions for various types of applications. For the reasons mentioned above, we chose FNN for developing our Stress Detection Model.

4.2.6.1 Stress Detection Model: Architecture, weights, activation functions, compilations

Below in Figure 4.6 is the visualisation of the developed SDM's architecture using Netron (<https://github.com/lutzroeder/netron>), a Python programming-based library, which is freely available under MIT licence.

Figure 4.6 Stress Detection Model visualisation with all layers, weights and activations



The SDM model (code is shown in Appendix G) was created using Keras. Keras (<https://github.com/keras-team/keras>) is a powerful, open source, and easy to use ML library that provides a Python-based application programming interface (API) for building and evaluating artificial DNN models, allowing for efficient experimental and implementation use

cases. It is licensed by Apache 2.0, and it is widely used by CERN, NASA, NIH and other scientific organisations all over the world.

The SDM was created using sequential API in Keras, and it consists of one input layer, two hidden layers and a final output layer. The first layer (dense) is a dense layer with 24 nodes or neurons, input dimension of MFCCs = 32. All the input units are assigned a separate weight to each output unit. For n inputs and m outputs, the number of weights is $n*m$. Additionally, you have a bias for each output node, so you are at $(n+1)*m$ parameters. The formula to determine weight in a fully connected layer:

$$\text{Weights in a fully connected layer} = (\text{input} + 1) * \text{Number of units} \quad (1)$$

Therefore, the layer 1 has 792 weights $[(32+1) * 24]$ and has 24 biases for each input. The second layer (dense_1) is first hidden layer, has 300 weights $[(24+1) * 12]$, and 12 bias terms. The third layer (dense_2) is the second hidden layer in the network. It consists of 78 weights $[(12+1)*6]$ and 6 bias. All these first three (dense, dense_1, dense_2) layers are activated with ‘ReLU’ activation function. A ReLU activation function stands for Rectified Linear Unit, and mathematically it is defined as:

$$f(x) = \max(0, x) \quad (2)$$

ReLU activation is helpful for the SDM as it allows nonlinearity in the model, leading the model to learn from data, which is not linearly separable. It also makes the computation processes efficient due to thresholding operations, and lastly ReLU helps in mitigating the vanishing of gradient problem, that occurs in deep networks when gradient becomes very small and makes the update of the weights difficult. Despite issues such as dead neurons and unbounded activations being common with the use of ReLU, activating layers with ReLU is widely considered and still seems to be the best choice to include in input and hidden layers.

The last layer (dense_3) is the output layer in the SDM and it consists of seven weights [(6+1) *1] and 1 bias. A sigmoid function is used in the final output layer of the SDM for a binary classification. A sigmoid function outputs the value between 0 or 1, and it could be interpreted as an input belonging to a stressed or nonstressed class according to the value that comes through as a final prediction.

The SDM is then finally compiled using a binary cross entropy loss function, the Adam optimiser and accuracy as a metric to monitor during training. The binary cross entropy is a logistic loss function and is used in binary classification of any problem. A loss function measures the difference between the predicted binary outcomes, and actual binary outcomes. Therefore, it is used to measure performance of a classification model with sigmoid prediction. In the binary classification, the task is to classify two possible classes (labelled as 0 and 1), particularly suitable for our case with stressed MFCCs labelled as 1 and nonstressed MFCCs voice samples labelled as 0. A binary cross entropy and sigmoid activation is a potent combination for handling a binary classification problem by providing a probabilistic output, effective learning gradient to the model, and penalising the wrong but confident predictions. The equation 3, denotes the loss function, with y indicating the actual value, and $y^{\wedge}$ indicating a predicted probability that the input belongs to 1 or 0.

$$L(y,y^{\wedge})=-(y\log(y^{\wedge})+(1-y)\log(1-y^{\wedge})) \quad (3)$$

The optimised function used in training of model is Adam (Adaptive Moment Estimation). Adam moves and maintains the average of both gradient and squared gradient descent and adapts the learning rate for each parameter for model training.

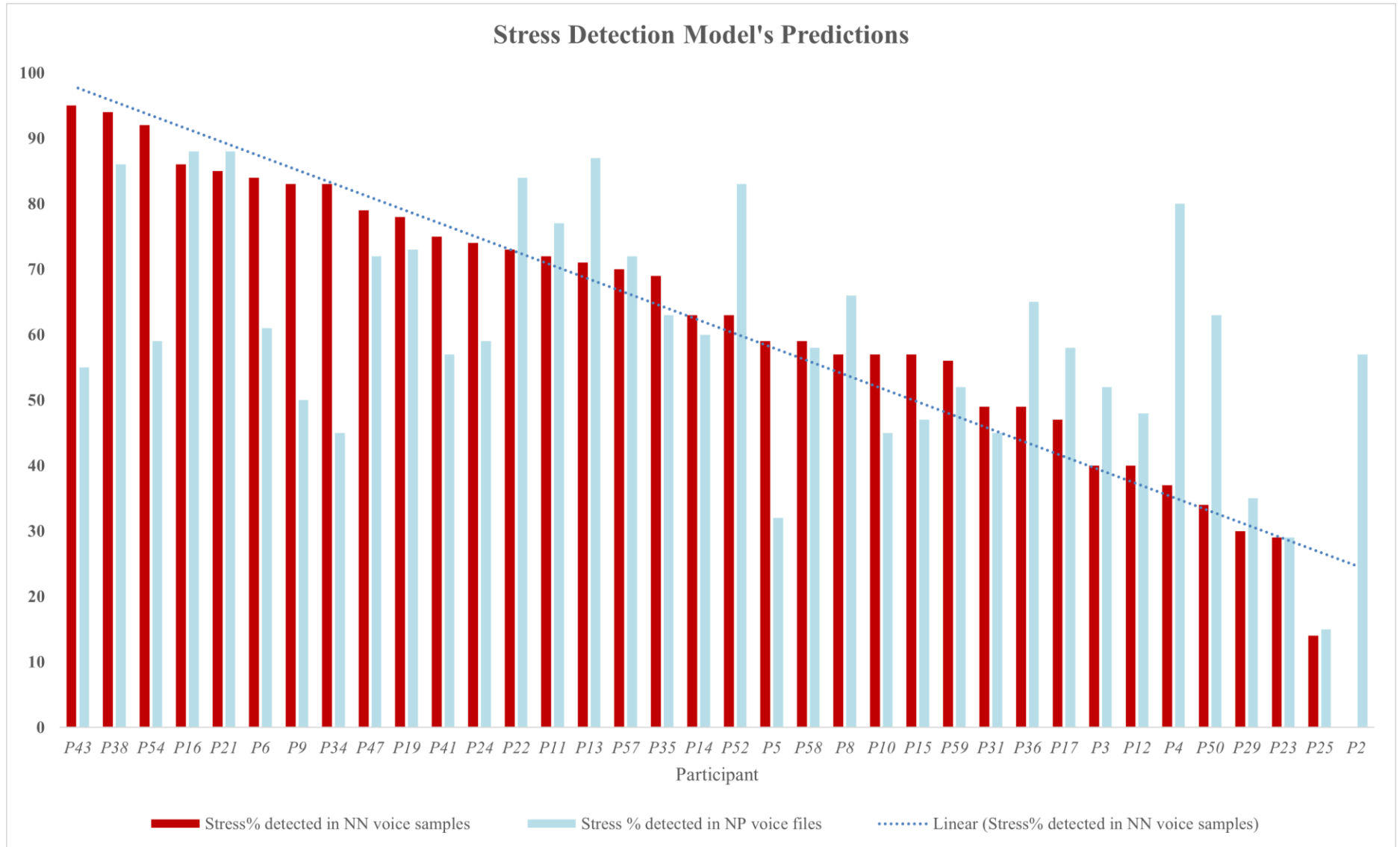
4.2.6.2 Training and testing of the Stress Detection Model

The model was trained on both the stressed voice samples (from NN condition) and nonstressed voice samples (from NP condition). The training data included a total of 36 voice sample's extracted MFCC features (18 per condition). The model was trained on 50 per cent (36 files – 18 NN and 18 NP MFCCs) of data and was tested on 50 per cent. Initially, different trials were carried out by manipulating varying sampling rates [time window and hop length] for extraction of MFCCs (essentially means the no. of samples recorded in one minute) for all the identified voice samples.

Considering the sample size, we set the frame size for MFCCs feature extraction to be 24,576 samples (which corresponds to a 3-second window) and a hop length of 8,192 (equal to 1s), allowing the ML model to analyse the last 3 seconds of data in a continuous and smooth progressive manner. The sample size, window length and hop length allows each analysis window to smoothly progress through the audio file, capturing overlapping segments to provide a more comprehensive understanding of the audio features, particularly those related to stress detection (Giannakopoulos et al., 2014; Sahidullah et al., 2012). There is no strict rule to split the data. However, when we have a small dataset, it is judicious to allocate the larger portion for testing (Joseph, 2022). The model training process involved 100 epochs with each epoch divided into a batch of 32 samples. The development of the ML model for stress detection involved different FNN configurations, in the end resulting in the described SDM as visualised in Figure 4.6 and making it suitable to test and integrate in the VR application.

The SDM was then tested on the untrained voice files which is previously unseen voice data from the VRPS dataset, to assess its ability to detect stress levels in real-time with a completely new set of voice data. The results of this testing are summarised in Figure 4.7.

Figure 4.7 Results of stress detection model classification



The blue-coloured bars in Figure 4.7 indicate stress detection percentage in the NP (nonstressed) voice sample files for respective participant, while the red-coloured bars indicate the stress detection percentage of the NN (stressed) of the respective participant. Overall, Figure 4.7 shows that the stress in the NN conditions was detected to be higher in comparison to NP conditions on 18 testing samples [P5, P6, P9, P14, P15, P19, P23, P24, P31, P34, P35, P38, P43, P47, P54, P58, P59] of the voice samples. The SDM was able to achieve an accuracy of 69.44 per cent in identifying stress in the testing dataset.

The blue line in Figure 4.7 represents stress detected in NN voice samples, shows that the model is generally consistent in identifying stress in the NN condition and was able to correctly classify stress 61.90 per cent of the time in NN conditions, which is promising for its application in real-time scenarios. The consistency of the blue line, which remains above 50 per cent for most participants, indicates that the model is effectively capturing significant stress-related features from the voice data. The SDM in NP (nonstressed) voice samples still exhibits greater variability. It could be inferred that the model sometimes detects higher levels of stress in NP condition samples, with P3, P8 and P25 showing higher stress in NP conditions, their vocal traits must have mimicked stress, or there may be instances of misclassification by the model.

Although this suggests the model can recognise some changes in the affective state of the voice samples from both conditions, the model could be less successful in classifying a new voice file beyond the training dataset. Due to this reason, to explore the model's behaviour on real-world speeches, we decided to diversify our test cases to include voice recordings from renowned public figures such as Barack Obama and Tulsi Gabbard, who are known for their confident public speaking. We acknowledge that the speakers' actual stress levels cannot be directly verified, these individuals' voice samples were selected due to their established reputation as effective orators under high-pressure situations. The SDM detected only 33-38

per cent stress in the sample files from both Obama and Gabbard's recordings. Although this is not definitive, the model's results align with the expectation that their voices should exhibit low stress levels and project confidence, ensuring the cross-validation. This exploratory outcome confirms the model's ability to generalise beyond the specific conditions it was trained on, reinforcing its potential utility in real-world applications.

Although recent deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers have shown strong performance in audio-based emotion recognition tasks, these models were not utilised in the current study. CNNs are typically used with 2D spectrogram representations, allowing the model to learn from spatial patterns. However, in this study, MFCCs were extracted per frame across different actors, resulting in input representations that may vary in temporal length and visual structure. This would have had made spatial learning inconsistent and potentially misleading. CNNs do not inherently model temporal dependencies, which are important in affective speech analysis. Although RNNs are more suitable for sequence-based inputs like audio, are computationally intensive and prone to issues such as vanishing gradients on small or variable-length datasets. Transformers, though more robust in capturing long-range dependencies, require substantially larger datasets to generalize effectively and are computationally expensive.

The primary goal of this study was to investigate voice features indicative of stress and develop a model using a custom-built VRPS database focused on a public speaking environment. Therefore, traditional feature engineering and machine learning methods and a Feedforward Neural Network (FNN) were chosen to enable transparent evaluation of specific voice features (e.g., RMS, AE, MFCCs). It provided interpretability, which was necessary for this research.

While we considered making further adjustments to the model, such as modifying the number of layers, adjusting weights, altering batch sizes or increasing the number of epochs, we ultimately decided to maintain the original configuration. The reasoning behind this decision was that the SDM consistently produced reliable results for confident voice samples, suggesting that the model has reached a stable and effective state for its intended application.

The prediction results of the SDM, which includes both test data from the VRPS database and additional diversified test cases, suggest that the model has strong generalisation capabilities. The results demonstrate SDM's ability to make accurate predictions across various scenarios and cases, indicating that it has successfully learned meaningful patterns from the training data. It suggests that the model could be effectively applied to public speaking scenarios.

Chapter 5: Affective Computing Integrated VR Application for Public Speaking Training via Experiment 2

5.1 Rationale and Aim of Experiment 2

In the field of affective computing, the recognition of various affective states and emotions such as happiness, sadness, anger, disgust, fear and surprise has been extensively studied using voice (Lin & Wei, 2005; Nogueiras et al., 2001; Nwe et al., 2003b). However, as discussed in Chapter 1 [Introduction](#) and Chapter 2 [Affective Computing and Virtual Reality for Public Speaking](#), detecting stress via voice analysis in a public speaking context using VR in real-time remains less explored and underresearched (Healey & Picard, 2005; Sano & Picard, 2000). Effective detection of stress in this context can significantly enhance individuals' abilities to manage stress and deliver more impactful public speeches.

Stress encompasses psychological and emotional components and, therefore, research poses significant challenges for reliable detection and identification, especially in the public speaking context. Stress is detected using either nonverbal features (Han et al., 2014; Pfister & Robinson, 2011) or physiological features (McDuff et al., 2014; Tecles et al., 2014), or self-reported questionnaires (Barreto et al., 2007; Gillespie et al., 2017) during stressful events. Voice analysis techniques for analysing stress during public speaking were limited. However, even these studies have limitations in the lack of a comprehensive database of stress-related voice samples. To address this gap, Phase I [Chapter 3](#) established the groundwork for this research by designing and developing VR environments capable of inducing stress and thereby leading to creation of a stress-related voice samples database under two different conditions.

Phase II [Chapter 4](#) utilised the collected voice samples from Experiment 1 and identified the relevant voice features indicative of stress. These features were then used to

construct a machine learning-based stress detection model (SDM), resulting in an FFNN created with the use of MFCC features.

The present Phase II, Chapter 5 describes how the developed ML SDM can work within different VR scenarios to provide real-time feedback. This phase concludes with Experiment 2 (E2), which aims to test the stress-buffering effect of real-time virtual audience behaviour on detected stress in speakers across the two different scenarios. The aim is focused on comparing supportive and competitive feedback scenarios of the public speaking within the VR application. In particular, the two VR modes are designed to provide users with relevant real-time feedback, either in the form of social support (i.e. audience to always provide positive feedback regardless of stress levels) or competitive mode (i.e. audience to provide negative feedback when the speech performance is low and positive feedback only when the speech performance is identified as good).

E2 has several potential implications. First, the findings can contribute significantly to the existing literature on the stress-reduction capabilities of VR technology when integrated with affective computing to enhance public speaking abilities. If the VR prototype, with its real-time feedback from virtual audiences, can effectively reduce stress, it could become a valuable tool for improving stress management practices and creating long-term consistent training environments to enhance public speaking skills. Additionally, VR environments that incorporate ML models to provide real-time feedback could serve as innovative tools for enhancing public speaking skills in a broader audience, regardless of their stress management challenges.

5.2 Research Questions and Hypothesis of Experiment 2

Experiment 2 (E2) is focused on the objectives as mentioned in Chapter 1, i.e., design and develop a VR game-based training program to support individuals to self-regulate their

stress during public speaking. In particular, E2 is aimed to answer RQ₃ through testing the following derived hypotheses:

Research Question 3 (RQ₃): How can the Stress Detection Model (SDM), developed based on the identified set of voice features indicative of stress, be effectively used to provide real-time feedback?

More specifically:

- Can integrating the SDM into two different game modes within VR (supportive feedback [SF] and competitive feedback [CF]) be effectively used to provide real-time feedback and support to users?
- Can this integration help users reduce stress, and how effective is it in testing the stress-buffering effects of a virtual audience based on detected stress in speakers?

➤ **Hypothesis 4a (H_{4a}):** SF mode in real-time may reduce stress (perceived stress level) and may decrease blood pressure and HR because participants may feel motivated in CF mode

➤ **Hypothesis 4b (H_{4b}):** CF mode, in comparison to SF mode may increase stress (perceived stress levels) and may increase HR and blood pressure because participants may feel stressed in it.

The two hypotheses were merged to form a single hypothesis, **Hypothesis 4 (H₄):** The Supportive Feedback mode, using real-time feedback from the integrated Stress Detection Model (SDM), will more effectively reduce stress (as indicated by perceived stress levels, HR, and blood pressure) in speakers compared with the Competitive Feedback mode.

The purpose of this research is not to equate good public speaking performance solely with reduced stress. Instead, the focus is on addressing one of the primary challenges faced during public speaking, which is stress. Stress can lead to a sudden onset of various physiological changes, significantly impacting even the most well-prepared speakers,

particularly their voice. The goal of the SDM integrated within VR modes is to provide individuals with a tool to manage stress in real-time. By offering a safe and supportive environment with consistent and realistic feedback, the VR system aims to help speakers build resilience, and learn to regulate their stress in voice, improve their confidence and better prepare for public speaking scenarios.

5.3 Design of Experiment 2

The design of this experiment focuses on creating two distinct scenarios, or game modes, using VR, named SF and CF, respectively. The integrated SDM assesses stress in the speaker's voice in real-time, providing a continuous score that reflects the detected stress levels. Based on the score, appropriate feedback from the virtual audience is triggered, according to the selected game mode. The impact of real-time virtual audience feedback on the speaker's stress levels in both modes is assessed later via a questionnaire and physiological parameters. These conditions are detailed in Table 5.1.

Table 5.1 *SDM Integrated Game Modes in VR for Public Speaking*

Game Mode	Description
Supportive Feedback (SF)	The supportive feedback mode consistently offers positive reinforcement to the speaker by providing supportive response whenever the speaker reaches certain performance thresholds. This mode begins with the virtual audience appearing neutral but interested, creating a baseline of attentive listening. This mode focuses on encouraging the speaker regardless of their performance. As the speaker progresses, the virtual audience maintains a visually supportive demeanour, showing attentive listening and engagement. This positive reinforcement is designed to boost the speaker's confidence and reduce stress.
Competitive Feedback (CF)	The Competitive Feedback mode offers feedback to the speaker that directly correlates with their detected stress levels. In this mode, the

virtual audience starts with a neutral appearance, indicating a baseline level of attentive interest.

As the speaker's stress level increases, the virtual audience's behaviour automatically changes to reflect decreasing interest: first, they begin looking at their phone screens, then some start talking on their phones, and if the speaker remains highly stressed, the audience may eventually boo. In contrast, if the speaker demonstrates lower stress levels and performs well, the audience's visual feedback becomes more positive: they appear more attentive and quieter, eventually applauding when the speaker reaches a certain performance threshold.

In the SF mode, the audience's reactions are designed to support the speaker's emotional state. If the speaker demonstrates a high level of stress, the audience provides an encouraging response by clapping. The unconditional supportive behaviour in SF mode is aimed to reassure the speakers and motivate them to improve despite their stress (Pan et al., 2018; Schenkler et al., 1982). On the other hand, when the speaker performs exceptionally well and shows minimal stress, the audience's applause serves as a reward for their effective performance. The consistent positive feedback helps to create a nurturing environment, encouraging the speaker to focus on improving without the fear of negative judgement (Blackwell et al., 2007; Deci et al., 1991; Hattie et al., 2007). The SF mode emphasises a nurturing and positive atmosphere, aiming to help the speaker feel supported and confident regardless of their performance level. The audience's reactions are designed to uplift and motivate the speaker, ensuring that they receive consistent encouragement whether they are struggling with stress or excelling in their presentation.

Conversely, the CF mode is designed to simulate a scenario where the speaker is effectively competing against their own stress levels. The speaker is encouraged to manage and reduce their stress in real-time, with the audience's reactions providing immediate, context-

specific feedback. The CF mode aims to create a more challenging environment to help speakers improve their stress management skills and, ultimately, enhance their overall public speaking performance. Although the audience's applause serves as positive reinforcement when stress is managed effectively, it is important to note that speakers may not always be explicitly informed of this feedback mechanism. As a result, the impact of the CF mode may vary depending on whether participants recognise that managing their stress leads to positive audience reactions.

The SF mode provides feedback through consistent positive reinforcement, which acts as a support system for the speaker. Positive feedback encourages the speaker to continue trying and improving, with the virtual audience's supportive behaviours reinforcing their efforts. This approach helps build the speaker's confidence and reduce stress by creating a safe environment to practice public speaking. While the SF mode emphasises encouragement, the VR system also includes a CF mode, where the speaker can encounter more challenging, less engaged audience behaviours. This dual-mode design ensures that speakers gain confidence through positive reinforcement and also develop the ability to manage stress in more realistic scenarios, preparing them to handle disinterested or critical audiences in real-life situations.

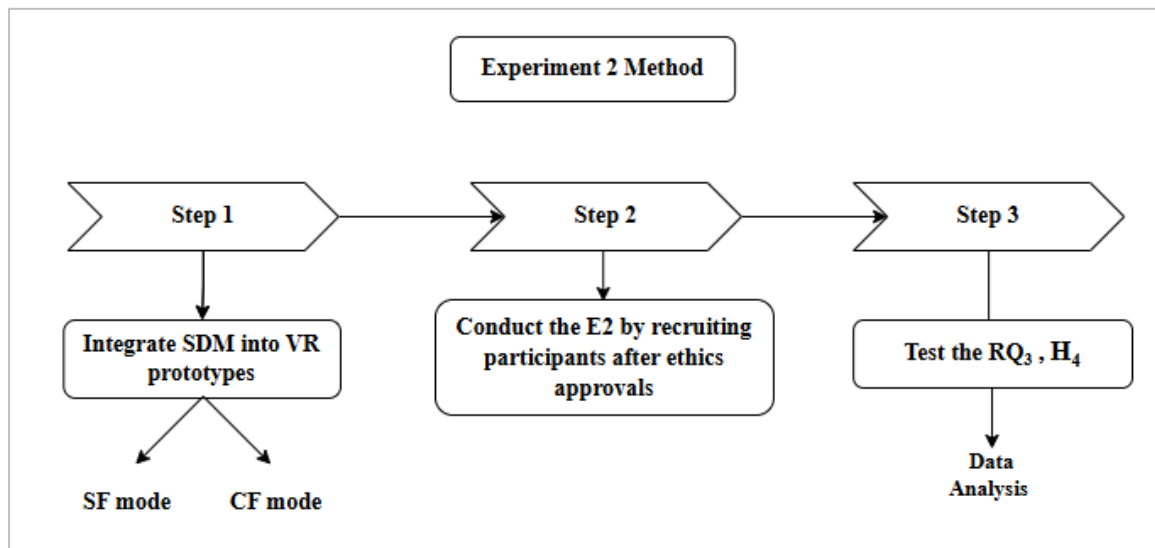
By providing real-time feedback based on stress levels, the experiment was intended to help speakers become more aware of their stress and learn voice techniques to manage stress effectively. The immediate and dynamic feedback from the audience was expected to motivate speakers to improve their performance by reducing stress, leading to better engagement and delivery. The speakers were expected to discover ways to manage their stress, such as adjusting their tone or incorporating pauses, through repeated playthroughs of the game. Specific techniques to reduce stress were not provided, as this was not the purpose of the experiment. Since stress impacts voice features such as tone, pitch and speech patterns, the SDM was designed to focus on these characteristics rather than linguistic proficiency. The SDM, trained

predominantly on voice samples from nonnative English speakers, can detect stress regardless of a speaker's fluency in English or even if they do not speak English. The experimental aim was to gather data on how different levels of stress affect a speaker's performance and how immediate feedback based on real-time analysis of stress can influence their behaviour and stress management strategies. By comparing the responses of speakers in the CF mode with those in the SF mode, the experiment sought to validate the effectiveness of supportive feedback verses competitive feedback in enhancing speakers' performance. For filling in of the questionnaires and demographic questionnaire, Qualtrics (2018) was used.

5.4 Methods of Experiment 2

The methods for E2 (Figure 5.1) involves integrating the developed machine learning-based SDM in the various prototypes of VR audiences to exhibit different behaviours, as defined by the previously established modes. These prototypes were then used to test hypotheses via E2.

Figure 5.1 *Methods for Experiment 2*



Note. SDM = Stress Detection Model, VR = Virtual Reality, SF = Supportive Feedback, CF = Competitive Feedback, E2 = Experiment 2, RQ3 = Research Question 3, H₄ = Hypothesis

In E2, the first step was to implement the developed SDM to work with both VR game modes and corresponding virtual audience behaviours. The second step involved conducting E2 itself, by recruiting participants and checking their engagement with the VR scenarios as well the impact of the corresponding feedback types on their stress levels. Participants were asked to give speeches in either one of the VR modes. During speeches, physiological parameters (systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (HR), self-reported perceived stress, arousal and voice samples) were recorded. The third step involved data analysis of physiological parameters and self-reported questionnaires (collected at before and after the experiment) to validate the stress arousal capability of the VR scenarios. Thus, E2 served to conclude the research by testing out the effectiveness of implementing the affective computing model into a VR scenario for public speaking.

5.5 Development of the Game Modes for E2

5.5.1 *Game Modes*

The final VR application comprised two game modes – SF and CF. These game modes were integrated with the real-time SDM. These modes were developed based on the hypothesis to be tested in E2 as described earlier in Section 5.2. Both modes share a similar visual design, including an auditorium setting with a high ceiling, as depicted in Figure 5.2. This design choice was made to enhance the realism and immersion of the scenario, particularly when the speaker looks around in the environment. Additionally, the virtual audience size in each mode is consistent with the specifications described in Section 3.5.1 of Chapter 3.

Figure 5.2 *Supportive Feedback and Competitive Feedback Game Mode Scene*

The auditorium model used in the VR application was purchased from the Unity online asset store, specifically from 3D artist Satish Parab (<https://assetstore.unity.com/packages/3d/environments/auditorium-141127>). The purchased model is a low-poly design ready to be imported in our game development engine of choice (i.e. Unity).

In addition to the auditorium model, new functionalities were introduced to enhance the game's interactivity. A timer was implemented in each game mode, as detailed in Appendix H, which was set to 5 minutes for the duration of the user's speech. The timer feature was added to provide a structured timeframe for the tasks within the VR experience. The timer functionality is managed by a script that is attached to the Canvas in Unity and is directly linked to a user interface (UI) text component named 'TimerText'. This script initialises the timer at 300 seconds and continuously updates the remaining time each frame by decrementing the time based on the frame rate ('Time.deltaTime'). The Update () function within the script ensures that the time value decreases smoothly until it reaches zero, at which point the timer stops. The remaining time is displayed in a 'MM:SS' format on the screen, providing clear, real-time

feedback to the user. This visual feedback helps the user manage their speech duration within the VR experience.

Furthermore, another score variable was incorporated into each game mode, a significant enhancement compared to the virtual prototypes discussed in Chapter 4 Experiment 1. This scoring mechanism allows for real-time feedback and performance tracking, adding an element of gamification to the user experience and enabling more precise evaluation of the user's engagement and stress responses during the experiment.

5.5.2 Score System

The score in the two game modes is directly proportional to the predictions generated by the developed SDM as discussed earlier in Chapter 4 Section 4.2.6. The SDM was written in Python programming language, and it operated through a live testing file (Code in Appendix I) containing the last 10 seconds of audio and was updated every 2 seconds. The live testing script continuously monitors a directory for new '.wav' files that are real-time voice recordings of the speakers (Code in Appendix J).

The live audio script processes the most recent voice file to extract the 32 MFCC features (4.2.6.1) and uses the pretrained SDM to predict the stress levels in the speaker's voice, as illustrated in Table 5.2, which presents the final prediction values output by the Feedforward Neural Network (FFNN) (explained in 4.2.6.1). When any speaker is stressed, the SDM will return the predictions like those shown bold in Table 5.2. When the speaker is stressed, the SDM will return predictions that exceed the threshold of 0.5 and are closer to 1, as shown in bold in Table 5.2. For voice sample 1, the SDM score of 2 indicates that only two values in the predictions exceeded the threshold, suggesting that the person speaking is not stressed. In contrast, for voice sample 2, more than 50 per cent of the predictions (seven values, as highlighted in bold) exceed the threshold, which indicates that the speaker is stressed.

Table 5.2 *Live Stress Detection on Two separate on Voice Samples*

Values indicating low stress in the voice sample 1	Values indicating high stress in the voice sample 2
0.01864511	1.92695E-08
0.03858998	0.99948466
0.001026214	0.9999881
0.005747156	0.99683458
0.00150659	0.71633381
0.026477879	0.002423578
0.001896939	0.008920915
0.000641325	0.97548699
0.003970414	0.9999854
0.90726006	0.000188922
0.9995579	0.55333972
2	7

The stress prediction values from the SDM are then saved in a readable text file. This text file is read by a separate Unity script responsible for score management, written in C#. This adjustment script adjusts the score range (Code also provided in Appendix K), and updates the final score displayed in the UI. This works basically in the same way for both the SF and CF scenarios.

The score adjustment script includes initialisation methods, such as start and reset functions. There are few key variables in this script, such as score (read from text file), which is adjusted based on specific thresholds:

- (i) n (initially set to 5 in beginning), to check if there was a change in the value of score and old score (set to 0 in initially). If these conditions are met, the score is increased or decreased based on other defined factors such as the following.
- (ii) The variable 'step' (set to 10) represents penalty or bonus in score.
- (iii) The variable 'c' is a numerical value and serves like a penalty or bonus, if the score in comparison to the old score keeps decreasing and is used as bonus.

- (iv) Additionally, a ‘help factor’ (set to 2) is used to prevent the score from becoming negative, assisting the speaker. The old score is then updated with the current new score.

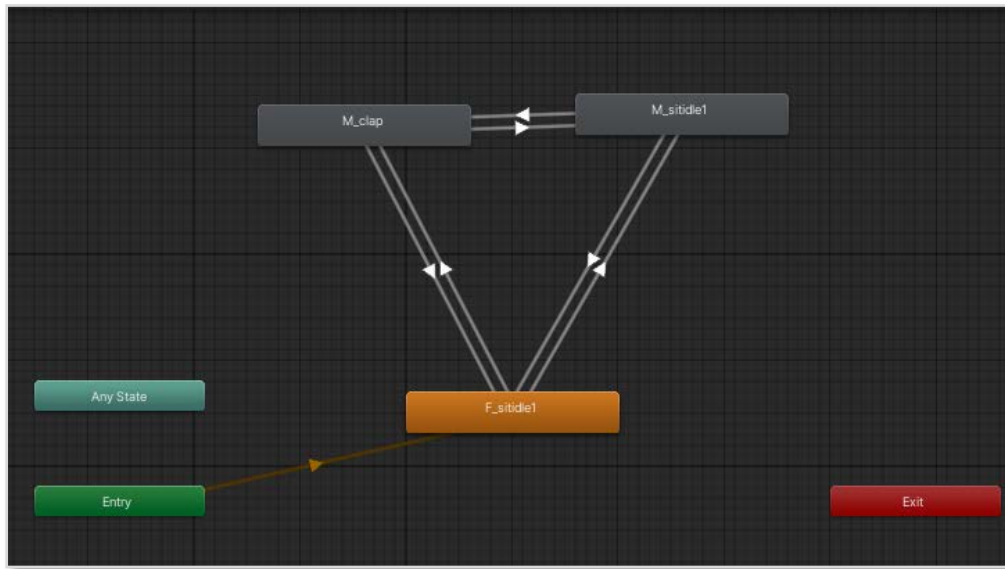
The same score adjustment script is then linked to update the CF mode’s UI elements, score, corresponding animations and audio. However, because the SF mode is intended to be more supportive than the CF mode, the same score is used to create a separate SF scoring system in Appendix L. The SF score system script incorporates additional variables such as ‘help factor’, ‘step’, ‘c’, and ‘n’ which add a dynamic, flexible aspect to the SF scoring mechanism. This dynamic scoring system in the SF mode reflects the speaker’s progress in real-time, based on data from the SDM, and is linked to another script that updates the UI elements, score, corresponding virtual audience animations and audio in the SF mode.

5.5.3 Animation and Audio

The SF mode provides only positive supportive feedback regardless of how much stress is detected from the speaker’s voice, or how low the score goes for a particular individual. Accordingly, the animations and audio in SF mode are positive, and attentive only. The speech is supposed to last for 5 minutes at most. This mode starts with a score of 0, with the virtual audience in the VR game starting in a quietly sitting animation and can take values between -100 (extreme stress detected) and +100 (extreme confidence). At the same time, the SDM starts analysing the stress in the speaker’s voice in the background and updates it in the score text file. The score adjustment script then reads the text from the score text file, and accordingly the animator parameters of the attached controller are triggered leading to different animations to set in motion and play in real-time with their supporting audios.

To determine the virtual audience behaviour in the game modes, a set of parameters are used as triggers. Specifically, the SF mode included three parameters, which were managed by the live score adjustment script as described in Appendix L. According to the changes in the

score, if any of the parameters (M_clap, M_sitidle1, or F_sitidle1) was triggered, then the respective animation in Unity is played. All animation states of the virtual audiences are illustrated in Figure 5.3.

Figure 5.3 Animation States of the SF modes based on the stress levels predicted

For example,

- (i) If a new score is equal or greater than 0, the *sit idle* animation is played
- (ii) If the new score is less than 0, and greater than -79, a different *attentive sitting* animation is played
- (iii) If the new score is less than -80 and is greater than or equal to -100, *clap* animation is played to show support and encouragement despite the high stress detected in the performance

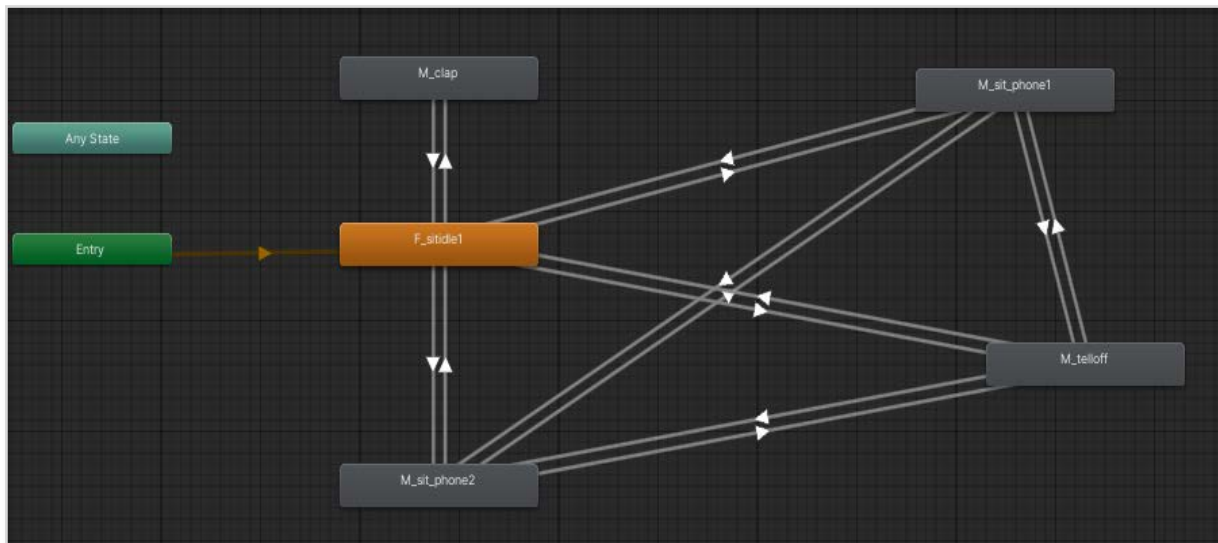
On the other hand, CF mode provides appropriate feedback for both good and bad performances. It includes SF (when the speaker shows no stress in the voice) and negative feedback (when the speaker shows stress in the voice). Accordingly, the animations in CF mode are designed to provide both supportive and unsupportive feedback, reflecting a range of reactions from encouragement to criticism.

The score in the CF scene changes in VR according to the stress levels predicted from the SDM. Similar to SF mode, the speech is supposed to last for no more than 5 minutes. This mode also starts with a score of 0, and therefore the virtual audience in the VR game sits quietly for the time being. As the speaker begins, the SDM in the background starts analysing the stress

in the speaker's voice and updates the score file. Accordingly, the CF behaviour setting script triggers different animations to be set in motion and play in real-time with their corresponding audios.

The parameters for CF mode are slightly different from those used in the SF mode. For example, in the CF mode, according to the changes in the score, the parameters (here *M_clap*, *M_sit_phone1*, *M_sit_phone2*, *M_telloff*, *F_sitidle*) are eventually triggered, leading to the respective animations (*M_clap*, *M_sitidle1*, *F_sitidle1*) being played. All animation states for the CF mode are illustrated in Figure 5.4.

Figure 5.4 Animation controller in the Unity game engine for the CF mode



In CF mode, the score also goes from -100 to 100.

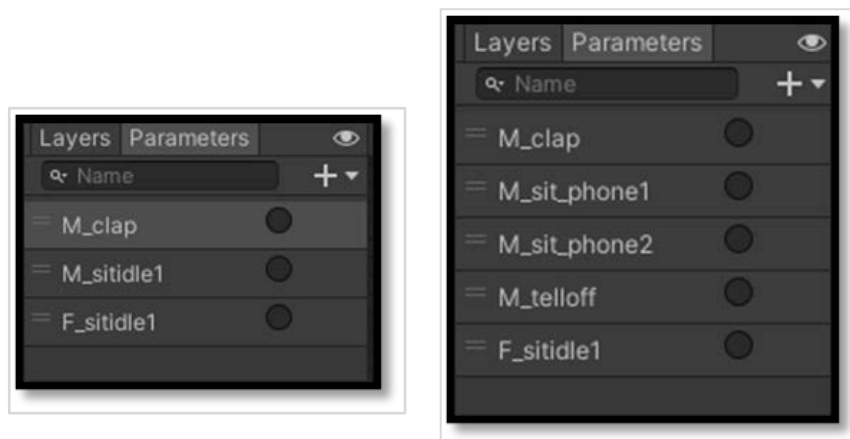
- (i) if a score equals or lies between 65-100, the virtual audience acknowledges a good performance and then starts *clapping*
- (ii) if the new score is between or equals 0 and 64, the virtual audience will *sit idle and listen attentively* to the speaker
- (iii) if the score goes below 0 and stays until -75, the virtual audience will *start checking their phones*, showing disinterest in the speaker's performance

- (iv) if the score lies between -75 and -80, they start *making calls on their phones* and *start chatting* in the auditorium
- (v) lastly, if the new score is lower than -80, the virtual audience starts *booing* the speaker.

In both SF mode and CF mode, the animation parameters were set as trigger parameters (indicated by the round circles beside each defined parameter), rather than using Boolean, integer or float values. However, the specific parameters differ between the two modes.

For SF mode (Figure 5.5), the parameters include M_clap, M_sitidle1, and F_sitidle1. These triggers are managed by the live score adjustment script to transition between different supportive animation states in the SF animator controller. In contrast, CF mode (Figure 10) utilises a different set of trigger parameters, including M_clap, M_sit_phone1, M_sit_phone2, M_telloff, and F_sitidle1. These parameters, also controlled by the live score adjustment script, enable transitions through a range of both supportive and unsupportive animation states in the CF animator controller. Both modes use trigger parameters for dynamic animation control, but they differ in the specific animations they manage to reflect the unique feedback styles of each mode.

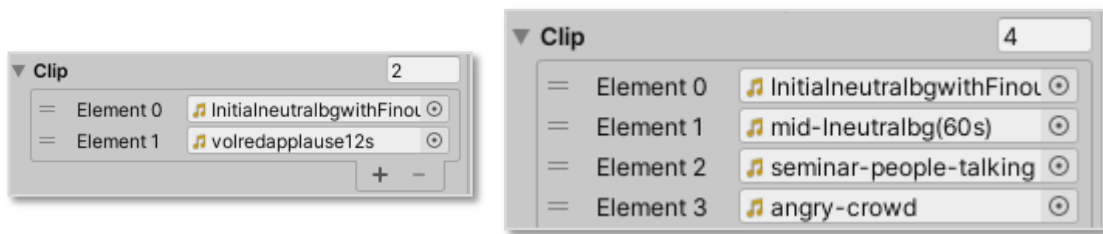
Figure 5.5 *Parameters of SF mode on left and CF mode on right*



In both SF and CF modes, the initial audio includes a neutral and quiet auditorium background sound, which is played when the audience is sitting quietly and attentively.

However, the subsequent audio clips differ significantly between the two modes, reflecting the distinct nature of each. In SF mode, the audio transitions to an applause sound effect whenever the applause animation is triggered, reinforcing positive feedback and encouragement for the speaker. On the other hand, in CF mode, the audio evolves to create a more challenging environment. After the neutral background, the second audio introduces chatting and murmuring sounds, followed by a third clip featuring people talking loudly on their phones. The final audio clip escalates to extreme and rude booing, intensifying the pressure on the speaker. These differences in audio are managed by the Audio Manager script (Appendix L) and are illustrated side by side in Figure 5.6, with SF mode on the left and CF mode on the right.

Figure 5.6 *Audio clips for SF mode on left and CF mode on right*



5.5.4 Scripts

Different scripts were created as per need of creating the SDM and integrating the model in both VR application modes. In Python programming-based files for developing the SDM, libraries were called off and various functions were defined. In Unity based C# scripts, different Boolean parameters were first called off. The defined scripts were used to trigger the animation states based on the respective animation controller; respective audios followed by respective time to stop playing the animations. The scripts included calling on various audio parameters to trigger the animator controller. Different game objects were created and attached together so as to form an object to be used in the game. The scripts also included end time for each of the animations that needed to be played.

The SF mode script (See Appendix M) in the Unity game manages and triggers animations and audios based on the player's score. Several variables are initiated in the beginning of the script such as 'static int newscore', 'public Animator anim', 'AudioSource audienceclip', 'public AudioClip[] clip' and 'public GameObject SoundManager'. The variable 'static int newscore' holds the player's current score, retrieved from Unity's 'PlayerPrefs'. The 'AudioSource audienceclip' variable holds the audio source component and is responsible for playing audio clips. The audio clip is an array of the sound clips which were defined earlier for the SF mode. The sound manager variable contains the reference to the script for managing the audio playbacks.

Various methods such as void Start (), and void Update () are used to initialise the script when the game begins; these are called once per frame and are responsible for checking and updating animations and audio based on the player's score respectively. The integer values t1, t2, t3 and t4 represent the threshold values for different score ranges, which are used to determine the animation and audio clips to be played for providing the speaker's real-time visual feedback from the virtual audiences. The SF mode script essentially adjusts the behaviour of the virtual audience (animations and audio) based on the player's performance score and provides dynamic feedback that is dedicated to providing the speaker with relevant emotional support and encouragement, in order to motivate the speaker to improve the performance.

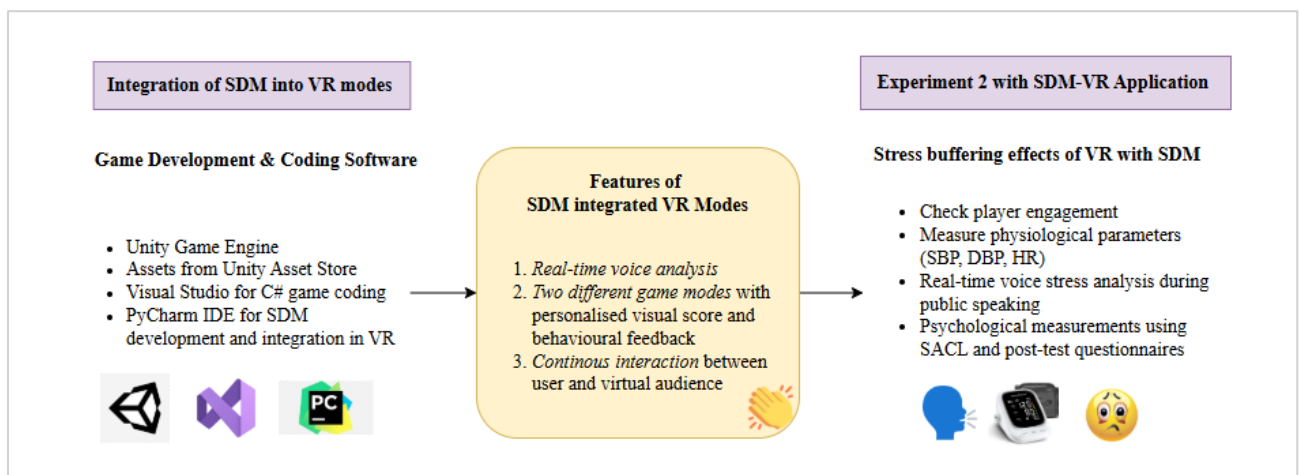
The CF mode script (Appendix M) manages, and triggers animations and audios based on the player's score reflecting the critical and negative responses. Similar variables were initiated in these scripts as that of SF mode. Functions such as void Start (), and void Update () initialise the script when the game begins and is called once per frame and is responsible for checking and updating animations and audio based on the player's score respectively. The integer values t1, t2, t3, t4 and t5 represent the different threshold values for different score

ranges, which are then used to determine the animations along and audio clips (as defined earlier in the animation states, animation controller, sound clips for CF mode) to be played for providing the speaker real-time visual feedback from the virtual audiences.

The CF mode script essentially adjusts the behaviour of the virtual audience (animations and audio) based on the player's performance score and provides real-time feedback based on the speaker performance and stress detected in their voices. The CF script is designed to provide critical and sometimes harsh feedback in response to the player's performance. As the score decreases, the audience's reactions become increasingly negative, helping to convey the consequences of poor performance, and when the score increases, the audience's reactions become positive, helping to convey the appreciation of a better performance.

The overall developmental work done for E2 prototypes is summarised in Figure 5.7. The features developed in the VR game scenario include the virtual audiences and the auditorium. Two different prototype game modes were developed using the software Unity game engine, Visual Studio and C# and Python programming language.

Figure 5.7 Overall work done to develop the prototype for Experiment 2



In the next section, we evaluate the effectiveness of the CF and SF mode prototypes.

These prototypes were designed to provide dynamic feedback based on the player's performance, with the CF mode delivering critical and negative responses to poor performance,

and the SF mode offering more supportive and encouraging reactions. The E2 aim was to assess how these contrasting feedback mechanisms influence the player's stress levels, engagement, and overall experience within the VR environment by focusing on [Research Questions and Hypothesis of Experiment 2](#). By testing these modes, we seek to understand the impact of different real-time based feedback styles on user behaviour and emotional response.

5.6 Experiment 2 (E2)

5.6.1 Participants

Participants in E2 consisted of 61 individuals aged between 18 and 40 years, recruited from James Cook University, Singapore (JCUS). This age range was chosen to capture a demographic typically in the early to midstages of their careers, where the development and refinement of public speaking skills are highly relevant. Individuals within this age group are frequently engaged in professional and societal activities that require effective communication, making them an ideal population for this study (Marinho et al., 2017).

All participants were able to understand and speak English fluently, ensuring they could fully engage with the experimental tasks. Importantly, participants were excluded for any conditions that could impact their ability to participate in the VR environment, specifically motion sickness and susceptibility to epileptic seizures, neither of which were reported during the study.

The sample size was determined based on a G-power analysis for MANOVA with between-subjects design, which recommended a minimum of 56 participants to achieve a power of 0.80. However, to increase statistical power and the robustness of our findings, a total of 61 participants were included in the study. Although this larger sample size does not align with the norms in prior research, it exceeds the typical range used in the human-computer interaction and affective computing fields, where studies usually involve 10-40 participants for

evaluating VR scenarios aimed at improving public speaking (Hartanto et al., 2014; Lindner et al., 2021).

Participants were recruited using a similar approach to that used in Experiment 1. We sought assistance from course lecturers, who were asked to display a research poster in their online and face-to-face classes (Appendix N *Experiment 2 Poster, Email to Lecturers and Social Media Post*). Additionally, participants were also recruited through word of mouth, personal contacts and social media platforms such as LinkedIn and the JCUS Telegram Psychology group, where the study poster and a related social media post were shared. The study poster featured a QR code that directed interested individuals to a Google form. This form collected essential information, including their email address and availability for participation in the experiment. Participation in the study was entirely voluntary, and participants were informed that they could withdraw at any time without any prejudice (See consent form and information sheet for Experiment 2 in Appendix N). An approval to conduct the experiment was obtained from the Human Research Ethics Committee of James Cook University prior to commencement (Approval Number: H8298; Appendix O).

From the sample of 61 participants, one participant was excluded from the dataset because they did not complete the demographic form, thereby failing to meet the inclusion criterion of providing complete demographic information. This resulted in a final sample size of 60 participants (51.66 per cent female), with most of them from JCUS ($n = 44$) and very few from social media and contact connections. Participants' mean age was 25 years ($SD = 6.34$), ranging between 18-40. Most participants were from Singapore (30 per cent), the rest were from China (15 per cent), Myanmar (13.3 per cent), India (10 per cent), Malaysia (6.7 per cent), Germany (3.3 per cent), Vietnam (3.3 per cent), Philippines (1.7 per cent), Indonesia (1.7 per cent), USA (1.7 per cent), Poland (1.7 per cent), Belgium (1.7 per cent), France (1.7 per cent), Venezuela (1.7 per cent), Brunei (1.7 per cent), Russia (1.7 per cent), and Japan (1.7 per cent).

Fewer than half (41.7 per cent) had English as their mother tongue. The majority (68.3 per cent) of the participants in the experiment were full-time students from JCUS, while the remaining 31.7 per cent were full-time employed or self-employed individuals who were either part-time students at JCUS or industry professionals recruited through the researcher's network, including JCU Alumni.

5.6.2 Measurements and Equipment

Stress Arousal Check List (SACL): The SACL is a 20-item checklist measuring the experience of stress and arousal experienced immediately after the task (King et al., 1983). The SACL includes a stress subscale and arousal subscale with 10 items for each subscale. Participants rated each item on a scale from 0-3; where (0) = Not at all, (1) = A little, (2) = Quite a bit, and (3) = Very much. Higher scores indicate a greater intensity of feelings. Sample items in the checklist include 'calm', 'comfortable', 'uneasy', 'tense' and 'aroused'. The SACL has been shown to have a high internal consistency of 0.86 for the stress subscale (King, Burrows, & Stanley, 1983). For the present study, the Cronbach's alpha for the 10-item stress subscale of the SACL was greater than .70. Specifically, for both the pretest and posttest, the Cronbach's alpha calculated for the stress subscale was .84. The Cronbach's alpha for the pretest arousal 10-item subscale was .77, and for the posttest arousal it was .75.

Heart Rate & Blood Pressure: HR and blood pressure are physiological parameters that are reported to be raised under conditions of stress (Ursin et al., 1993). They can easily be measured with little inconvenience. HR and blood pressure were measured using the OMRON Automatic Blood Pressure Monitor HEM-907. The OMRON Automatic Blood pressure monitor allowed evaluation of the effect of the virtual audience's responses in the VR game on the speaker's HR and blood pressure and manual recording of the readings in real-time. A separate Polar H10 HR monitor as shown in Appendix D was also used to record the HR (and HR variability) for the complete 30-minute experiment duration. Both the OMRON HEM-907

automatic blood pressure monitor (El Assaad et al., 2002) and the Polar H10 (Speer et al., 2020) are clinically validated and for that reason were employed as suitable devices for getting the respective average and continuous readings.

Posttest Questionnaire: The posttest questionnaire was specifically developed for this study to assess participants' personal opinions about public speaking and VR application and the developed scenarios. It is a 21-item questionnaire. Items are scored on a scale of 1-7, where (1) = Strongly disagree, (2) = Disagree, (3) = Somewhat disagree (4) = Neutral, (5) = Somewhat agree, (6) = Agree, (7) = Strongly agree. This questionnaire included items such as 'During the speech, I was not expecting the sudden responses from the audiences', 'During the speech, I felt that I was evaluated by the audience', 'I felt pleasant about the speech I gave', and 'During the speech, I had sudden feelings of uncertainty'. Ten items related to perceived feedback from the virtual audience. The questionnaire also included 11 statements about the general experience of the speaker using VR and public speaking such as 'I have tried virtual reality (VR) before' and 'I usually have difficulty in giving a public speech' and it was given at the end of the speaking task in each condition (Appendix D).

Demographic Questionnaire: The demographic questionnaire as shown in Appendix N was also included in E2. The items include age, occupation, country of birth, mother tongue and number of months remaining to course completion. These questions allowed checks for the effects of age, occupation, country and first language on SBP and DBP, HR, stress arousal checklist responses and posttest questionnaires of the participants.

Devices: An HP OMEN gaming laptop and HTC Vive were used to play the VR scenarios during the experiment sessions. While the laptop recorded participants' voices in real-time every 10 seconds for analysis by the SDM, an iPhone 7 was also used to record their speeches. The iPhone recordings were kept as a backup in case separate analysis of the audio data was needed later.

5.6.3 Procedure

The experiment was conducted using a one-way between-subject design. Each participant went through only one of the modes (either social SF or CF). The study was conducted at the psychophysiology laboratory at JCU Singapore. Upon the arrival of participants, 5 minutes were allocated to introduce the project, task, and devices (Polar H10 HR monitor, OMRON Automatic Blood Pressure monitor HEM – 907, iPhone 7, HP OMEN gaming laptop, and HTC Vive) that would be used in the experiment, as well as how these devices would be utilised.

After an initial introduction to the experiment, the participant provided consent for their voice to be recorded, to give a public speech, to complete the questionnaires and for their physiological parameters of HR and blood pressure to be recorded.

Next, the participant was requested to wear the Polar H10 HR monitor around their chest as per instructions and its operability was checked. After this, the participant was asked to sit in upright posture on a chair, keeping their left arm straight on the table, to facilitate placement of the OMRON automatic blood pressure monitor's cuff on the participant's left arm. Blood pressure measurements are often taken on the same arm each time to ensure consistency and comparability of results (in our case, we kept it to left arm). After waiting 5 minutes, two baseline readings of HR and blood pressure (both systolic and diastolic) were taken using the OMRON monitor. Next, the SACL was used to measure the participant's stress and arousal.

Following this, the Polar H10 was set to 'on'. Each participant was randomly assigned to either the SF or CF mode. The experimenter highlighted that the public speaking task was designed to assess certain psychological capabilities. The experimenter highlighted that research has shown that public speaking skills, specifically verbal intelligence, tone of voice, breathing pattern, poise, and overall performance on these tasks predicts how successful a

person will be in their relationships or occupation, whatever that may be. The participant was then assigned to speak on the topic of ‘why pornography should be banned’. The given topic was suggested to induce stress in the speaker and no other topic choice was given. The experimenter emphasised the importance of speaking clearly for the entire duration. The participants were then given 2 minutes and 30 seconds of time to prepare the points they wanted to speak about on the topic. Following this, the experimenter took a back seat in the same room. No writing materials were provided for speech preparation. The preparation time allowed participants to clear their minds about specific points they were going to make in their speech and ensured them speaking for the full duration of 5 minutes. The participants did not have access to any written materials or notes to assist them while speaking, as they were wearing the VR HTC Vive headset during the task. Once the speech preparation timer was off, and the participant was ready, the experimenter provided the HTC Vive VR headset and helped in fixing it properly on the participant. Next, the experimenter started the VR mode in the HP OMEN gaming laptop and asked the participant whether the VR scenario could be seen in the VR headset view, and audio was also tested. After everything was set in motion and the participant agreed to being fully ready to speak, the VR simulation scenario was begun. The experimenter set the voice recording in the iPhone 7 to ‘on’.

The participant spoke either in SF mode or CF mode. In both modes, the speaker’s stress level was visually displayed in real-time as a score ranging from -100 to 100. According to the score and the game mode in which the speaker was speaking, a real-time response was displayed by the virtual audience.

The scoring system in the SF and CF mode was designed to continuously indicate the speaker’s stress levels throughout the task and provide the relevant feedback (Section 5.5.2). This scoring persisted until the end of the 5-minute speaking period. In the SS mode, a third reading of the participant’s SBP, DBP, and HR was taken at the first 60 seconds of the speech

within the VR simulation mode, followed by three additional readings at approximately 120 seconds, 180 seconds and 240 seconds into the task. Similar in the third, fourth, fifth and sixth SBP, DBP and HR readings were taken at around 60 seconds, 120 seconds, 180 seconds and 240 seconds into the CF mode. Regarding the scoring, when the score increased, it reflected a decrease in stress levels, resulting in fewer triggered supportive behaviours. A summary of how long participants displayed positive versus negative scores at each minute was not recorded, as the real-time SDM could only indicate the stress levels experienced by the individual in the present moment. Once the participant finished speaking in the assigned condition, the iPhone recording (which was recoded in case further analysis was needed) was turned off and saved. The HTC Vive VR headset was then removed, and the participant was provided with the SACL and a questionnaire to complete as a posttest.

Different subsets of the virtual audience provide different feedback. The score system was specifically developed to control the behaviour of the virtual audience such that it did not fluctuate too much during the speech. Stress levels were only indicated via score, and behaviour was only triggered with extreme changes in the stress levels.

The voice of each participant giving a speech was recorded in both modes (as a third dependent variable to be used for further analysis, in case it was needed). The voice data would be quantified by analysing various feature parameters MFCCs using the SDM model. Next, a set of demographic questionnaires was administered towards the end of the experiment. After the completion of the study, participants received SGD10 Grab e-voucher via email or physical SGD10 NTUC FairPrice voucher on the spot.

5.7 Results

5.7.1 Data Preparation

The baseline readings of the SBP, DBP and HR were averaged using the first baseline and second baseline reading (Baseline_SBP, Baseline_DBP, Baseline_HR). The scores of the same parameters were also calculated by computing an average of the four readings recorded (Speech_SBP, Speech_DBP, Speech_HR) when the person gave a speech in the SF or CF mode at 60 seconds, 120 seconds, 180 seconds and 240 seconds (the third, fourth, fifth and sixth readings).

The scores for the stress and arousal checklist items for the baseline normal readings (Pre_stress, Pre_arousal) in the SF and CF modes were computed by taking the average of the stress subscale items and arousal subscale items. The stress subscale contained 10 items, and the anxiety subscale contained 10 items. Similarly, the scores for the stress and arousal checklist items for the baseline normal readings (Post_stress, Post_arousal) in the SF and CF modes were computed by taking the average of the stress subscale items and arousal subscale items.

The final readings to see the changes in the average stress (Average_stress) and average arousal (Average_arousal) were calculated by subtracting the Pre_stress and Pre_arousal readings from the Post_stress and Post_arousal readings.

For the stress and arousal readings, the scores for the subscales were reversed as per the need. Then an overall average was computed for both the stress and arousal subscale for game modes.

5.7.2 Assumption Testing

All results were analysed using SPSS version 26 with alpha level set at .05. Prior to analysing the data, several assumptions were tested. Each participant took part only once in the experiment and was asked not to share the details of the experiment with others. This was done to meet the assumption of independence of errors.

The assumption of normality was examined using the Shapiro Wilk statistic and histograms. Results showed that the normality assumption for physiological parameters (Baseline_SBP, Baseline_DBP, Baseline_HR, Speech_SBP, Speech_DBP, Speech_HR,) and computed average scores of stress and arousal (Pre_stress, Pre_arousal, Post_stress, Post_arousal, Average_stress, Average_arousal) from the SACL subscale items was met except for the Pre_Stress, and Speech_HR. A further visual inspection of the histograms showed approximate normal distribution, and no outliers were observed.

Second, the absence of univariate outliers' assumption was checked using boxplots. The boxplots results indicated no outliers for dependent variables baseline, speech DVs. Third, the Mahalanobis Distance coefficient was 22.80. This value was greater than the critical chi-square value, $\chi^2 (5) = 20.51$, at an alpha level of .001, implying that the assumption for absence of multivariate outliers was not met. Therefore, we removed one participant's data (P29) with a Mahalanobis Distance higher than 20.51. Fourth, the inspection of the scatterplot between each of the dependent variables indicated the assumption of linearity was met. Multicollinearity between dependent variables was not of concern by assessing the scatterplot of standardised predicted and residual scores.

5.7.3 Preliminary Analysis

Preliminary analyses were carried out to check any effect of the variables such as age, occupation, gender, first language, country of birth, number of months to course completion and the posttest questionnaire on dependent variables (DVs). The DVs are Baseline_SBP, Baseline_DBP, Baseline_HR, Speech_SBP, Speech_DBP, Speech_HR, Pre_Stress, Pre_Arousal Speech_SBP, Speech_DBP, Speech_HR, Post_Stress, and Post_Arousal.

To assess the relationship between age and DVs, bivariate Pearson's product moment correlation coefficients (r) were calculated. The bivariate correlations between age and all the DVs were not significant (Table 5.3). A Pearson's product moment correlation coefficient (r)

was calculated to assess the relationship between number of months left to complete the course and DVs. None was significant.

An independent samples *t*-test was used to examine the effect of gender on the DVs in the SS and CF modes respectively. All results were not significant, indicating there were no significant effects of gender on the DVs. Furthermore, multiple one-way between-subjects ANOVAs were carried out to examine the effects of nationality, first language and occupation on the DVs. These results showed that these factors had no significant effects on the DVs (Table 5.4). Lastly, a separate correlation test was conducted with the posttest questionnaire items across the responses obtained from both SS and CF conditions (Tables 5.5 and 5.6) to help understand the data.

Several questionnaire responses in CF mode such as Q14, 'During the speech, I felt threatened' shows a strong positive correlation with post stress, suggesting that feeling threatened during the speech is strongly associated with higher poststress levels. Another significant positive correlation in CF mode included Q15, 'During the speech, I felt challenged' which correlates positively with post stress, indicating that feeling challenged during the speech is linked to higher poststress levels.

The results in the SF mode had several significant correlations such as Q2, 'I am interested in improving my public speaking skills' had a significant positive correlation with post arousal. This indicates that participants who were more interested in improving their public speaking skills reported higher levels of arousal experienced during the speech. Another question Q17, 'During the speech, I felt that I was evaluated by the audience' is positively correlated with post arousal, indicating that feeling evaluated by the audience is associated with higher arousal levels after the speech. Several other questions such as Q15, 'During the speech, I felt challenged' is positively correlated with Speech HR and speech arousal, indicating that feeling challenged during the speech is linked to higher heart rate and arousal during the speech.

Another Q18 in SF mode, ‘During the speech, I had sudden feelings of uncertainty’ shows a significant positive correlation with post stress, suggesting that feelings of uncertainty during the speech are associated with higher stress levels afterward.

Table 5.3 *Correlation Results*

Variables/ Pearson's correlation	Baseline _SBP	Baseline _DBP	Baseline _HR	Pre _Stress	Pre _Arousal	Speech _SBP	Speech _DBP	Speech _HR	Post _Stress	Post _Arousal	Average _Stress	Average _Arousal
Age	.40	.17	-.13	-.40**	.13	.14	.09	-.12	-.15	.09	-.25	.17
No. of Months to Course Completion	.00	.01	.01	.05	-.58	-.12	.04	.29*	.00	.23	-.02	.23

Note. This table presents Pearson's correlation coefficients between Age and Number of Months to Course Completion with various physiological and psychological measures at baseline (Baseline), before the speech (Pre), during the speech in both conditions (Speech) and average (Average) change in the pre and post readings. The measures include SBP, DBP, HR, stress and arousal levels. Correlation coefficients marked with statistically significant * $p < .05$, ** $p < .01$.

Table 5.4 *T-Test and ANOVA Results*

Variables/ <i>F</i> -statistic	Baseline _SBP	Baseline _DBP	Baseline _HR	Pre _Stress	Pre _Arousal	Speech _SBP	Speech _DBP	Speech _HR	Post _Stress	Post _Arousal	Average _Stress	Average _Arousal
Country of Birth	1.02	.54	.70	2.13	1.26	.43	.29	.50	.67	.15	.86	1.30
First Language	.00	.01	.01	2.54	1.71	.48	.77	.47	.40	1.93	1.93	2.88
Occupation	.21	.19	.34	1.98	.34	1.65	.42	.06	.71	.30	2.37	.47
Gender	1.2	7.6	1.2	1.4	.96	1.37	2.95	2.94	2.50	.08	.89	3.81

Note. This table presents the *F*-statistics from t-tests and ANOVA analyses for the effects of country of birth, first language, occupation and gender on various physiological and psychological measures at baseline, before the speech (Pre), and during the speech. Measures include SBP, DBP, HR, stress and arousal levels. Significant results are indicated with * $p < .05$, ** $p < .01$

Table 5.5 *Correlation Results for Posttest Questionnaire Items in CF condition*

Item/Pearson Correlation (<i>r</i>)	Baseline _SBP	Baseline _DBP	Baseline _HR	Pre _Stress	Pre _Arousal	Speech _SBP	Speech _DBP	Speech _HR	Post _Stress	Post _Arousal
Q1	0.13	0.19	0.06	0.09	-0.01	0.05	0.10	-0.00	-0.05	0.09
Q2	-.37*	-0.25	-0.20	-0.32	.36*	-.39*	-0.33	-0.03	-0.06	0.21
Q3	-.47**	-.43*	-0.24	-0.06	-0.05	-.52**	-.46**	-0.21	-0.06	0.10
Q4	-0.07	-0.17	-0.06	0.10	0.09	0.01	-0.24	0.01	-0.27	0.11
Q5	-0.12	-0.28	0.05	-.46**	.36*	-0.14	-0.35	0.17	-0.27	0.23
Q6	-0.02	-0.13	0.15	-0.23	0.19	0.01	-0.20	0.33	0.03	-0.00
Q7	-0.29	-0.34	-0.09	0.01	-0.05	-0.19	-0.3	0.02	-0.06	-0.06
Q8	0.04	0.18	0.03	0.09	-0.12	-0.00	0.22	-0.13	0.10	-0.03
Q9	-0.09	0.04	-0.13	-0.26	-0.21	-0.11	0.00	-0.06	0.26	-0.13
Q10	0.06	0.27	-.46*	0.11	-.40*	0.05	0.08	-.51**	-0.22	-0.08
Q11	0.1	0.16	-0.21	.42*	-0.32	0.00	0.07	-0.26	-0.19	-0.22
Q12	0.07	0.02	-0.02	-0.34	.43*	0.09	0.00	0.06	-0.03	0.34
Q13	-0.08	0.06	-0.23	0.03	-0.31	-0.17	-0.08	-.41*	-0.21	0.02
Q14	-0.00	0.17	0.09	-0.09	-0.02	0.11	0.29	0.19	.66**	0.05
Q15	-0.02	0.00	0.25	-0.33	-0.02	0.00	0.03	0.32	.38*	0.34
Q16	0.11	0.07	0.17	0.12	-0.19	-0.01	-0.13	0.00	-0.07	0.00
Q17	-0.12	-0.00	-0.02	-0.17	0.19	-0.21	-0.09	-0.10	0.04	0.16
Q18	0.28	0.28	0.13	-0.32	0.27	0.21	0.18	0.30	0.25	0.18
Q19	-0.35	-.36*	0.01	-0.10	0.15	-0.31	-0.30	0.18	0.16	0.06
Q20	-0.31	-0.19	-0.03	-0.02	-0.0	-0.21	-0.07	0.17	0.31	-0.17
Q21	-0.21	-0.13	-0.10	0.16	-0.21	-0.19	-0.26	-0.02	-0.09	-0.11

Table 5.6 *Correlation Results for Posttest Questionnaire Items in SF condition*

Item/Pearson Correlation (<i>r</i>)	Baseline _SBP	Baseline _DBP	Baseline _HR	Pre _Stress	Pre _Arousal	Speech _SBP	Speech _DBP	Speech _HR	Post _Stress	Post _Arousal
Q1	-0.09	-0.08	-0.11	-0.23	0.01	0.06	0.06	-0.17	-.51**	-0.07
Q2	0.08	0.10	-0.13	-0.06	-0.17	0.09	0.10	-0.17	-0.29	.52**
Q3	0.02	0.01	-0.01	0.02	0.22	-0.08	0.03	-0.08	-0.03	0.03
Q4	-0.16	-0.24	-0.25	-.37*	0.11	-0.12	-0.27	-0.29	-.44*	-0.05
Q5	0.12	0.21	-0.15	0.08	-0.29	0.05	0.21	-0.07	-0.10	0.22
Q6	0.31	.41*	0.05	-0.21	0.01	.45*	.44*	0.17	0.16	0.13
Q7	0.22	0.35	-0.01	-0.12	-0.00	0.21	0.24	0.15	-0.02	0.09
Q8	-0.10	-0.17	0.10	-0.04	-0.17	-0.04	-0.13	0.24	0.08	0.05
Q9	-0.10	-0.03	-0.08	0.20	0.01	-0.28	-0.24	0.03	.38*	0.00
Q10	0.13	0.18	-0.08	-0.28	0.09	0.28	0.29	0.00	0.22	-0.02
Q11	-0.00	-0.12	-0.15	0.23	-0.12	-0.01	-0.08	0.00	-0.06	-.41*
Q12	-0.27	-0.13	0.02	0.01	0.16	-0.04	-0.10	0.11	0.07	-0.08
Q13	0.14	0.11	-0.15	-0.06	0.11	0.21	0.32	-0.06	-.37*	-0.25
Q14	-0.07	0.07	0.14	0.02	-.39*	-0.03	0.01	0.11	0.24	-0.16
Q15	-0.18	-0.13	.46*	-0.03	-0.04	-0.14	-0.14	.43*	0.01	-0.03
Q16	-0.28	-0.08	-0.12	-0.13	0.01	-0.20	-0.04	-0.13	-.53**	-0.08
Q17	0.08	0.16	0.10	0.07	-0.07	-0.25	-0.26	0.03	-0.00	.54**
Q18	-0.01	-0.01	-0.10	0.02	0.14	-0.11	-0.06	-0.10	.39*	-0.10
Q19	0.12	0.05	0.08	0.19	-0.21	-0.12	-0.08	0.16	0.24	0.07
Q20	-0.18	-0.10	-0.03	-0.00	0.20	-0.14	-0.20	-0.01	.39*	-0.07
Q21	-0.18	-0.18	-0.01	0.02	0.07	-0.07	-0.05	-0.03	-0.02	-0.13

Chapter 5: Phase II, Experiment 2

Note. These tables above present Pearson's correlation coefficients for posttest questionnaire items (Q1 to Q21) with various physiological and psychological measures at baseline, before the speech (Pre), and during the speech. Q1 represents 'During the speech, I was confident', and Q21 represents 'During the speech, I desired to speak slowly several times' et cetera. Other questionnaire items assess participants' attitudes and experiences related to public speaking in both conditions SF and CF, as well as their reactions during the speech. Correlation coefficients marked with an asterisk (*) are statistically significant at * $p < .05$, ** $p < .01$.

5.7.4 Main Results Hypothesis: Mixed MANOVA

Two separate mixed MANOVAs were conducted to determine the effect of the real-time SDM integrated into VR scenarios on the speaker's stress levels. The stress levels were measured using both physiological parameters (SBP, DBP, HR) and self-reported questionnaire responses (stress and arousal). The two independent variables (IVs) were the VR modes (between-subjects factor): (i) SF or (ii) CF; and timing (baseline, speech), a within-subject factor.

Due to unequal sample sizes among groups of between-subject factors, Pillai's trace is used to interpret the results. Data are expressed as (before experiment mean $\pm$ standard deviation, during speech experiment mean $\pm$ standard deviation) for SF and CF respectively.

5.7.4.1 Physiological parameters

The first MANOVA examined the effects of VR modes and timing on physiological readings (SBP, DBP and HR). The assumptions of homogeneity were tested using Box's M test, which was significant ($p = .001$), indicating that the variance-covariance matrices are not equal across groups. The potential violation of assumption was addressed by using Pillai's trace, a more robust multivariate test statistic under such conditions, to interpret the MANOVA results. The homogeneity of variances was verified with Levene's Test of Homogeneity of Variance, which was nonsignificant ($p > .05$), confirming that the variances were equal across groups. The sphericity assumption was assumed for SBP, DBP and HR.

The analysis revealed that the difference between the SDM-Integrated VR modes (SF and CF) on the scores across all DVs (Baseline SBP, Baseline DBP, Baseline HR, Speech DBP, Speech SBP, Speech HR) was not statistically significant, $F(3,57) = 1.06$, $p = .37$; Pillai's trace = .057; partial $\eta^2 = .057$.

However, there was a statistically significant difference in the effect of the within-subject factor (Baseline vs Speech) on the physiological readings, $F(3,57) = 6.78, p < .001$; Pillai's trace = .278; partial $\eta^2 = .278$. The significant multivariate effect indicates that the timing of the measurements (before and during the speech task) influenced physiological readings. A closer examination of the univariate tests table indicated that this effect was primarily driven by significant differences in SBP, ($p = .027$) and DBP ($p = .000$), while HR did not show a significant difference ($p = .774$). Participant experienced an increase in SBP from baseline during the speech task and rise in DBP from baseline to during the speech task, however, the HR was not significantly changed from baseline to during the speech task as summarised in Table 5.7.

Table 5.7 *Systolic Blood Pressure, Diastolic Blood Pressure and Heart rate Results*

DV's	Independent Variables conditions			
	Supportive Feedback (SF)		Competitive Feedback (CF)	
	Mean	SD	Mean	SD
Baseline _ SBP	111.41	12.79	118.27	16.54
Speech _ SBP	124.53	14.55	128.41	18.77
Baseline _ DBP	70.74	9.93	72.63	12.26
Speech _ DBP	84.13	8.89	84.96	11.95
Baseline _ HR	80.26	10.15	80.42	11.85
Speech _ HR	87.89	10.93	91.15	15.85

5.7.4.2 Self-reported questionnaire responses

The second MANOVA examined the effects of VR modes and timing on psychological readings (stress and arousal). The homogeneity of variance-covariance matrices is assessed by Box's M test ($p = .001$), and homogeneity of variances was assessed by Levene's Test of Homogeneity of Variance ($p > .05$) were both demonstrated. The potential violation of homogeneity of variance-covariance matrices assumption was addressed by using Pillai's

trace, a more robust multivariate test statistic under such conditions, to interpret the MANOVA results. The analysis revealed the differences between VR modes (SF and CF) on psychological variables (stress and arousal) were not statistically significant, $F(3,57) = .669$, $p = .516$; Pillai's trace = .023; partial $\eta^2 = .023$. The significant multivariate effect indicates that the timing of the measurements (before and during the speech task) influenced the psychological variables, $F(2,57) = 25.35$, $p < .001$; Pillai's trace = .475; partial $\eta^2 = .475$. The investigation of univariate table indicated a significant difference in stress ($p < .001$) and arousal ($p < .001$). Participants, regardless of the VR modes, experienced an increase in stress as evidenced by the rise in stress from baseline to during the speech task, however, the arousal did not increase from baseline to during the speech task as summarised in Table 5.8.

Table 5.8 *Stress and Arousal Results*

DV's	Independent Variables (IV) conditions			
	Supportive Feedback (SF)		Competitive Feedback (CF)	
	Mean	SD	Mean	SD
Baseline _ Stress	.86	.56	.76	.47
Speech _ Stress	1.35	.67	1.56	.67
Baseline _ Arousal	1.62	.59	1.57	.49
Speech _ Arousal	2.03	.42	1.83	.54

5.8 Discussion

The current experiment examined the stress-buffering impact of two different SDM-integrated VR game modes, SF and CF, on speakers' stress levels as measured by both physiological parameters (SBP, DBP, HR and self-reported stress and arousal). The hypothesis to be tested in this experiment was that 'When increased stress is detected in the modes, social supportive mode will reduce the stress in comparison to competitive mode'. The hypothesis predicted that when increased stress is detected in the speakers, the SF mode would reduce the speaker's stress more effectively than the CF mode, due to the positive support provided by the

audience in real-time. The results revealed several key insights into how real-time feedback providing VR environments can influence stress and arousal during a public speaking task.

The mixed MANOVA results showed that there was no significant difference between the two VR game modes (SF and CF) for any of the DVs (SBP, DBP, HR, stress, arousal) when compared directly. This indicates that the type of real-time feedback provided by the SDM-integrated VR modes did not significantly alter the physiology and psychology outcomes of the individuals. However, a further analysis revealed significant effects in the measurements of the SBP, DBP, HR, stress and arousal, in respect to timings of recording and measurement of these DVs (Baseline vs during the speech).

Specifically, the within-subject (with respect to timings: Baseline and during the speech) analysis indicated significant effect and changes on participants' physiological measures including increase in SBP, DBP readings from baseline to during the speech task in both VR modes. These findings were consistent with the previous research (Bodie, 2010; Kothgassner et al., 2016). Interestingly, it is worth mentioning that stress levels were heightened during the speech task despite no significant difference in SBP, DBP, HR between the two game modes. This suggests that while both environments elicited a stress response, the social supportive feedback in the SF mode may have played a role in mitigating the physiological stress responses, even if not to a statistically significant degree.

Similarly, the self-reported questionnaire responses regarding stress and arousal did not show a significant difference between the two game modes during the baseline and speech tasks. This further suggest that lack of significant difference supports the physiological findings, indicating that while the supportive feedback in the SF mode may have impacted physiological stress markers, it did not lead to a marked difference in the subjective experience of stress

While the MANOVA results indicated no statistically significant differences between the two conditions, this does not necessarily mean that participants were equally stressed or equally not stressed. The lack of significance suggests that the variation in stress and arousal levels across the SF and CF conditions was not large enough to reach statistical thresholds. Despite these not significant results, it is clear that both SDM-integrated VR modes can impact participants' affective states during the speech. Regardless of the specific VR mode, the real-time feedback from the virtual audience, combined with the scoring system, resulted in an overall increase in stress levels, particularly in the CF mode, as evidenced by elevated systolic and DBP and HR.

Additionally, slight trends can be observed, such as higher speech-related stress and arousal levels in the CF condition compared to the SF condition. While the SF mode did not significantly reduce stress levels, it is noteworthy that participants reported higher post-speech arousal and a subjective sense of lower stress in this mode, suggesting that the social support provided by the SF mode may have fostered a more positive and motivated mindset in individuals, even if the physiological measures did not reflect a significant reduction in stress. These findings may imply that both conditions elicited stress, but the experimental manipulations did not create sufficiently distinct stress responses between the two groups. This could be due to individual differences in stress perception, the relatively short duration of the intervention, or the natural variability in how participants respond to feedback.

An interesting observation of post_arousal levels during the speech in the SF mode is the significant strong positive correlation between posttest Q2: 'I am interested in improving my public speaking skills' and Post_Arousal levels which had significantly different ratings in both the CF and SF modes.

In the context of this particular observation, the higher arousal in the SF mode may indicate that participants felt more motivated and energised by the positive feedback, leading

to a heightened state of alertness that was not necessarily linked to negative stress. This could suggest that the SF mode due to its flexible and positive environment, while reducing stress, also fosters a more engaged and dynamic environment, which could be beneficial for improving and training of public speaking. This specific finding is consistent with previous research that suggests high arousal levels combined with low stress are typically associated with a state of heightened alertness and energy, without the negative emotional strain that often accompanies stress. This state of arousal can be linked to positive emotions like excitement, enthusiasm, and motivation to perform tasks (Sapolsky, 2004, 2007; Yerkes et al., 1908).

Another key observation from the mixed MANOVA analysis is the greater spread in the data within the CF mode. The greater spread of data is reflected in the larger standard deviation values for the SBP, DBP, HR and arousal, suggesting a potentially destabilising effect of mixed feedback (positive and negative in short period of time), contributing to the higher variability in stress responses observed and ultimately leading to less consistency, cohesion and greater fluctuations in participants responses and resulting in nonsignificant findings.

These correlations suggest that participants' individual characteristics and prior experiences played a role in how they responded to the different VR modes. While the SF mode tended to foster a more supportive environment that was linked to higher arousal and lower stress, the CF mode's more variable feedback may have led to greater fluctuations in stress, particularly for those less prepared for the competitive nature of the task.

Another alternative approach would be to conduct a long-term study with regular training sessions using both VR modes with larger sample sizes, assessing the participants' performance over time, as previous studies have shown that repeated use of VR can reduce public speaking anxiety and improve communication skills (Takac et al., 2019). Such findings align with the concept of social support as a crucial factor in stress reduction, consistent with

previous research demonstrating that virtual social support can buffer stress responses (Kothgassner et al., 2019).

5.9 Theoretical and Practical Implications

The present experimental results have important implications for the design of the VR environments aimed at improving public speaking skills. The results suggest that integrating SF mechanisms into VR scenarios can effectively reduce stress and create a more positive and motivating environment for users. The SF mode's ability to lower stress while simultaneously increasing arousal highlights its potential as a powerful tool for public speaking training. However, the current study did not find significant differences between the two VR modes in their efficacy in reducing the stress levels. The lack of significant differences between the VR modes in the direct comparisons of physiological and psychological measures suggests that the effects of these environments might be more nuanced and dependent on individual differences in stress responses and is related to the type of feedback provided.

Future research should explore these individual differences more deeply, potentially examining how long-term use of VR can influence responses to supportive versus CF for lowering stress and improving public speaking skills. Additionally, recruiting a larger sample size can provide robust comparative results, even though the present study included a decent sample size as calculated with G*power. Moreover, perhaps in future it would be beneficial to include a control group. This control group could be exposed to a nonstressful speaking task, such as reading from a prepared script or describing a scene in front of them, to compare against the stress-inducing and stress-buffering VR modes.

This study provided new insights into our understanding of integrating affective state detection models to incorporate different feedback styles into VR environments for influencing stress and arousal management during public speaking tasks. Despite both the SF and CF

modes leading to nonsignificant increases in physiological and psychological stress markers during the speech task, the SF mode appeared to offer a more encouraging and motivating environment to individuals as evident from the correlations, reducing both physiological and self-reported stress. The higher arousal observed in the SF mode suggests that this feedback type may also enhance engagement and motivation, offering a promising approach for public speaking training. Future research should continue to refine these findings with larger sample sizes, exploring the complex interplay between stress, arousal and feedback in virtual environments.

5.10 Summary

The goal of this chapter was to discuss experiment 2 in detail. This chapter has provided in-depth information about development of the SDM and its integration into VR modes. This was followed by discussion of the procedure, method of study such as participant, data collection and ethical concerns were also outlined in this chapter for this specific experiment 2. The results were discussed in terms of assumption testing, and hypothesis testing.

Chapter 6: Discussion and Conclusion

This research study aimed to develop a machine learning, voice-based stress detection model specifically for public speaking scenarios, for its integration in Virtual Reality scenarios for improving public speaking skills. During the development procedure, I addressed several key research questions which were identified from the literature search and based on the existing gaps. These three research questions were then converted into hypotheses and were systematically explored through development processes and respective experiments.

Research Question 1 (RQ₁): Can the different behaviours demonstrated by a virtual audience in both supportive and nonsupportive scripted VR scenarios, as provided in our specifically developed simulation, impact a speaker's stress level irrespective of their performance, potentially leading to stress induction in the users?

Research Question 2 (RQ₂): What set of voice features (e.g., prosodic, spectral), correlated with physiological parameters (e.g., blood pressure, heart rate), can be reliably analysed in real-time from recorded voice to detect a speaker's stress level and create a voice-based stress detection model?

Research Question 3 (RQ₃): How can the stress detection model, based on the identified voice features of stress, be effectively used to provide real-time feedback? Additionally, can we test the stress-buffering effects of a virtual audience based on the detected stress levels in speakers?

6.1 Key findings

The current research resulted in significant findings across the two experiments, alongside with the development of a stress detection model that was integrated into a virtual reality application prototype.

6.1.1 Phase I Findings

Experiment 1 (E1) investigated the applicability of VR technology in inducing stress among speakers, essentially to answer RQ₁. The experiment tested the hypothesis that two different types of VR scenarios involving a variety of behaviours exhibited by a virtual audience, could change the affective state of the individuals by inducing varying levels of stress. The hypothesis was assessed using a combination of self-reported stress arousal checklists, task relevant questionnaires, and physiological measures, including SBP, DBP and HR of the participants. E1 was dedicated to testing the stress arousal capability of the VR scenarios – neutral-positive (NP) and neutral-negative (NN), as a means to collect stress voice data for further research and development of the affective state detection model.

Participants in the NN condition, with the virtual audiences exhibiting the negative scripted behaviours, experienced a higher increase in the SBP compared to those in the NP condition, where the virtual audience displayed the positive scripted behaviour. The difference was significant between the two conditions. DBP also increased more in the NN condition in comparison to the NP condition, however, this difference was less pronounced. HR followed a similar pattern to SBP, and was significantly higher in the NN condition, in contrast to the NP condition. Regarding the psychological responses to self-reported questionnaires, negative behaviours from the virtual audience in the NN condition, such as booing or constant chattering, led to increased perceived arousal, and perceived stress in comparison to the NP condition, findings that align with the physiological responses observed.

E1 therefore highlights the impacts of audience behaviour on participants' perceived stress and physiological responses, confirming the effectiveness of the VR scenarios in impacting, modulating and inducing the affective states such as stress. Additionally, I was able to collect voice samples of participants giving speeches under conditions with varying stress levels, which provided valuable stress voice samples.

The development of the VR prototypes with different virtual audience behaviours was integral to establishing these stress-inducing conditions and capabilities, which were essential and crucial for collecting stress-related voice data. This voice data was subsequently utilised in Phase II to identify stress indicative voice features, develop a ML model and ultimately integrate this model into the VR application to help reduce and manage stress in public speaking scenarios, thereby aiding effective communication development, leading to several other significant findings.

6.1.2 Phase II Findings

The development of the SDM was a critical and highly significant component of this research work, aimed at identifying the most effective voice features for detecting stress, as outlined in RQ₂. This process began with a preliminary analysis of the publicly available RAVDESS database and was subsequently extended to the VRPS database, which contained voice samples collected during E1. The insights gained from this analysis were crucial and set the stage for subsequent refinement using the VRPS database, which contained real-world voice samples collected during E1.

To address RQ₂, from the RAVDESS analysis of voice features, preliminary findings revealed that Mel-frequency cepstral coefficients (MFCCs) were the most effective features (showing a 23 per cent significant difference) for distinguishing between confident, calm vocal expressions and those marked by fear or stress. Additional features such as root mean square (18 per cent significant difference) and amplitude envelope (16 per cent significant difference), also proved useful, enhancing the model's ability to capture the nuances of stress vocal expressions. These features were initially employed to develop a stress prediction model, using various algorithmic models. Among these, K-Nearest Neighbours, Random Forests both achieved a notable accuracy of 82 per cent while Support Vector Machines showed a lower accuracy of 57 per cent. These preliminary voice features, and algorithmic results, underscored

the potential of MFCCs and other selected features in forming the basis of a stress prediction model.

The findings from the VRPS database, which featured voice samples from E1, the significance of MFCCs (14 per cent accuracy) was further validated in differentiating between stressed and nonstressed voice samples. The VRPS database, characterised by diversity and real-world data samples, reinforced that MFCCs were consistent and their ability to capture the nuanced spectral energy changes that indicate stress in the voice, makes them reliable indicators of vocal quality despite variations in emotional content and differences between professional actors and university students.

For the final model, 32 MFCCs were extracted and selected, excluding the first eight MFCCs to include and focus on delta coefficients to capture dynamic speech patterns, and these were utilised in developing a feedforward neural network model. This SDM (achieving 69.44 per cent accuracy with testing data) was designed for real-time stress detection and for its integration in the VR prototype, enabling the monitoring of speakers' stress levels and providing real-time feedback through a visual score and the virtual audience in public speaking scenarios.

The integration of the SDM into the VR environment was a significant advancement, enabling real-time monitoring of speakers' stress levels and optimising the system for improving public speaking performance. The integration was crucial for addressing RQ₃, which aimed to explore how SDM-integrated VR scenarios could be effectively utilised to enhance public speaking skills.

The effectiveness of the SDM-integrated VR application for improving public speaking was tested using two distinct game modes: SF and CF, to answer RQ₃. In the SF mode, the virtual audience provides consistent positive reinforcement, designed to encourage and motivate speakers regardless of their performance. CF mode combines both positive and

critical feedback, challenging speakers to adapt and manage stress in a more dynamic and competitive environment. The effectiveness of these modes in improving public speaking performance through real-time feedback and stress reduction was evaluated in Experiment 2 (E2). E2 specifically tested the hypothesis that the SF mode would be more effective at reducing stress than the CF mode.

The results of Experiment 2 indicated that participants in both the SF and CF conditions experienced an increase in stress from baseline to during the speech, however, there was no significant difference in the stress-buffering effects between the two modes. This was evident across all physiological measures, including systolic and DBP, HR and self-reported stress levels, measured at the baseline and during speech task readings. Both modes showed consistent increases in these indicators, reflecting heightened stress during the speech task.

Interestingly, while stress levels increased across both modes during speech tasks in comparison to baseline, arousal levels were unexpectedly higher in the SF mode compared with the CF mode. This finding suggests that the relationship between stress reduction and the type of feedback provided by the VR environment is more complex than initially anticipated. These results highlight the need for further research, potentially with a larger sample size, to better understand how different types of real-time feedback influence stress and arousal in public speaking scenarios.

6.2 Theoretical Implications

The current study advances research in stress recognition and detection, specifically within the context of public speaking where it is a common barrier to effective communication. By focusing on the analysis of stress within the framework of affective computing focusing on voice analysis, this research contributes to a deeper theoretical understanding of how stress impacts voice features. The development and integration of voice analysis-based ML models

for stress detection and feedback within VR environments represents a key contribution to the field of affective computing and immersive technologies dedicated to public speaking. The current research study also introduces significant experimental work and sheds light on different VR scenarios featuring varying virtual audience behaviours, such as supportive and critical feedback, in stress induction, stress recognition and stress reduction. Furthermore, by utilising VR technology integrated with a foundational real-time SDM, these innovations deepen understanding of stress-related voice features during public speaking in immersive settings. They also offer important theoretical implications for the integration of affective computing within immersive technologies, particularly when combined with social support and the repeated use of VR for training. This approach paves the way for the development of effective and interactive solutions for stress management for detection of stress via voice analysis and improving public speaking.

6.2.1 Technical Implications: Voice Analysis Based Stress Detection

In the context of public speaking, stress is a subjective response that triggers various psycho-physiological reactions (Dowd et al., 2010; Droppleman & McNair, 1971; Goyal et al., 2016), ultimately impacting a speaker's voice (Hansen & Patil, 2007; Pisanski & Sorokowski, 2021). Stress leads to cardiorespiratory changes, such as increased HR and irregular breathing, which alter vocal characteristics like pitch, tone, and energy (Pisanski et al., 2018). While significant research has focused on detecting stress through physiological parameters (Sano & Picard, 2000; Zhai et al., 2008) and self-reported questionnaire responses (Feldman et al., 2004; Kemeny, 2003; Westenberg et al., 2009), real-time voice-based stress detection in the context of public speaking has been given little attention and underexplored.

This study addressed this gap by identifying vocal stress indicators and developing a ML model for real-time stress detection in VR public speaking environments. Recent research suggests the role of voice as a key marker of stress (Van Puyvelde et al., 2018), with changes

in spectral, temporal and prosodic features, making it a valuable tool for detecting affective states during public speaking and providing insights into the speaker's state (Bhavan et al., 2019; Juslin & Laukka, 2003; Tolkmitt & Scherer, 1986).

Stress indicative voice features such as fundamental frequency across different studies remain inconsistent, especially in context of public speaking (Giddens et al., 2013). The subjective nature of stress (Koroleva et al., 2014) and the variability in how each individual responds further complicates its detection in real-time. Traditional speech recognition models are often unable to capture the dynamic vocal changes stress induces, leading to the need for specialised stress detection models tailored to public speaking.

A significant challenge has been the lack of context-specific voice databases for public speaking that can lead to further voice analysis research. Widely used databases like the Berlin Database of Emotional Speech (EmoDB), the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC), or the Speech Under Simulated and Actual Stress (SUSAS) focus on general emotional states or high-stress environments like military tasks, or pilots stress speech voice samples, but are not suited for public speaking contexts. To contribute to this gap, this research has developed a unique voice database called the Virtual Reality Public Speaking (VRPS) database, which contains voice data from 60 participants who gave speeches in stressful and nonstressful VR-based public speaking scenarios. These scenarios were designed to simulate real-world public speaking conditions, offering highly context-specific and accurate voice data.

By using the VRPS database, this study also contributed to another significant gap in current research, which lies in the identification of stress-related voice features specifically for public speaking scenarios. While a lot of research has been conducted on identifying relevant indicative voice features predominantly focused on detecting a range of other emotions, such as happiness, sadness, fear, disgust, anger and surprise in different contexts (Nogueiras et al.,

2001; Nwe et al., 2003b; Yi-Lin et al., 2005), minimal research has been dedicated to detect nonverbal voice features indicative of stress in the context of public speaking (Han et al., 2014; Pfister & Robinson, 2011).

Using the VRPS database, this study identified MFCCs as the most consistent voice features for recognising, distinguishing and detecting stress in public speaking scenarios. This finding fills a gap in the literature, where real-time voice-based stress indicative features remain underexplored in the context of public speaking. Previous research by Hansen and Patil, (2007) has shown that stress changes energy spectra and formant frequencies, and this study builds on that by demonstrating the relevance of MFCCs in capturing those changes. Specifically, MFCCs provide insights into shifts in the energy distribution, which reflect stress-related vocal changes (Hansen & Patil, 2007). The dispersal of resonances in formants, linked to high arousal (Echternach et al., 2016), aligns stress responses closer to emotional states like fear. Research by Hansen & Patil (2007) highlighted that speech under stress exhibit similarities to those found in emotional states like anger and fear. This is similar to what was found in the preliminary work carried out using the RAVDESS database voice samples between fearful (stressed) and calm (confident) voice samples (Arushi et al., 2021).

MFCCs, computed via a Mel filter bank, mimic the human auditory perception of sound by analysing the power spectral envelope. MFCCs are well-suited to detecting these subtle changes in the acoustic properties of speech, such as reduced airflow and changes in vocal tension, which often accompany stress (Godino-Llorente et al., 2006). In this case, 32 MFCCs were selected out of 40 possible features to enhance the model's ability to detect stress. Specifically, MFCCs 13-25 were chosen for their sensitivity to finer spectral details and timbre, while MFCCs 26-39 captured dynamic changes in the spectrum over time (Rabiner & Juang, 1993). This combination was critical because it allowed the model to detect subtle shifts in spectral energy that occur under stress, significantly improving its effectiveness in real-world

applications, especially for public speaking stress detection. This makes MFCCs valuable for identifying stress-related variations in voice.

The selected stress-indicative MFCCs were implemented into a Feedforward Neural Network (FFNN), which was chosen for its ability to capture complex, nonlinear relationships in the data. FFNNs outperform models like SVMs, KNN or Random Forests, which may struggle with high dimensional and nonlinear data such as the MFCC patterns. The context-specific developed Stress Detection Model (SDM) (with 67 per cent accuracy with testing data) using FFNNs allowed for fast inference, flexibility in architecture and quick processing, making it ideal for real-time applications like stress detection in live public speaking scenarios (Bishop, 2006; Haykin, 1998).

The SDM could become more robust by incorporating multiple modalities beyond just MFCCs. Adding features such as physiological signals (e.g., HR, skin conductance) or visual cues (e.g., facial expressions) could enhance its ability to detect stress more accurately. Expanding to other voice features like prosodic elements (e.g., pitch, duration) and using hybrid models or ensemble models approach that combine MFCCs with additional features could improve generalisation across various stress-inducing scenarios. The development of the SDM with MFCCs and FFNNs has improved the model's accuracy and also facilitated its deployment in practical settings where real-time feedback and response is needed every few seconds in the virtual reality public speaking application. The current research highlights the effectiveness in the combination of spectral sensitivity provided by the MFCCs with the robust processing capability of neural networks, making SDM a powerful tool for real-time stress detection in public speaking scenarios. This approach ensures that the model can be easily deployed and integrated into various platforms, providing real-time feedback and stress management solutions.

The final contribution of this research, compared with traditional public speaking classes where feedback is limited, is that the SDM, embedded in VR public speaking scenarios, provides real-time behavioural feedback throughout the training. This continuous feedback, combined with a serious game element that displays scores based on detected stress levels, creates a sustainable and engaging training environment. The VR application allows speakers to practice managing stress in public speaking, accurately reflecting real-world stressors while offering a safe and immersive platform for consistent improvement in confidence and communication skills.

This research contributes to the field of voice analysis-based affective computing by advancing real-time stress detection technology. It bridges a key gap in existing research by offering a VRPS database, recognising stress indicative voice features, and developing a model specifically designed for recognising public speaking stress. The SDM's integration with immersive VR technology offers both speed and accuracy in detecting subtle, stress-induced vocal changes in voice, and providing a platform to individuals for improving public speaking.

6.2.2 Experiment 1 Implications: VR for Public Speaking Stress Induction

Phase I, consisting of Experiment 1 (E1), examined how VR can effectively induce stress in public speaking scenarios using supportive and nonsupportive virtual audiences. The outcomes confirmed the stress-induction capabilities of VR, leading to the creation of the VRPS database, which captures stress-related voice data. This foundational experiment highlighted how VR can simulate stress-inducing environments and laid the groundwork for future research in public speaking stress detection models.

Traditionally, VR technology has been employed to study, induce anxiety (Carl et al., 2019; Gerardi et al., 2010; Mostajeran et al., 2020) and fear of public speaking (Slater et al., 2006; Slater et al., 1999), often focusing on the interactions and behaviours of virtual audiences. However, E1 in this research distinguishes itself by focusing on stress associated

with public speaking within VR (El-Yamri et al., 2019; Kotlyar et al., 2008) contributing significantly to the literature. E1's findings provide empirical evidence that different behaviours exhibited by virtual audiences – negative or unsupportive – can significantly heighten stress, arousal, physiological parameters (HR, SBP), while positive supportive behaviours offer a stress-buffering effect.

These results align with and extend cognitive appraisal theory (Roseman et al., 2001), which emphasises that emotional responses are shaped by how individuals interpret social cues. Previous studies in VR have similarly indicated that social support is crucial in managing stress in high-stress scenarios (Kothgassner et al., 2019). In this context, E1 demonstrates that VR environments can effectively engage users emotionally and impact their affective states. E1 confirms that VR not only functions as a stress induction tool but also as an active interactive medium that engages users emotionally, further validating the immersive capacity of VR in shaping affective states (Westenberg et al., 2009; Wiemers et al., 2015).

The insights gained from E1 deepen the understanding of how virtual environments can be leveraged both for stress induction and for stress management through social support mechanisms. For instance, positive encouragement could serve as an effective tool in VR-based public speaking training to help individuals manage their stress and build confidence.

E1's implications are especially significant for public speaking training. The findings show that VR can simulate realistic public speaking scenarios while allowing controlled manipulation of stress levels. This capability is invaluable in helping individuals develop effective coping mechanisms and refine their ability to manage stress in real-world public speaking situations. Beyond public speaking, these results also open avenues for research and application in adaptive stress management training for leadership development, interview preparation, public speaking and presentations in both educational and professional settings.

6.2.3 Experiment 2: *SDM Integrated VR for Stress Reduction*

Based on the findings from E1, which demonstrated the effectiveness of VR in inducing varying levels of stress in controlled public speaking environments and led to the creation of the comprehensive VRPS database of stress-related voice samples, the SDM was developed using voice features derived from those samples. Experiment 2 (E2) took a step further by integrating the SDM into the VR system. This experiment explored how personalised feedback, specifically the SF mode, could potentially reduce stress by providing real-time feedback based on the user's detected stress levels from voice. The aim was to test whether SF could buffer stress and enhance public speaking performance.

While E2 demonstrated that the SDM effectively provided real-time feedback and behavioural adjustments in the virtual audience based on stress detection, the expected reduction in stress through the SF mode in comparison to CF mode was not conclusively proven as the differences were not significant. Higher arousal levels ($SD = 0.42$) combined with lower stress levels ($SD = 0.67$) and the notably high variability in stress within the SF mode suggested that, in some cases, social support may increase engagement or motivation rather than only alleviating stress. This combination of heightened arousal and moderate stress levels indicates that participants in both the NP of E1 and SF of E2 conditions were likely experiencing a challenge response. In addition to this, in the SF condition both DBP and HR increased from baseline to speech. Although the increases were slightly lower compared with the CF mode, they still suggest a physiological response linked to effort and engagement. This aligns with previous research indicating that physiological measures like HR can reflect effort or motivation in addition to stress (Mikus et al., 2009). Rather than perceiving the task as threatening, they may have viewed it as an opportunity to improve performance, driven by the SF. The motivated engagement suggests that supportive environments in VR may manage stress and also enhance focus and readiness to perform, making them a valuable tool for public

speaking training. Additionally, social support can still play a motivational role, encouraging speakers to improve their communication skills. Further experiments with larger sample sizes are needed to clarify the stress-buffering effects of the SS mode and to confirm these results.

These findings emphasise the complexity of stress responses in VR settings. The variability in individual stress responses, even with SF, highlights the need for further research to better understand the interplay between different types of feedback and individual arousal levels. The partially supported hypothesis calls for future experiments with larger sample sizes to explore the stress-buffering effects of the SF mode and refine the interaction between supportive feedback and real-time stress detection.

Despite the lack of full support for the hypothesis, E2 makes an important contribution by showcasing the successful real-time integration of affective computing technologies within VR environments. The SDM's ability to provide immediate, stress-sensitive scoring and behavioural adjustments of the virtual audience marks a significant advance in public speaking training. This integration of voice analysis with immersive VR environments represents a novel approach to enhancing stress management and communication skills.

The methods used in this experiment hold great potential for further development in high-stress fields beyond public speaking, where adaptive VR training could provide personalised, context-specific interventions to help individuals cope with stress and improve performance.

6.2.4 E1 and E2 Replicability: Implications for Public Speaking Training

This study explored the integration of a voice-based ML SDM in VR for public speaking training scenarios. Both Experiment 1 (E1) and Experiment 2 (E2) helped develop a comprehensive understanding of how VR scenarios can be used to induce, detect, and manage stress in real time, providing technical contributions and insights into human behaviour in stressful situations.

In E1, the focus was on inducing stress through virtual audience behaviours: Neutral-Positive (NP) and Neutral-Negative (NN), to generate stress-related voice data and build the VRPS database. This experiment provides a replicable framework for stress induction using controlled virtual environments, adaptable to other high-stress tasks such as job interviews, performance assessments or educational settings. The success of E1 in inducing measurable stress responses, confirmed by physiological and self-reported data, offers a high degree of replicability, which can be applied to future studies in similar performance-based tasks.

One of the key technical contributions from E1 was the creation of the VRPS database, and consistent identification of MFCCs as the primary voice features for stress detection. These features were central to the development of the SDM, demonstrating their reliability across different datasets, including in real-world stress contexts like public speaking. The SDM was designed based on these voice features, making it adaptable for real-time applications in public speaking and other high-stress tasks.

E2 built on E1's findings by integrating the SDM into the VR system, allowing real-time feedback based on participants' stress levels during public speaking. This experiment tested the effects of SF and CF on public speaking performance and stress management. The real-time integration of the SDM, which utilised MFCCs and feedforward neural networks (FFNNs), provides a replicable and adaptable process for immersive environments and stress-inducing tasks.

The findings of E1 and E2 show inconsistencies in stress and arousal responses, highlighting the complexity of stress induction and reduction in different VR environments. In E1, clear and consistent stress responses were observed when participants were exposed to NP and NN audience behaviours, demonstrating the replicability of stress induction through controlled VR environments. On the other hand, E2 revealed greater variability in stress and arousal levels between SF and CF modes. The variability in E2 responses was influenced by

the interactive and mixed nature of real-time feedback in CF mode and motivating effect of effects of supportive behaviour in SF mode results. E1 focused on inducing stress through predefined audience behaviours at certain predefined sets of timestamps (for example 60 seconds, at 120 seconds, and 180 seconds) and E2 included dynamic, real-time feedback that elicit continuously changing and individualised responses. The interplay between continuous visual scores shown to speakers, continuous behavioural feedback provided in form of SF and mixed critical and SF in E2 introduces a layer of complexity that does not present in E1. The unexpected inconsistent results suggest that predefined and automated real-time behaviour of a virtual audience can induce varying responses and impact on the users, further indicating that the type of virtual audience's feedback, UI design mechanisms can have impact on stress levels and arousal levels. However, the variability in stress and arousal responses between the SF and CF modes, compared with NP and NN in E1, highlights the complexity of stress responses in social contexts. The replicability of stress responses may depend on the specific design and context of the VR scenarios, suggesting that more research is needed to fully understand the interplay between real-time feedback mechanisms, and affective state detection systems integration in VR, UI design for best user experience, and their impact on arousal, and stress.

Future studies should aim to replicate these findings with larger and more diverse datasets to further validate the results of E1 and explore the stress-buffering effects observed in E2. Longitudinal studies could provide more insight into the effectiveness of the SDM-integrated VR application, examining its generalisability across different public-speaking scenarios and user populations.

6.3 Practical Implications

The findings from this study on voice analysis-based stress detection and management in public speaking scenarios using VR offer valuable insights with significant practical

implications across various fields. The results from Experiment 1 (E1) suggest that VR can be a powerful tool for inducing and managing stress in controlled settings. This opens new avenues for VR-based therapy and stress management programs.

VR has already been widely adopted in mental health interventions, particularly in exposure therapy for conditions like social anxiety and public speaking fears (Kothgassner et al., 2019). Existing studies have shown that VR can provide a controlled yet immersive environment where individuals gradually confront anxiety-provoking situations. For example, research by Takac et al. (2019) demonstrated the effectiveness of VR in public speaking training, with participants reporting significant reductions in anxiety and improved coping mechanisms through repeated VR exposure (Takac et al., 2019). Similarly, Hartanto et al. (2014) found that VR environments can help patients manage social stress by providing a safe space for repeated exposure to feared situations, such as speaking before an audience (Hartanto et al., 2014).

This study further contributes to these findings by integrating real-time stress detection via the SDM, offering mental health professionals a flexible tool to tailor VR experiences to individual needs. With real-time adjustments in audience behaviour, therapists can provide exposure therapy for social anxiety or public speaking fears, ensuring the experience matches the client's stress levels. Moreover, the SDM's real-time feedback can be adapted for relaxation and mindfulness training, helping users recognise and mitigate stress as it occurs.

Findings from Experiment 2 (E2) provide important considerations for designing stress-responsive systems. The counterintuitive result that SF can both reduce stress and increase arousal more effectively than CF indicates that feedback systems in VR and other interactive platforms must consider the nuanced ways in which users respond to different types of support. Practitioners and designers should be mindful when implementing SF mechanisms, ensuring they are appropriately matched to the context and user needs. This insight is particularly

valuable for developers of educational technologies, therapeutic tools and other applications where user feedback is crucial to the experience.

VR has proven effective in enhancing public speaking skills but its current applications lack integrated stress detection mechanisms, limiting its ability to address stress management. Recent studies (Premkumar et al., 2021; Sülter et al., 2022) continues to further explore anxiety and fear during public speaking. The present study uniquely contributes to understanding of underlying affective state of stress during public speaking, particularly through the development of a voice analysis-based stress detection system and feedback mechanism within a VR environment. These advances reinforce the significance of integrating affective computing with immersive technologies to address public speaking stress. One of the most impactful practical implications of this research is the use of immersive technologies to improve public speaking training. The integration of the SDM into a VR-based public speaking simulation offers a groundbreaking approach to enhancing this skill. This application provides users with real-time feedback on their stress levels, allowing them to become more aware of their emotional states during presentations and adjust their performance accordingly. The ability to simulate a range of audience behaviours both supportive and nonsupportive within the VR environment creates a controlled setting where users can practice managing stress in response to different social cues. This innovative approach offers consistent feedback, support and learning opportunities, which are often not readily available in traditional public speaking courses. By incorporating VR-based modules that simulate high-pressure scenarios, educational institutions and corporate training programs can significantly benefit. These modules help students and professionals refine their public speaking and leadership skills, building resilience and confidence for challenging situations. The availability of such a comprehensive training tool makes it easier for learners to practice and improve continuously, something that real-life public speaking courses may not always provide.

For AI and VR developers, this research provides valuable insights into the integration of affective computing approaches based on ML models within interactive virtual environments. The successful application of Mel-frequency cepstral coefficients (MFCCs) for real-time stress detection highlights the potential for incorporating advanced audio analysis techniques into VR applications. Developers can build on this foundation to create more sophisticated VR experiences that dynamically respond to users' emotional states, enhancing engagement and personalisation. This is particularly relevant for the gaming industry, where adaptive gameplay based on emotional responses could lead to more immersive and emotionally engaging experiences.

The replicability and adaptability of the SDM across different contexts suggest that this technology could be expanded beyond public speaking. Similar models could be developed for customer service training, where managing stress and maintaining composure are crucial. Additionally, this technology could be applied in fields where stress management is key to success, such as in nursing homes or hospitals, where it could help detect stress in nurses and doctors from their voices, allowing for adjustments in duties to reduce stress. The ongoing refinement and testing of these systems will be essential to ensuring their effectiveness and broadening their applicability across various industries.

6.4 Limitations and Suggestions for Future Work

Despite the significant findings and contributions of this study, several limitations must be acknowledged. The experiments conducted in this study involved a relatively small and homogenous sample, primarily consisting of university students. This limits the generalisability of the findings to broader populations with diverse demographic backgrounds. Different professional experiences may trigger different responses to stress in VR environments, and these variations were not fully explored in this study. The VR scenarios

developed for E1 were highly controlled, with scripted audience behaviours. While this was necessary for experimental consistency specifically to validate stress arousal capability, it may not fully capture the complexity and variability of real-world public speaking situations. Real-world settings often involve more dynamic and unpredictable audience interactions, which could affect stress levels differently.

The SDM primarily focused on vocal features, particularly Mel-frequency cepstral coefficients (MFCCs). While these features were effective in detecting stress, other physiological and behavioural indicators, such as facial expressions, body language and skin conductance, were not included in the model. This narrow focus may limit the comprehensiveness of the stress detection system. Traditional feature selection and model development methods were used in this study due to their suitability for working with a contextual dataset and their alignment with the research's primary aim of investigating stress-indicative voice features in the context of public speaking. The approach allowed a transparent analysis of how individual voice features varied between stressed and non-stressed speech samples, supporting interpretability in an interdisciplinary research setting. Future research may benefit from the integration of state-of-the-art deep learning architectures, such as Recurrent Neural Networks (RNNs) or Transformers, for stress detection model development in similar contexts. Comparing these deep learning methods against the current framework is recommended to evaluate potential performance improvements and to explore the advantages of end-to-end learning approaches for improving affective computing-based voice analysis.

E2 results indicated that arousal levels in the SF were unexpectedly higher than in the CF mode. This suggests that arousal and stress, while related, may require more nuanced differentiation in future studies with larger sample to distinguish types of arousal (e.g., excitement vs anxiety) experienced by the speakers when social support is provided.

In both E1 and E2, another limitation was the experimental design where participants were required to remain seated throughout the VR public speaking tasks to ensure accurate measurement of physiological signals (e.g., heart rate, systolic and diastolic blood pressure), leading to restricted body movements and gestures. However, several participants exhibited strong affective responses, including visible signs of freezing and in some few cases, removing the VR headset earlier due to stress, suggesting strong ecological validity in terms of emotional engagement. Incorporating motion-tolerant physiological sensors along with VR simulations in future studies would be helpful to assess whether the observed affective responses, behaviours and speech patterns remain ecologically valid under more naturalistic conditions.

Future research should aim to include a larger and more diverse sample to enhance the generalisability of the findings. Including participants from various demographic backgrounds, such as different age groups, cultural contexts and professional experiences, would provide a more comprehensive understanding of how stress is induced and managed in VR public speaking scenarios. To better simulate real-world public speaking experiences, future studies could incorporate more dynamic and unpredictable audience behaviours into the VR scenarios in the stress induction process. This could involve unscripted audience interactions or varying the composition and size of the virtual audience, providing a more realistic and challenging environment for participants. Future iterations of the SDM should explore the integration of additional physiological and behavioural indicators, such as facial expressions, body language and skin conductance. A multimodal approach with ensemble algorithmic models with more exclusive prosody indicative features could enhance the accuracy and reliability of stress detection by capturing a broader range of stress responses. Further research is needed to better understand the relationship between arousal and stress, particularly in distinguishing between different types of arousal. Developing more refined tools and methodologies to differentiate

between positive arousal (e.g., excitement) and negative arousal (e.g., anxiety) could provide deeper insights into how different feedback modes influence stress levels in VR environments.

Future research could benefit from conducting longitudinal studies to assess the long-term effects of repeated exposure to VR public speaking scenarios. A limitation of this study is the short duration of exposure to the VR game modes with participants receiving only a single 5-minute training session, providing valuable insights into the immediate effects of the SF and CF modes, but it is unclear whether these results translate into longer-term benefits. Future research should investigate the impact of extended and repeated exposure to these SDM integrated VR training environments. Longer and more frequent sessions could potentially yield more pronounced and enduring effects on stress management, public speaking performance and the ability to regulate stress responses. This would help determine whether the stress-reducing effects of the SF, or any other interventions, have lasting impacts on public speaking performance and stress management over time. Given the mixed results regarding the effectiveness of SF and CF modes, future studies could explore alternative feedback mechanisms or combinations of feedback types, specifically focusing on long term effects with training. Investigating how personalised feedback, based on individual stress profiles, impacts stress reduction and public speaking performance, could lead to more tailored and effective training tools incorporating eye-tracking technology to gain deeper insights into participant behaviour such as engagement and attention during VR training sessions. For instance, in the SF condition, it would be useful to examine whether participants focus more on the stress score or the supportive audience. Similarly, in the CF condition, eye-tracking data could reveal whether participants are drawn to critical feedback elements or other environmental cues. Information such as this could further elucidate how different feedback mechanisms influence participants' stress responses, motivation and overall interaction with the VR scenario. By addressing these limitations and pursuing the suggested areas for future research, the field can

continue to advance our understanding of stress induction and management in VR environments, ultimately leading to more effective and widely applicable interventions.

6.5 Conclusion

The current thesis has set out to explore and address the challenge of stress in public speaking by developing a machine learning-based voice stress detection model integrated into a VR environment. The study was driven by three core research questions that focused on the impact of virtual audience behaviours on stress levels, the identification of voice features that indicate stress, and the application of real-time feedback to manage stress during public speaking with the SDM.

The findings from this research have made significant contributions to both theoretical understanding and practical applications in the fields of affective computing, VR and public speaking training. Experiment 1 (E1) provided strong evidence that different behaviours exhibited by a virtual audience can significantly influence a speaker's stress levels, as measured by both self-reported arousal and physiological responses. Positive audience behaviours were shown to reduce stress and negative behaviours increased stress, underscoring the potential of VR as a tool for controlled stress induction in research and training contexts.

Building on previous E1 findings, Experiment 2 (E2) tested the effectiveness of the SDM in providing real-time feedback during public speaking scenarios. The model successfully identified key voice features, particularly MFCCs, that could reliably differentiate between stressed and nonstressed vocal expressions. When integrated into the VR environment, the SDM allowed for real-time stress monitoring and feedback, with the SF and CF modes offering insights into different strategies for stress management. While the SF mode was expected to reduce stress more effectively, the results revealed a more complex relationship between feedback type, arousal and stress, suggesting avenues for further exploration.

The practical implications of this work are substantial, offering new tools for public speaking training, VR-based therapy and stress management. The integration of ML and VR provides a powerful platform for simulating realistic public speaking scenarios, offering users the opportunity to practice and improve their skills in a controlled environment. The potential applications of the SDM extend beyond public speaking, with possibilities in customer service training, and other fields where stress management is crucial. However, this study also identified several limitations, including the need for more diverse participant samples, the incorporation of real-world variability in VR scenarios, and the exploration of multimodal stress detection approaches. Addressing these limitations in future research will be essential for further refining the SDM and enhancing its applicability across different contexts.

To conclude, this thesis has demonstrated the feasibility and effectiveness of using VR and ML to address the challenges of stress in public speaking. The findings contribute to the growing body of knowledge on stress induction and management in virtual environments and offer practical solutions that can be adapted and expanded in future work. As technology continues to evolve, the integration of advanced affective computing, specifically voice analysis-based stress detection and feedback systems into everyday applications holds great promise for improving individual performance and wellbeing in a variety of settings.

References

- Abe, A., Yamamoto, K., & Nakagawa, S. (2015, 2015). Robust speech recognition using DNN-HMM acoustic model combining noise-aware training with spectral subtraction.
<https://doi.org/10.21437/Interspeech.2015-599>
- Agarwalla, S., & Sarma, K. K. (2016). Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech. *Neural Networks*, 78, 97-111.
<https://doi.org/10.1016/j.neunet.2015.12.010>
- Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken Emotion Recognition Using Hierarchical Classifiers. *Computer Speech & Language*, 25(3), 25.
<https://doi.org/10.1016/j.csl.2010.10.001>
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8, 1-74.
<https://doi.org/10.1186/s40537-021-00444-8>
- American Psychological Association. (2018a). Anxiety. In *APA Dictionary*. Retrieved 7/2/2022, from <https://dictionary.apa.org/anxiety>
- American Psychological Association. (2018b). Fear. <https://dictionary.apa.org/fear>
- American Psychological Association. (2018c). Stress. In *APA Dictionary*.
<https://dictionary.apa.org/stress>
- Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features And Classifiers For Emotion Recognition From Speech: A Survey From 2000 To 2011. *Artificial Intelligence Review*, 43(2), 155-177. <https://doi.org/10.1007/s10462-012-9368-5>
- Arnold, J. K. (2018). *Stress to success: Public speaking anxiety and its relationship to perceived leadership* <https://researchonline.stthomas.edu/esploro/outputs/991015132289003691>
- Arushi, Dillon, R., & Teoh, A. N. (2021, 17-20 Aug. 2021). Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills. 2021 IEEE Conference on Games (CoG). <https://doi.org/10.1109/CoG52621.2021.9618989>

- Atkinson, J. M. (1984). Public speaking and audience responses: some techniques for inviting applause. In. <https://tinyurl.com/2u2cdwrc>
- Ayres, J. (1996). Speech preparation processes and speech apprehension. *Communication Education*, 45(3), 228-235. <https://doi.org/10.1080/03634529609379051>
- Azwar, H., Alam, N., Batool, M., & Khan, S. A. (2016). Virtual Reality Based Immersion Systems. *International Journal of Technology and Research*, 4(4), 82-84. <https://api.semanticscholar.org/CorpusID:115104212>
- Bachmann, M., Subramaniam, A., Born, J., & Weibel, D. (2023). Virtual reality public speaking training: effectiveness and user technology acceptance. *Frontiers in Virtual Reality*, 4, 1242544. <https://doi.org/10.3389/frvir.2023.1242544>
- Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., & Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78(5), 5571-5589. <https://doi.org/10.1007/s11042-017-5292-7>
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2012, 2012). 3D constrained local model for rigid and non-rigid facial tracking. <https://doi.org/10.1109/CVPR.2012.6247980>.
- Bandela, S. R., & Kumar, T. K. (2019). Speech emotion recognition using semi-NMF feature optimization. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(5), 3741-3757. <https://doi.org/10.3906/elk-1903-121>
- Banse, R., & Scherer, K. R. (1996). Acoustic Profiles In Vocal Emotion Expression. *Journal of personality and social psychology*, 70(3), 614. <https://doi.org/10.1037//0022-3514.70.3.614>
- Barreto, A., Zhai, J., & Adjouadi, M. (2007). *Non-Intrusive Physiological Monitoring For Automated Stress Detection In Human-Computer Interaction* International Workshop on Human-Computer Interaction. https://doi.org/10.1007/978-3-540-75773-3_4
- Bassett, J. R., Marshall, P. M., & Spillane, R. (1987). The physiological measurement of acute stress (public speaking) in bank employees. *International Journal of Psychophysiology*, 5(4), 265-273. [https://doi.org/10.1016/0167-8760\(87\)90058-4](https://doi.org/10.1016/0167-8760(87)90058-4)

- Bengio, Y. (2009). *Learning deep architectures for AI* (Vol. 2). Now Publishers, Inc.<https://www.nowpublishers.com/article/Details/MAL-006>
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828. <https://doi.org/10.1109/TPAMI.2013.50>
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157-166. <https://doi.org/10.1109/72.279181>
- Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1991). Autonomic determinism: the modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychological review*, 98(4), 459. <https://doi.org/10.1037/0033-295x.98.4.459>
- Besbes, S., & Lachiri, Z. (2019, 21-24 March 2019). Stressed speech recognition using MMFCC and kernel based classification. 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD). <https://doi.org/10.1109/SSD.2019.8893273>
- Bhagat, K. K., Liou, W.-K., & Chang, C.-Y. (2016). A Cost-Effective Interactive 3d Virtual Reality System Applied To Military Live Firing Training. *Virtual Reality*, 20(2), 127-140. <https://doi.org/10.1007/s10055-016-0284-x>
- Bhavan, A., Chauhan, P., Hitkul, & Shah, R. R. (2019). Bagged Support Vector Machines For Emotion Recognition From Speech. *Knowledge-Based Systems*, 184, 104886. <https://doi.org/10.1016/j.knosys.2019.104886>
- Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 2). Springer New York, NY.<https://tinyurl.com/bdykyj34>
- Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-Level Spectral Features For Emotion Recognition. *Speech communication*, 52(7-8), 613-625. <https://doi.org/10.1016/j.specom.2010.02.010>
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development*, 78(1), 246-263. <https://doi.org/10.1111/j.1467-8624.2007.00995.x>

- Bodie, G. D. (2010). A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education*, 59(1), 70-105.
<https://doi.org/10.1080/03634520903443849>
- Bouchlaghem, D., Shang, H., Whyte, J., & Ganah, A. (2005). Visualisation in architecture, engineering and construction (AEC). *Automation in Construction*, 14(3), 287-295.
<https://doi.org/10.1016/j.autcon.2004.08.012>
- Bowers, B. (2001). *Sir Charles Wheatstone FRS: 1802-1875*. IET. <https://shop.theiet.org/sir-charles-1802-1875-2nd-ed>
- Breakey, L. K. (2005). Fear of public speaking-the role of the SLP. 26, 107-117.
<https://doi.org/10.1055/s-2005-871206>
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
<https://doi.org/10.1023/A:1010933404324>
- Brooks, A. W. (2013). Get Excited: Reappraising Pre-Performance Anxiety as Excitement. *Journal of Experimental Psychology: General*, 143(1), 000. <https://doi.org/10.1037/a0035325>
- Brundage, S. B., & Hancock, A. B. (2015). Real Enough: Using Virtual Public Speaking Environments To Evoke Feelings And Behaviors Targeted In Stuttering Assessment And Treatment. *American Journal Of Speech-Language Pathology*, 24(2), 139-149.
https://doi.org/10.1044/2014_ajslp-14-0087
- Burdea, G. C., & Coiffet, P. (2017). *Virtual reality technology*. John Wiley & Sons. <https://www.wiley.com/en-us/Virtual+Reality+Technology%2C+3rd+Edition-p-9781394306947>
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). *A database of German emotional speech*. <http://emodb.bilderbar.info/docu/>
- Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O* NET). *ETS Research Report Series*, 2013(2), i-55. <https://doi.org/https://doi.org/10.1002/j.2333-8504.2013.tb02328.x>

- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). *IEMOCAP: Interactive emotional dyadic motion capture database*. <https://sail.usc.edu/iemocap/>
- Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4), 582-596. <https://doi.org/10.1109/TASL.2008.2009578>
- Busso, C., Mariooryad, S., Metallinou, A., & Narayanan, S. (2013). Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE transactions on Affective computing*, 4(4), 386-397. <https://doi.org/10.1109/T-AFFC.2013.26>
- Byron, G., Strahm, A., Bagne, A., & Hilmert, C. (2023). Using virtual reality to study the impact of audience size on cortisol responses to the Trier Social Stress Test. *International Journal of Virtual Reality*, 23(1), 1-16. <https://doi.org/10.20870/IJVR.2023.23.1.7146>
- Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18-37. <https://doi.org/10.1109/T-AFFC.2010.1>
- Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4), 106-124. <https://doi.org/10.2307/1415404>
- Cannon, W. B. (1934). *The wisdom of the body*. Nature <https://doi.org/10.1038/133082a0>
- Carl, E., Stein, A. T., Levihn-Coon, A., Pogue, J. R., Rothbaum, B., Emmelkamp, P., Asmundson, G. J. G., Carlbring, P., Powers, M. B., Stockholms, u., Samhällsvetenskapliga, f., Psykologiska, i., & Klinisk, p. (2019). Virtual Reality Exposure Therapy For Anxiety And Related Disorders: A Meta-Analysis Of Randomized Controlled Trials. *Journal of Anxiety Disorders*, 61, 27-36. <https://doi.org/10.1016/j.janxdis.2018.08.003>
- Carnegie, D., & Esenwein, J. B. (2017). *The art of public speaking*. Courier Dover Publications. <https://www.funeralhelp.co.uk/files/The%20Art%20Of%20Public%20Speaking.pdf>

- Carrillo, E., Moya-Albiol, L., Gonzalez-Bono, E., Salvador, A., Ricarte, J., & Gomez-Amor, J. (2001). Gender differences in cardiovascular and electrodermal responses to public speaking task: the role of anxiety and mood states. *International Journal of Psychophysiology*, 42(3), 253-264. [https://doi.org/10.1016/s0167-8760\(01\)00147-7](https://doi.org/10.1016/s0167-8760(01)00147-7)
- Casale, S., Russo, A., & Serrano, S. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech communication*, 49(10-11), 801-810. <https://doi.org/10.1016/j.specom.2007.04.012>
- Chan, Y. T., & Ho, K. C. (1994, 1994). An efficient closed-form localization solution from time difference of arrival measurements. <https://doi.org/10.1109/ICASSP.1994.389638>
- Chanwimalueang, T., Aufegger, L., von Rosenberg, W., & Mandic, D. P. (2016, 2016). Modelling stress in public speaking: evolution of stress levels during conference presentations. <https://doi.org/10.1109/ICASSP.2016.7471788>
- Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A., & Scherer, S. (2015). Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan. <https://doi.org/10.1145/2750858.2806060>
- Chollet, M., Kalin, S., Helmut, P., & Stefan, S. (2015, 2015). Public speaking training with a multimodal interactive virtual audience framework - Demonstration. <https://doi.org/10.1145/2818346.2823294>
- Cicero, M. T. (1986). *Cicero on oratory and orators*. SIU Press. <https://www.hup.harvard.edu/books/9780674993839>
- Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6), 698-712. <https://doi.org/10.1037/amp0000399>
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications. https://www.ucg.ac.me/skladiste/blog_609332/objava_105202/fajlovi/Creswell.pdf

- Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., & Ouyang, M. (2022). A review: Music-emotion recognition and analysis based on EEG signals. *Frontiers in Neuroinformatics*, 16, 997282. <https://doi.org/10.3389/fninf.2022.997282>
- Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media psychology*, 19(2), 272-309. <https://doi.org/10.1080/15213269.2015.1015740>
- Cummings, K. E., & Clements, M. A. (1992, 1992). Improvements to and applications of analysis of stressed speech using glottal waveforms. <https://doi.org/10.1109/ICASSP.1992.226129>
- Curtis, K., Jones, G. J. F., & Campbell, N. (2015). *Effects of Good Speaking Techniques on Audience Engagement* Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA. <https://doi.org/10.1145/2818346.2820766>
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-Dependent Pre-Trained Deep Neural Networks For Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 30-42. <https://doi.org/10.1109/TASL.2011.2134090>
- Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. *Brain research reviews*, 26(2-3), 83-86. [https://doi.org/10.1016/S0165-0173\(97\)00064-7](https://doi.org/10.1016/S0165-0173(97)00064-7)
- Darwin, C., & Prodger, P. (1998). *The Expression Of The Emotions In Man And Animals*. Oxford University Press, USA. <https://doi.org/10.1093/oso/9780195112719.001.0001>
- Dascalu, M.-I., Bagis, S., Nitu, M., Ferche, O.-M., & Alin Dragos Bogdan, M. (2017). Experiential Learning VR System for Studying Computer Architecture. *Romanian Journal of Human - Computer Interaction*, 10(3), 197-215. <https://rochi.utcluj.ro/rrioc/articole/RRIOC-10-3-Dascalu.pdf>
- Dechant, M., Trimpl, S., Wolff, C., Mühlberger, A., & Shiban, Y. (2017). Potential Of Virtual Reality As A Diagnostic Tool For Social Anxiety: A Pilot Study. *Computers in Human behavior*, 76, 128-134. <https://doi.org/10.1016/j.chb.2017.07.005>
- Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational psychologist*, 26(3-4), 325-346. https://doi.org/10.1207/s15326985ep2603&4_6

- Demenko, G. (2008, 2008). Voice stress extraction. <https://doi.org/10.21437/speechprosody.2008-12>
- Deng, L. (2006). *Dynamic speech models: theory, algorithms, and applications* (Vol. 2). Morgan & Claypool Publishers. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Deng-Book2006.pdf>
- Dillon, R. (2001, 2001). Extracting audio cues in real time to understand musical expressiveness. https://doi.org/10.1007/11751069_6
- Dillon, R. (2003, 2003). A statistical approach to expressive intention recognition in violin performances. https://doi.org/10.1007/11751069_6
- Dillon, R., Wong, G., & Ang, R. (2006). Virtual orchestra: An immersive computer game for fun and education. Proceedings of the 2006 international conference on Game research and development. <https://doi.org/10.5555/1234341.1234375>
- Docan-Morgan, T., & Nelson, L. L. (2015). Chapter 11: The Benefits and Necessity of Public Speaking Education – Tony Docan-Morgan & Laura L. Nelson. In *Public Speaking for the Curious*. https://www.researchgate.net/publication/332974615_The_Benefits_and_Necessity_of_Public_Speaking_Education
- Dolka, H., Vm, A. X., & Juliet, S. (2021). Speech emotion recognition using ANN on MFCC features. <https://doi.org/10.1007/s10772-024-10095-8>
- Dongrui, W., Courtney, C. G., Lance, B. J., Narayanan, S. S., Dawson, M. E., Oie, K. S., & Parsons, T. D. (2010). Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals: Virtual Reality Stroop Task. *IEEE Transactions on Affective Computing*, 1(2), 109-118. <https://doi.org/10.1109/T-AFFC.2010.12>
- Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., & Batliner, A. (2007). The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. International conference on affective computing and intelligent interaction. https://doi.org/10.1007/978-3-540-74889-2_43

- Dowd, H., Dowd, H., Zautra, A., Zautra, A., Hogan, M., & Hogan, M. (2010). Emotion, Stress, and Cardiovascular Response: An Experimental Test of Models of Positive and Negative Affect. *International Journal of Behavioral Medicine*, 17(3), 189-194.
<https://doi.org/10.1007/s12529-009-9063-3>
- Droppleman, L. F., & McNair, D. M. (1971). An experimental analog of public speaking. *Journal of Consulting and Clinical Psychology*, 36(1), 91. <https://doi.org/10.1037/h0030420>
- Dwyer, K. K., & Davidson, M. M. (2012). Is public speaking really more feared than death? *Communication Research Reports*, 29(2), 99-107.
<https://doi.org/10.1080/08824096.2012.667772>
- Echternach, M., Burk, F., Burdumy, M., Traser, L., & Richter, B. (2016). Morphometric Differences of Vocal Tract Articulators in Different Loudness Conditions in Singing. *PloS one*, 11(4), e0153792. <https://doi.org/10.1371/journal.pone.0153792>
- Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60), 16.
<https://doi.org/10.1002/0470013494.ch3>
- El-Assaad, M. A., Topouchian, J. A., Darné, B. M., & Asmar, R. G. (2002). Validation of the Omron HEM-907 device for blood pressure measurement. *Blood Press Monit*, 7(4), 237-241.
<https://doi.org/10.1097/00126097-200208000-00006>
- El-Yamri, M., Romero-Hernandez, A., Gonzalez-Riojo, M., & Manero, B. (2019). Designing a VR game for public speaking based on speakers features: a case study. *Smart Learning Environments*, 6(1), 12. <https://doi.org/10.1186/s40561-019-0094-1>
- El Assaad, M. A., Topouchian, J. A., Darne, B. M., & Asmar, R. G. (2002). Validation of the Omron HEM-907 device for blood pressure measurement. *Blood pressure monitoring*, 7(4), 237-241.
<https://doi.org/10.1097/00126097-200208000-00006>
- El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3), 572-587.
<https://doi.org/10.1016/j.patcog.2010.09.020>

- Elfering, A., & Grebner, S. (2011). Ambulatory assessment of skin conductivity during first thesis presentation: lower self-confidence predicts prolonged stress response. *Appl Psychophysiol Biofeedback*, 36(2), 93-99. <https://doi.org/10.1007/s10484-011-9152-3>
- Fawcett, S. B., & Miller, L. K. (1975). TRAINING PUBLIC - SPEAKING BEHAVIOR: AN EXPERIMENTAL ANALYSIS AND SOCIAL VALIDATION 1. *Journal of Applied Behavior Analysis*, 8(2), 125-135. <https://doi.org/10.1901/jaba.1975.8-125>
- Feldman, P. J., Cohen, S., Hamrick, N., & Lepore, S. J. (2004). Psychological stress, appraisal, emotion and cardiovascular response in a public speaking task. *Psychology & Health*, 19(3), 353-368. <https://doi.org/10.1080/0887044042000193497>
- Fidell, L. S., & Tabachnick, B. G. (2003). Preparatory data analysis. *Handbook of psychology: Research methods in psychology*, 2, 115-141. <https://doi.org/10.1002/0471264385.wei0205>
- Fisher, S. S., McGreevy, M., Humphries, J., & Robinett, W. (1987, 1987). Virtual environment display system. <https://doi.org/10.1145/319120.319127>
- Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes. *Journal of personality and social psychology*, 50(5), 992. <https://doi.org/10.1037/0022-3514.50.5.992>
- Fredrick, S. S., Demaray, M. K., Malecki, C. K., & Dorio, N. B. (2018). Can social support buffer the association between depression and suicidal ideation in adolescent boys and girls? *Psychology in the Schools*, 55(5), 490-505. <https://doi.org/10.1002/pits.22125>
- Furmark, T., Tillfors, M., Everz, P. O., Marteinsdottir, I., Gefvert, O., & Fredrikson, M. (1999). Social phobia in the general population: prevalence and sociodemographic profile. *Social psychiatry and psychiatric epidemiology*, 34, 416-424. <https://doi.org/10.1007/s001270050163>
- Furness Iii, T. A. (1986). The super cockpit and its human factors challenges.
- Gallego, A., McHugh, L., Penttonen, M., & Lappalainen, R. (2022). Measuring public speaking anxiety: self-report, behavioral, and physiological. *Behavior Modification*, 46(4), 782-798. <https://doi.org/10.1177/0145445521994308>

- Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved 7/2/2022 from <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>
- Garcia-Palacios, A., Hoffman, H., Carlin, A., Furness, T. A., & Botella, C. (2002). Virtual Reality In The Treatment Of Spider Phobia: A Controlled Study. *Behaviour Research and Therapy*, 40(9), 983-993. [https://doi.org/10.1016/S0005-7967\(01\)00068-7](https://doi.org/10.1016/S0005-7967(01)00068-7)
- Gareis, E. (2006). Guidelines for public speaking. *Department of Communication Studies, Baruch College/CUNY, New York*. <https://weissman.baruch.cuny.edu/wp-content/uploads/sites/20/2020/09/Guidelines.pdf>
- Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating Virtual Reality And Augmented Reality Training For Industrial Maintenance And Assembly Tasks. *Interactive Learning Environments*, 23(6), 778-798. <https://doi.org/10.1080/10494820.2013.815221>
- Gebara, C. M., Barros-Neto, T. P. d., Gertsenchtein, L., & Lotufo-Neto, F. (2016). Virtual Reality Exposure Using Three-Dimensional Images For The Treatment Of Social Phobia. *Revista Brasileira De Psiquiatria (Sao Paulo, Brazil : 1999)*, 38(1), 24-29. <https://doi.org/10.1590/1516-4446-2014-1560>
- Gerardi, M., Cukor, J., Difede, J., Rizzo, A., & Rothbaum, B. O. (2010). Virtual Reality Exposure Therapy for Post-Traumatic Stress Disorder and Other Anxiety Disorders. *Current Psychiatry Reports*, 12(4), 298-305. <https://doi.org/10.1007/s11920-010-0128-4>
- Giannakopoulos, T. (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one*, 10(12), e0144610. <https://doi.org/10.1371/journal.pone.0144610>
- Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to audio analysis: a MATLAB® approach*. Academic Press. <https://doi.org/10.1016/C2012-0-03524-7>
- Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: a review. *Journal of voice*, 27(3), 390-e321. <https://doi.org/10.1016/j.jvoice.2012.12.010>

- Gigante, M. A. (1993). 1 - Virtual Reality: Definitions, History and Applications. In R. A. Earnshaw, M. A. Gigante, & H. Jones (Eds.), *Virtual Reality Systems* (pp. 3-14). Academic Press.
<https://doi.org/10.1016/B978-0-12-227748-1.50009-3>
- Gillespie, S., Moore, E., Laures-Gore, J., Farina, M., Russell, S., & Logan, Y.-Y. (2017, 2017).
 Detecting stress and depression in adults with aphasia through speech analysis.
<https://doi.org/10.1109/ICASSP.2017.7953136>
- Gillespie, S., Moore, E., Laures-Gore, J., Farina, M., Russell, S., Logan, Y. Y., & Ieee. (2017).
 Detecting Stress And Depression In Adults With Aphasia Through Speech Analysis. In *2017 Ieee International Conference on Acoustics, Speech and Signal Processing* (pp. 5140-5144).
<https://doi.org/10.1109/ICASSP.2017.7953136>
- Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Jauregui, D. A. G., Eyharabide, V., Filaire, E., Le Scanff, C., & Devillers, L. (2013, 2013). Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses.
<https://doi.org/10.1109/ACII.2013.75>
- Godino-Llorente, J. I., Gomez-Vilda, P., & Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE Transactions on Biomedical Engineering*, 53(10), 1943-1953.
<https://doi.org/10.1109/TBME.2006.871883>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
<https://mitpress.mit.edu/9780262035613/deep-learning/>
- Goyal, A., Singh, S., Vir, D., & Pershad, D. (2016). Automation of stress recognition using subjective or objective measures. *Psychological Studies*, 61, 348-364. <https://doi.org/10.1007/s12646-016-0379-1>
- Greene, S., Thapliyal, H., & Caban-Holt, A. (2016). A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, 5(4), 44-56. <https://doi.org/10.1109/MCE.2016.2590178>

- Grieve, R., Woodley, J., Hunt, S. E., & McKay, A. (2021). Student fears of oral presentations and public speaking in higher education: a qualitative survey. *Journal of Further and Higher Education*, 45(9), 1281-1293. <https://doi.org/10.1080/0309877X.2021.1948509>
- Hakim, A., Marsland, S., & Guesgen, H. W. (2013, 2013). Computational analysis of emotion dynamics. <https://doi.org/10.1109/ACII.2013.37>
- Hamid, O. K. (2018). Frame blocking and windowing speech signal. *Journal of Information, Communication, and Intelligence Systems (JICIS)*, 4, 87-94.
https://www.researchgate.net/publication/331635757_Frame_Blocking_and_Windowing_Speech_Signal
- Han, H., Byun, K., & Kang, H.-G. (2018, 2018). A Deep Learning-based Stress Detection Algorithm with Speech Signal. *AVSU'18 International Multimedia Conference*.
<https://doi.org/10.1145/3264869.3264875>
- Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech 2014*. <https://doi.org/10.21437/Interspeech.2014-57>
- Hancock, A. B., Stone, M. D., Brundage, S. B., & Zeigler, M. T. (2010). Public speaking attitudes: does curriculum make a difference? *Journal of voice*, 24(3), 302-307.
<https://doi.org/10.1016/j.jvoice.2008.09.007>
- Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., & Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. *Eurospeech*.
<https://doi.org/10.21437/Eurospeech.1997-494>
- Hansen, J. H. L., & Patil, S. (2007). Speech Under Stress: Analysis, Modeling and Recognition. In C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods* (pp. 108-137). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74200-5_6
- Hansen, J. H. L., & Womack, B. D. (1996). Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on speech and audio processing*, 4(4), 307-313.
<https://doi.org/10.1109/89.506935>

- Hao, M., Cao, W.-H., Liu, Z.-T., Wu, M., & Xiao, P. (2020). Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing*, 391, 42-51. <https://doi.org/10.1016/j.neucom.2020.01.048>
- Harris, D., Arthur, T., Wilson, M., & Vine, S. (2023, 10-13 Sept. 2023). Eye tracking for affective computing in virtual reality healthcare applications. 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). <https://doi.org/10.1109/ACIIW59127.2023.10388102>
- Hartanto, D., Kampmann, I. L., Morina, N., Emmelkamp, P. G., Neerincx, M. A., & Brinkman, W.-P. (2014). Controlling social stress in virtual reality environments. *PloS one*, 9(3), e92804. <https://doi.org/10.1371/journal.pone.0092804>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR
- Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2), 156-166. <https://doi.org/10.1109/TITS.2005.848368>
- Heidemann, K. (2016). A System for Describing Vocal Timbre in Popular Song. *Music Theory Online*, 22(1). <https://doi.org/10.30535/mt.22.1.2>
- Heilig, M. (1998). Beginnings: sensorama and the telesphere mask. In *Digital illusion: entertaining the future with high technology* (pp. 343-351). <https://doi.org/10.5555/274450>
- Heimberg, R. G., Salzman, D. G., Holt, C. S., & Blendell, K. A. (1993). Cognitive—behavioral group treatment for social phobia: Effectiveness at five-year followup. *Cognitive therapy and Research*, 17(4), 325-339. <https://psycnet.apa.org/doi/10.1007/BF01177658>
- Hellhammer, D. H., Stone, A. A., Hellhammer, J., & Broderick, J. (2010). Measuring Stress. In G. F. Koob, M. L. Moal, & R. F. Thompson (Eds.), *Encyclopedia of Behavioral Neuroscience* (pp. 186-191). Academic Press. <https://doi.org/10.1016/B978-0-08-045396-5.00188-3>

- Herrera, F., Oh, S. Y., & Bailenson, J. N. (2020). Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence*, 27(2), 163-182.
https://doi.org/10.1162/pres_a_00324
- Hilfert, T., & König, M. (2016). Low-Cost Virtual Reality Environment For Engineering And Construction. *Visualization in Engineering*, 4(1), 1-18. <https://doi.org/10.1186/s40327-015-0031-5>.
- Hilmy, M. S. H., Asnawi, A. L., Jusoh, A. Z., Abdullah, K., Ibrahim, S. N., Ramli, H. A. M., & Azmin, N. F. M. (2021). Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning. <https://doi.org/10.1109/ICCCE50029.2021.9467176>
- Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92, 84-89. <https://doi.org/10.1007/s00421-004-1055-z>
- Hu, H., Xu, M., & Wu, W. (2007, 15-20 April 2007). GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. <https://doi.org/10.1109/ICASSP.2007.366937>
- IBM. (2020). *Recurrent Neural Networks*. Retrieved 7/2/2022 from <https://www.ibm.com/cloud/learn/recurrent-neural-networks>
- Ilmi, J., Suharsono, Y. Y., & Ingarianti, T. T. M. (2017, 2017). Self-hypnosis training to improve self confidence in students for speaking in public. <https://doi.org/10.2991/icet-17.2017.23>
- Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical signal processing and control*, 59, 101894.
<https://doi.org/10.1016/j.bspc.2020.101894>
- Jackson, E. M., & Dishman, R. K. (2006). Cardiorespiratory fitness and laboratory stress: A meta - regression analysis. *Psychophysiology*, 43(1), 57-72. <https://doi.org/10.1111/j.1469-8986.2006.00373.x>

- Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*.
<https://doi.org/10.48550/arXiv.2002.07590>
- James, W. (1884). II.—WHAT IS AN EMOTION ? *Mind, os-IX*(34), 188-205.
<https://doi.org/10.1093/mind/os-IX.34.188>
- Jezova, D., Hlavacova, N., Dicko, I., Solarikova, P., & Brezina, I. (2016). Psychosocial Stress Based On Public Speech In Humans: Is There A Real Life/Laboratory Setting Cross-Adaptation? *Stress, 19*(4), 429-433. <https://doi.org/10.1080/10253890.2016.1203416>
- Jiang, H., Vimalasvaran, S., Wang, J. K., Lim, K. B., Mogali, S. R., & Car, L. T. (2022). Virtual reality in medical students' education: scoping review. *JMIR medical Education, 8*(1), e34860. <https://doi.org/10.2196/34860>
- Jönsson, P., Wallergård, M., Österberg, K., Hansen, Å. M., Johansson, G., & Karlson, B. (2010). Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the Trier Social Stress Test: A pilot study. *Psychoneuroendocrinology, 35*(9), 1397-1403.
<https://doi.org/10.1016/j.psyneuen.2010.04.003>
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 15*(4), 531-538. <https://doi.org/https://doi.org/10.1002/sam.11583>
- Junek, W. (2007). Mind reading: The interactive guide to emotions. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 16*(4), 182.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC2247465/>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull, 129*(5), 770-814.
<https://doi.org/10.1037/0033-2909.129.5.770>
- Kaiser, J. F. (1993). Some useful properties of Teager's energy operators. 1993 IEEE international conference on acoustics, speech, and signal processing.
<https://doi.org/10.1109/ICASSP.1993.319457>
- Kalatzis, A., Stanley, L., & Prabhu, V. G. (2021, 15-17 Nov. 2021). Affective State Classification in Virtual Reality Environments Using Electrocardiogram and Respiration Signals. 2021 IEEE

- International Conference on Artificial Intelligence and Virtual Reality (AIVR).
<https://doi.org/10.1109/AIVR52153.2021.00037>
- Kallet, M. (2014). *Think smarter: critical thinking to improve problem-solving and decision-making skills*. John Wiley & Sons
- Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, 27(2), 237-265.
<https://doi.org/10.3758/s13423-019-01701-x>
- Kandali, A. B., Routray, A., & Basu, T. K. (2008). Emotion recognition from Assamese speeches using MFCC features and GMM classifier. <https://doi.org/10.1109/TENCON.2008.4766487>
- Kappas, A., Hess, U., & Scherer, K. R. (1991). Voice and emotion. *Fundamentals of nonverbal behavior*, 200. <https://psycnet.apa.org/record/1991-98576-006>
- Karthikeyan, R., McDonald, A. D., & Mehta, R. K. (2023). Stress Detection During Motor Activity: Comparing Neurophysiological Indices in Older Adults. *IEEE Transactions on Affective Computing*, 14(3), 2224-2237. <https://doi.org/10.1109/TAFFC.2022.3148234>
- Kassem, M., Benomran, L., & Teizer, J. (2017). Virtual Environments For Safety Learning In Construction And Engineering: Seeking Evidence And Identifying Gaps For Future Research. *Visualization in Engineering*, 5(1), 1-15. <https://doi.org/10.1186/s40327-017-0054-1>
- Kemeny, M. E. (2003). The psychobiology of stress. *Current directions in psychological science*, 12(4), 124-129. <https://doi.org/10.1111/1467-8721.01246>
- Kenny, P. (1982). *A handbook of public speaking for scientists and engineers*. CRC Press.<https://www.routledge.com/A-Handbook-of-Public-Speaking-for-Scientists-and-Engineers/Kenny/p/book/9780852745533>
- Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation*, 15(3), 235.
<https://doi.org/10.30773/pi.2017.08.17>
- Kim, K., Kim, C.-H., Kim, S.-Y., Roh, D., & Kim, S. I. (2009). Virtual Reality for Obsessive-Compulsive Disorder: Past and the Future. *Psychiatry investigation*, 6(3), 115-121.
<https://doi.org/10.4306/pi.2009.6.3.115>

- Kimani, E., & Bickmore, T. (2019). *Addressing Public Speaking Anxiety in Real-time Using a Virtual Public Speaking Coach and Physiological Sensors* Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, Paris, France.
<https://doi.org/10.1145/3308532.3329409>
- King, M. G., Burrows, G. D., & Stanley, G. V. (1983). Measurement of stress and arousal: Validation of the stress/arousal adjective checklist. *British journal of psychology*, 74(4), 473-479.
<https://doi.org/10.1111/j.2044-8295.1983.tb01880.x>
- Knight, M. L., Johnson, K. G., & Stewart, F. (2016). Reducing Student Apprehension of Public Speaking: Evaluating Effectiveness of Group Tutoring Practices. *Learning Assistance Review*, 21(1), 21-54. <https://shorturl.at/lfL6p>
- Ko, S. H., & Rogers, J. (2021). Functional materials and devices for XR (VR/AR/MR) applications. In (Vol. 31, pp. 2106546). <https://doi.org/10.1002/adfm.202106546>
- Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: speech database for emotion analysis. International conference on contemporary computing.
<https://doi.org/10.1002/adfm.202106546>
- Koolagudi, S. G., & Rao, K. S. (2012). Emotion Recognition From Speech: A Review. *International journal of speech technology*, 15(2), 99-117. <https://doi.org/10.1002/adfm.202106546>
- Koroleva, M., Bakhchina, A., Shyshalov, I., Parin, S. B., & Polevaia, S. A. (2014). Influence of the context of public speaking on human functional state. *International Journal of Psychophysiology*, 2(94), 230-231. <https://doi.org/10.1016/j.ijpsycho.2014.08.901>
- Kothgassner, O. D., Felnhofer, A., Hlavacs, H., Beutl, L., Palme, R., Kryspin-Exner, I., & Glenk, L. M. (2016). Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. *Computers in Human behavior*, 62, 124-135.
<https://doi.org/10.1016/j.chb.2016.03.081>
- Kothgassner, O. D., Goreis, A., Kafka, J. X., Kaufmann, M., Atteneeder, K., Beutl, L., Hennig-Fast, K., Hlavacs, H., & Felnhofer, A. (2019). Virtual social support buffers stress response: an experimental comparison of real-life and virtual support prior to a social stressor. *Journal of*

- behavior therapy and experimental psychiatry*, 63, 57-65.
<https://doi.org/10.1016/j.jbtep.2018.11.003>
- Kotlyar, M., Donahue, C., Thuras, P., Kushner, M. G., O'Gorman, N., Smith, E. A., & Adson, D. E. (2008). Physiological Response To A Speech Stressor Presented In A Virtual Reality Environment. *Psychophysiology*, 45(6), 1034-1037. <https://doi.org/10.1111/j.1469-8986.2008.00690.x>
- Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons. <https://doi.org/10.1002/9781444395068>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25. <https://doi.org/10.1145/3065386>
- Kroczek, L. O. H., & Mühlberger, A. (2023). Public speaking training in front of a supportive audience in Virtual Reality improves performance in real-life. *Scientific Reports*, 13(1), 13968. <https://doi.org/10.1038/s41598-023-41155-9>
- Kuchibhotla, S., Vankayalapati, H. D., Vaddi, R., & Anne, K. R. (2014). A comparative analysis of classifiers in emotion recognition through acoustic features. *International journal of speech technology*, 17(4), 401-408. <https://doi.org/10.1007/s10772-014-9239-3>
- Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., & Lis, A. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics*, 13(15), 2582. <https://doi.org/10.3390/diagnostics13152582>
- Kumar, G. S. S., Sampathila, N., & Martis, R. J. (2023). Classification of human emotional states based on valence-arousal scale using electroencephalogram. *Journal of Medical Signals & Sensors*, 13(2), 173-182. https://doi.org/10.4103/jmss.jmss_169_21
- Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. Eighth European conference on speech communication and technology. <https://doi.org/10.21437/Eurospeech.2003-80>

- Laver, K. E., George, S., Thomas, S., Deutsch, J. E., & Crotty, M. (2015). Virtual reality for stroke rehabilitation. *The Cochrane database of systematic reviews*(2), CD008349.
<https://doi.org/10.1002/14651858.CD008349.pub3>
- Lazarus, R. S. (2000). Evolution of a model of stress, coping, and discrete emotions. *Handbook of stress, coping, and health: Implications for nursing research, theory, and practice*, 195-222.
https://books.google.com.sg/books/about/Handbook_of_Stress_Coping_and_Health.html?id=TMVxBf10IHwC&redir_esc=y
- Le, Q. T., Le, Q. T., Pedro, A., Pedro, A., Park, C. S., & Park, C. S. (2015). A Social Virtual Reality Based Construction Safety Education System for Experiential Learning. *Journal of Intelligent & Robotic Systems*, 79(3), 487-506. <https://doi.org/10.1007/s10846-014-0112-z>
- Lech, M., & He, L. (2014). Stress and Emotion Recognition Using Acoustic Speech Analysis. In M. Lech, I. Song, P. Yellowlees, & J. Diederich (Eds.), *Mental Health Informatics* (pp. 163-184). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38550-6_9
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
<https://doi.org/10.1038/nature14539>
- Lefter, I., Luxton, D. D., Baird, A., Chaspari, T., Hammal, Z., Mahmoud, M., & Salah, A. A. (2023, 10-13 Sept. 2023). Affective Computing for Mental Wellbeing: Challenges, Opportunities, and Promising Synergies. 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW).
<https://doi.org/10.1109/ACIIW59127.2023.10388209>
- Levy, F., Leboucher, P., Rautureau, G., & Jouvent, R. (2016). E-Virtual Reality Exposure Therapy In Acrophobia: A Pilot Study. *Journal of Telemedicine and Telecare*, 22(4), 215-220.
<https://doi.org/10.1177/1357633X15598243>
- Li, R., & Liu, Z. (2020). Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making*, 20, 1-10. <https://doi.org/10.1186/s12911-020-01299-4>
- Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features.
<https://doi.org/10.1109/ICASSP.2007.367261>

- Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07.
<https://doi.org/10.1109/ICASSP.2007.367261>
- Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. 2005 international conference on machine learning and cybernetics.
<https://doi.org/10.1109/ICMLC.2005.1527805>
- Lindner, P., Dagöö, J., Hamilton, W., Miloff, A., Andersson, G., Schill, A., & Carlbring, P. (2021). Virtual Reality exposure therapy for public speaking anxiety in routine care: a single-subject effectiveness trial. *Cognitive Behaviour Therapy*, 50(1), 67-87.
<https://doi.org/10.1080/16506073.2020.1795240>
- Lindner, P., Miloff, A., Fagnäs, S., Andersen, J., Sigeman, M., Andersson, G., Furmark, T., Carlbring, P., Linköpings, u., Institutionen för beteendevetenskap och, l., Filosofiska, f., & Psykologi. (2019). Therapist-led and self-led one-session virtual reality exposure therapy for public speaking anxiety with consumer hardware and software: A randomized controlled trial. *Journal of Anxiety Disorders*, 61, 45-54. <https://doi.org/10.1016/j.janxdis.2018.07.003>
- Lipp, M. N., & Anderson, D. E. (1999). Cardiovascular reactivity to simulated social stress. *Stress Medicine*, 15(4), 249-257. [https://doi.org/10.1002/\(SICI\)1099-1700\(199910\)15:4<249::AID-SMI821>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1099-1700(199910)15:4<249::AID-SMI821>3.0.CO;2-3)
- Lister, H. A., Piercey, C. D., & Joordens, C. (2010). The effectiveness of 3-D video virtual reality for the treatment of fear of public speaking. *Journal of CyberTherapy and Rehabilitation*, 3(4), 375. <https://shorturl.at/4Gs9h>
- Liu, J., Wu, G., Luo, Y., Qiu, S., Yang, S., Li, W., & Bi, Y. (2020). EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder [Original Research]. *Frontiers in Systems Neuroscience*, 14.
<https://www.frontiersin.org/articles/10.3389/fnsys.2020.00043>

- Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5), e0196391. <https://doi.org/10.1371/journal.pone.0196391>
- Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2010). Detection Of Clinical Depression In Adolescents' Speech During Family Interactions. *IEEE Transactions on Biomedical Engineering*, 58(3), 574-586. <https://doi.org/10.1109/TBME.2010.2091640>
- Lucas, S. E., & Stob, P. (2020). *The art of public speaking*. McGraw-Hill
- Lynch, J. J., Thomas, S. A., Long, J. M., Malinow, K. L., Friedmann, E., & Katcher, A. H. (1982). Blood pressure changes while talking. *Isr J Med Sci*, 18(5), 575-579. <https://pubmed.ncbi.nlm.nih.gov/7096040/>
- Mandel, S. (2000). *Effective presentation skills a practical guide for better speaking*. Crisp Learning. <https://www.amazon.sg/Effective-Presentation-Skills-Practical-Speaking/dp/156052202X>
- Marinho, A. C. F., de Medeiros, A. M., Gama, A. C. C., & Teixeira, L. C. (2017). Fear of public speaking: Perception of college students and correlates. *Journal of voice*, 31(1), 127-e127. 10.1016/j.jvoice.2015.12.012
- Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., & Shapiro, A. (2013). Virtual character performance from speech. Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. <https://doi.org/10.1145/2485895.2485900>
- Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. 22nd International Conference on Data Engineering Workshops (ICDEW'06). <https://doi.org/10.1109/ICDEW.2006.145>
- Martin., D. J. J. H. (2021). *Hidden Markov Models*. <https://web.stanford.edu/~jurafsky/slp3/A.pdf>
- Maskey, M., Lowry, J., Rodgers, J., McConachie, H., & Parr, J. R. (2014). Reducing specific phobia/fear in young people with autism spectrum disorders (ASDs) through a virtual reality environment intervention. *PloS one*, 9(7), e100374. <https://doi.org/10.1371/journal.pone.0100374>

- McCroskey, J. C. (2015). *An introduction to rhetorical communication*. Routledge
<https://doi.org/10.4324/9781315663791>
- McDuff, D., Gontarek, S., & Picard, R. (2014, 2014). Remote measurement of cognitive stress via heart rate variability. <https://doi.org/10.1109/EMBC.2014.6944243>
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. Proceedings of the 14th python in science conference. <https://doi.org/10.25080/Majora-7b98e3ed-003>
- Mejbri, N., Essalmi, F., Jemni, M., & Alyoubi, B. A. (2022). Trends in the use of affective computing in e-learning environments. *Education and Information Technologies*, 1-23.
<https://doi.org/10.1007/s10639-021-10769-9>
- Menzel, K. E., & Carrell, L. J. (1994). The relationship between preparation and performance in public speaking. *Communication Education*, 43(1), 17-26.
<https://doi.org/10.1080/03634529409378958>
- Mihelj, M., Novak, D., & Beguš, S. (2014). Virtual reality technology and applications.
<https://doi.org/10.1007/978-94-007-6910-6>
- Mikus, C. R., Earnest, C. P., Blair, S. N., & Church, T. S. (2009). Heart rate and exercise intensity during training: observations from the DREW Study. *British journal of sports medicine*, 43(10), 750-755. <https://doi.org/10.1136/bjsm.2008.046342>
- Miloff, A., Lindner, P., Hamilton, W., Reuterskiöld, L., Andersson, G., Carlbring, P., Stockholms, u., Samhällsvetenskapliga, f., Psykologiska, i., & Klinisk, p. (2016). Single-session gamified virtual reality exposure therapy for spider phobia vs. traditional exposure therapy: study protocol for a randomized controlled non-inferiority trial. *Trials*, 17(1), 60.
<https://doi.org/10.1186/s13063-016-1171-1>
- Mishra, S. P., Warule, P., & Deb, S. (2023). Speech emotion recognition using MFCC-based entropy feature. *Signal, Image and Video Processing*. <https://doi.org/10.1007/s11760-023-02716-7>
- Mitchell, R. L., & Ross, E. D. (2013). Attitudinal Prosody: What We Know And Directions For Future Study. *Neuroscience & Biobehavioral Reviews*, 37(3), 471-479.
<https://doi.org/10.1016/j.neubiorev.2013.01.027>

- Mohapatra, S. K., Kanna, R. K., Arora, G., Sarangi, P. K., Mohanty, J., & Sahu, P. (2022). Systematic Stress Detection in CNN Application. <https://doi.org/10.1109/ICRITO56286.2022.9964761>
- Molina, K. I., Ricci, N. A., de Moraes, S. A., & Perracini, M. R. (2014). Virtual reality using games for improving physical functioning in older adults: a systematic review. *Journal of neuroengineering and rehabilitation*, 11(1), 156. <https://doi.org/10.1186/1743-0003-11-156>
- Montazer, G. A., Giveki, D., Karami, M., & Rastegar, H. (2018). Radial basis function neural networks: A review. *Comput. Rev. J*, 1(1), 52-74. <https://europub.co.uk/articles/-A-433738>
- Mostajeran, F., Balci, M. B., Steinicke, F., Kühn, S., & Gallinat, J. (2020, 22-26 March 2020). The Effects of Virtual Audience Size on Social Anxiety during Public Speaking. 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). <https://doi.org/10.1109/VR46266.2020.00050>
- Motraghi, T. E., Seim, R. W., Meyer, E. C., & Morissette, S. B. (2014). Virtual Reality Exposure Therapy for the Treatment of Posttraumatic Stress Disorder: A Methodological Review Using CONSORT Guidelines. *Journal of Clinical Psychology*, 70(3), 197-208. <https://doi.org/10.1002/jclp.22051>
- Mowbray, R., & Perry, L. B. (2015). Improving lecture quality through training in public speaking. *Innovations in Education and Teaching International*, 52(2), 207-217. <https://doi.org/10.1080/14703297.2013.849205>
- Nalini, N. J., & Palanivel, S. (2016). Music emotion recognition: The combined evidence of MFCC and residual phase. *Egyptian Informatics Journal*, 17(1), 1-10. <https://doi.org/10.1016/j.eij.2015.05.004>
- Naylor, P. A., Kounoudes, A., Gudnason, J., & Brookes, M. (2006). Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), 34-43. <https://doi.org/10.1109/TASL.2006.876878>
- Neiberg, D., Elenius, K., Laskowski, K., & Isca. (2006). *Emotion Recognition in Spontaneous Speech Using GMMs*. <https://doi.org/10.21437/Interspeech.2006-277>
- Nijholt, A. (2014). Breaking fresh ground in human–media interaction research. *Frontiers in ICT*, 1, 4. <https://doi.org/10.3389/fict.2014.00004>

- Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. Seventh European Conference on Speech Communication and Technology. <https://doi.org/10.21437/Eurospeech.2001-627>
- North, M. M., North, S. M., & Coble, J. R. (1998). Virtual reality therapy: an effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, 3(3), 1-6. <https://doi.org/10.20870/IJVR.1998.3.3.2625>
- Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003a). Detection of stress and emotion in speech using traditional and FFT based log energy features. Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint. <https://doi.org/10.1109/ICICS.2003.1292741>
- Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003b). Speech emotion recognition using hidden Markov models. *Speech communication*, 41(4), 603-623. <https://doi.org/10.1109/ICICS.2003.1292741>
- Ooi, C. S., Seng, K. P., Ang, L.-M., & Chew, L. W. (2014). A New Approach Of Audio Emotion Recognition. *Expert Systems With Applications*, 41(13), 5858-5869. <https://doi.org/10.1016/j.eswa.2014.03.026>
- Oprîș, D., Pinteș, S., García - Palacios, A., Botella, C., Szamosközi, Ş., & David, D. (2012). Virtual reality exposure therapy in anxiety disorders: a quantitative meta - analysis. *Depression and Anxiety*, 29(2), 85-93. <https://doi.org/10.1002/da.20910>
- Ordaz, N., Romero, D., Gorecky, D., & Siller, H. R. (2015). Serious Games and Virtual Simulator for Automotive Manufacturing Education & Training. *Procedia Computer Science*, 75, 267-274. <https://doi.org/10.1016/j.procs.2015.12.247>
- Palacios, D., Rodellar, V., Lázaro, C., Gómez, A., & Gómez, P. (2020). An ICA-based method for stress classification from voice samples. *Neural Computing and Applications*, 32(24), 17887-17897. <https://doi.org/10.1007/s00521-019-04549-3>
- Palmas, F., Cichor, J., Plecher, D. A., & Klinker, G. (2019, 2019). Acceptance and effectiveness of a virtual reality public speaking training. <https://doi.org/10.1109/ISMAR.2019.00034>

- Palo, H., Mohanty, M. N., & Chandra, M. (2015). Design of neural network model for emotional speech recognition. In *Artificial intelligence and evolutionary algorithms in engineering systems* (pp. 291-300). Springer. https://doi.org/10.1007/978-81-322-2135-7_32
- Pan, X., Gillies, M., Barker, C., Clark, D. M., & Slater, M. (2012). Socially anxious and confident men interact with a forward virtual woman: an experimental study. *PloS one*, 7(4), e32931. <https://doi.org/10.1371/journal.pone.0032931>
- Pan, X., & Hamilton, A. F. d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British journal of psychology*, 109(3), 395-417. <https://doi.org/10.1111/bjop.12290>
- Paredes, P., Sun, D., & Canny, J. (2013, 5-8 May 2013). Sensor-less sensing for affective computing and stress management technology. 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops. <https://doi.org/10.4108/icst.pervasivehealth.2013.252380>
- Parsons, T. D., & Rizzo, A. A. (2008). Affective Outcomes Of Virtual Reality Exposure Therapy For Anxiety And Specific Phobias: A Meta-Analysis. *Journal of behavior therapy and experimental psychiatry*, 39(3), 250-261. <https://doi.org/10.1016/j.jbtep.2007.07.007>
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013, 2013). On the difficulty of training recurrent neural networks. <https://doi.org/10.48550/arXiv.1211.5063>
- Peddinti, V., Povey, D., & Khudanpur, S. (2015, 2015). A time delay neural network architecture for efficient modeling of long temporal contexts. <https://doi.org/10.21437/Interspeech.2015-647>
- Pedro, A., Le, Q. T., & Park, C. S. (2016). Framework for Integrating Safety into Construction Methods Education through Interactive Virtual Reality. *Journal of Professional Issues in Engineering Education and Practice*, 142(2), 4015011. [https://doi.org/10.1061/\(ASCE\)EI.1943-5541.0000261](https://doi.org/10.1061/(ASCE)EI.1943-5541.0000261)
- Pellett, K., & Zaidi, S. F. M. (2019). *A Framework for Virtual Reality Training to Improve Public Speaking* 25th ACM Symposium on Virtual Reality Software and Technology, Parramatta, NSW, Australia. <https://doi.org/10.1145/3359996.3364727>

- Pertaub, D., Slater, M., & Barker, C. (2001). An Experiment On Fear Of Public Speaking In Virtual Reality. *Studies in health technology and informatics*, 372-378.
<https://doi.org/10.1037/e705412011-025>
- Pfister, T., & Robinson, P. (2011). Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2), 66-78. <https://doi.org/10.1109/T-AFFC.2011.8>
- Picard, R. W. (1997). Affective Computing. In: Citeseer.
<https://doi.org/10.7551/mitpress/1140.001.0001>
- Pisanski, K., Kobylarek, A., Jakubowska, L., Nowak, J., Walter, A., Błaszczński, K., Kasprzyk, M., Łysenko, K., Sukiennik, I., & Piątek, K. (2018). Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test. *Hormones and behavior*, 106, 52-61. <https://doi.org/10.1016/j.yhbeh.2018.08.014>
- Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. *Physiology & behavior*, 163, 234-238.
<https://doi.org/10.1016/j.physbeh.2016.05.018>
- Pisanski, K., & Sorokowski, P. (2021). Human stress detection: cortisol levels in stressed speakers predict voice-based judgments of stress. *Perception*, 50(1), 80-87.
<https://doi.org/10.1177/0301006620978378>
- Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4), 344-350. <https://www.proquest.com/scholarly-journals/nature-emotions/docview/215262319/se-2>
- Poeschl, S. (2017). Virtual reality training for public speaking—a QUEST-VR framework validation. *Frontiers in ICT*, 4, 13. <https://doi.org/10.3389/fict.2017.00013>
- Powers, M. B., Briceno, N. F., Gresham, R., Jouriles, E. N., Emmelkamp, P. M. G., & Smits, J. A. J. (2013). Do conversations with virtual avatars increase feelings of social anxiety?? *Journal of Anxiety Disorders*, 27(4), 398-403. <https://doi.org/10.1016/j.janxdis.2013.03.003>

- Premkumar, P., Heym, N., Brown, D. J., Battersby, S., Sumich, A., Huntington, B., Daly, R., & Zysk, E. (2021). The effectiveness of self-guided virtual-reality exposure therapy for public-speaking anxiety. *Frontiers in Psychiatry, 12*, 694610.
<https://doi.org/10.3389/fpsyt.2021.694610>
- Pribyl, C. B., Keaten, J., & Sakamoto, M. (2001). The effectiveness of a skills - based program in reducing public speaking anxiety. *Japanese Psychological Research, 43*(3), 148-155.
<https://doi.org/10.1111/1468-5884.t01-1-00171>
- Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286. <https://doi.org/10.1109/5.18626>
- Rahurkar, M. A., Hansen, J. H. L., Meyerhoff, J., Saviolakis, G., & Koenig, M. (2002, 2002). Frequency band analysis for stress detection using a teager energy operator based feature.
<https://doi.org/10.21437/ICSLP.2002-555>
- Raja, F. (2017). Anxiety Level in Students of Public Speaking: Causes and Remedies. *Journal of education and educational development, 4*(1), 94-110. <https://eric.ed.gov/?id=EJ1161521>
- Rao, K. S., Rao, K. S., Koolagudi, S. G., Koolagudi, S. G., Vempada, R. R., & Vempada, R. R. (2013). Emotion Recognition From Speech Using Global And Local Prosodic Features. *International journal of speech technology, 16*(2), 143-160.
<https://doi.org/10.1002/adfm.202106546>
- Rhodes, S. C., & Frandsen, K. D. (1975). Some effects of instruction in feedback utilization on the fluency of college students' speech. *Communications Monographs, 42*(1), 83-89.
<https://doi.org/10.1080/03637757509375881>
- Rivlin, R. (1948). Large elastic deformations of isotropic materials. I. Fundamental concepts. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 240*(822), 459-490. <https://doi.org/10.1098/rsta.1948.0002>

- Rong, J., Li, G., & Chen, Y.-P. P. (2009). Acoustic Feature Selection For Automatic Emotion Recognition From Speech. *Information Processing and Management*, 45(3), 315-328.
<https://doi.org/10.1016/j.ipm.2008.09.003>
- Roseman, I. J., & Smith, C. A. (2001). Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, 3-19. <https://psycnet.apa.org/record/2001-06810-001>
- Rosén, J., Kastrati, G., Reppling, A., Bergkvist, K., Åhs, F., Mittuniversitetet, Institutionen för psykologi och socialt, a., & Fakulteten för, h. (2019). The effect of immersive virtual reality on proximal and conditioned threat. *Scientific Reports*, 9(1), 1-8.
<https://doi.org/10.1038/s41598-019-53971-z>
- Ruscio, A. M., Brown, T. A., Chiu, W. T., Sareen, J., Stein, M. B., & Kessler, R. C. (2008). Social Fears And Social Phobia In The USA: Results From The National Comorbidity Survey Replication. *Psychological Medicine*, 38(1), 15-28.
<https://doi.org/10.1017/S0033291707001699>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6), 1161. <https://doi.org/10.1037/h0077714>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1), 145. <https://doi.org/10.1037/0033-295X.110.1.145>
- Ryerson Multimedia Research Lab., R. U. (2017). *RML Emotion Database*.
<http://shachi.org/resources/4965>
- Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech communication*, 54(4), 543-565. <https://doi.org/10.1016/j.specom.2011.11.004>
- Sano, A., & Picard, R. W. (2000, 2013). Stress recognition using wearable sensors and mobile phones. <https://doi.org/10.1109/ACII.2013.117>
- Sapolsky, R. M. (2004). *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping*. Holt paperbacks. <https://www.amazon.com/Why-Zebras-Dont-Ulcers-Third/dp/0805073698>

- Sapolsky, R. M. (2007). Stress, Stress-Related Disease, and Emotional Regulation. In *Handbook of emotion regulation*. (pp. 606-615). The Guilford Press. <https://psycnet.apa.org/record/2007-01392-030>
- Sawyer, C. R., & Behnke, R. R. (1999). State anxiety patterns for public speaking and the behavior inhibition system. *Communication Reports*, 12(1), 33-41.
<https://doi.org/10.1080/08934219909367706>
- Schenkler, B. R., & Leary, M. R. (1982). Social Anxiety and Self Presentation: A Conceptualization and Model. *Psychological bulletin*, 92(3), 641-669. <https://doi.org/10.1037/0033-2909.92.3.641>
- Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of voice*, 9(3), 235-248.
[https://doi.org/10.1016/S0892-1997\(05\)80231-0](https://doi.org/10.1016/S0892-1997(05)80231-0)
- Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, 137(3), 137-162.
https://www.researchgate.net/publication/247293598_Psychological_models_of_emotion
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social science information*, 44(4), 695-729. <https://doi.org/10.1177/0539018405058216>
- Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and emotion*, 15, 123-148. <https://doi.org/10.1007/BF00995674>
- Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model. *Psychological bulletin*, 92(3), 641. <https://doi.org/10.1037/0033-2909.92.3.641>
- Schmid Mast, M., Kleinlogel, E. P., Tur, B., & Bachmann, M. (2018). The Future Of Interpersonal Skills Development: Immersive Virtual Reality Training With Virtual Humans. *Human Resource Development Quarterly*, 29(2), 125-141. <https://doi.org/10.1002/hrdq.21307>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Schneider, J., Börner, D., Rosmalen, P. v., & Specht, M. (2015). *Presentation Trainer, your Public Speaking Multimodal Coach* Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA. <https://doi.org/10.1145/2818346.2830603>

- Schreiber, L. (2011). Informative speaking. *The Public Speaking Project (Ed.). Public speaking. The virtual text*, 15-11.
[https://socialsci.libretexts.org/Bookshelves/Communication/Public_Speaking/Public_Speaking_\(The_Public_Speaking_Project\)](https://socialsci.libretexts.org/Bookshelves/Communication/Public_Speaking/Public_Speaking_(The_Public_Speaking_Project))
- Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53(9), 1062-1087. <https://doi.org/10.1016/j.specom.2011.01.011>
- Schuller, B., Rigoll, G., & Lang, M. (2003a). *Hidden Markov Model-based Speech Emotion Recognition* (Vol. 2). <https://doi.org/10.1109/ICME.2003.1220939>
- Schuller, B., Rigoll, G., & Lang, M. (2003b). Hmm-Based Music Retrieval Using Stereophonic Feature Information And Framelength Adaptation. 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698).
<https://doi.org/10.1109/ICME.2003.1221716>
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1), 4-39. <https://doi.org/10.1016/j.csl.2012.02.005>
- Scikit Learn. (2022). *Support Vector Machines*. Retrieved 7/2/2022 from <https://scikit-learn.org/dev/modules/svm.html>
- Selvan, K. T., & Furse, C. M. (2022). Professional Development Ideas for Students and Young Professionals [Young Professionals]. *IEEE Antennas and Propagation Magazine*, 64(5), 122-127. <https://doi.org/10.1109/MAP.2022.3196844>
- Selye, H. (1950). Stress and the General Adaptation Syndrome. *British Medical Journal*, 1(4667), 1383. <https://doi.org/10.1136/bmj.1.4667.1383>
- Selye, H. (2013). *Stress in health and disease*. Butterworth-Heinemann. <https://shop.elsevier.com/books/stress-in-health-and-disease/selye/978-0-407-98510-0>

- Semwal, N., Kumar, A., & Narayanan, S. (2017). *Automatic Speech Emotion Detection System using Multi-domain Acoustic Feature Selection and Classification Models*.
<https://doi.org/10.1109/ISBA.2017.7947681>
- Shabahi, S. (2020). *my-voice-analysis 0.7*. <https://shahabks.github.io/my-voice-analysis/>
- Shouse, E. (2005). Feeling, emotion, affect. *M/c journal*, 8(6). <https://doi.org/10.5204/mcj.2443>
- Shubham Kumar. (2021). *Neural Networks:- What? Why? How?*
<https://sk7653722.medium.com/neural-networks-what-why-how-af1e06827801>
- Slater, M., Pertaub, D.-P., Barker, C., & Clark, D. M. (2006). An Experimental Study On Fear Of Public Speaking Using A Virtual Environment. *CyberPsychology & Behavior*, 9(5), 627-633.
<https://doi.org/10.1089/cpb.2006.9.627>
- Slater, M., Pertaub, D.-P., & Steed, A. (1999). Public Speaking In Virtual Reality: Facing An Audience Of Avatars. *IEEE Computer Graphics and Applications*, 19(2), 6-9.
<https://doi.org/10.1109/38.749116>
- Soltani, K., & Ainon, R. N. (2007). *Speech emotion detection based on neural networks*.
<https://doi.org/10.1109/ISSPA.2007.4555476>
- Somarathna, R., Bednarz, T., & Mohammadi, G. (2022). Virtual reality for emotion elicitation—a review. *IEEE Transactions on Affective Computing*.
<https://doi.org/10.1109/TAFFC.2022.3181053>
- Soyka, F., Leyrer, M., Smallwood, J., Ferguson, C., Riecke, B. E., & Mohler, B. J. (2016, 2016). Enhancing stress management techniques using virtual reality. <https://doi.org/10.1145/2931002.2931017>
- Speer, K. E., Semple, S., Naumovski, N., & McKune, A. J. (2020). Measuring heart rate variability using commercially available devices in healthy children: A validity and reliability study. *European Journal of Investigation in Health, Psychology and Education*, 10(1), 390-404.
<https://doi.org/10.3390/ejihpe10010029>
- Spieler, C., & Miltenberger, R. (2017). Using awareness training to decrease nervous habits during public speaking. *Journal of Applied Behavior Analysis*, 50(1), 38-47.
<https://doi.org/10.1002/jaba.362>

- Stein, M. B., Walker, J. R., & Forde, D. R. (1996). Public-speaking fears in a community sample: Prevalence, impact on functioning, and diagnostic classification. *Archives of general psychiatry*, 53(2), 169-174. <https://doi.org/10.1001/archpsyc.1996.01830020087010>
- Suh, A., & Prophet, J. (2018). The state of immersive technology research: A literature analysis. *Computers in Human behavior*, 86, 77-90. <https://doi.org/10.1016/j.chb.2018.04.019>
- Sülter, R. E., Ketelaar, P. E., & Lange, W.-G. (2022). SpeakApp-Kids! Virtual reality training to reduce fear of public speaking in children—A proof of concept. *Computers & Education*, 178, 104384. <https://doi.org/10.1016/j.compedu.2021.104384>
- Sun, R., & Moore, E. (2011). Investigating glottal parameters and teager energy operators in emotion recognition. International Conference on Affective Computing and Intelligent Interaction. https://doi.org/10.1007/978-3-642-24571-8_54
- Takac, M., Collett, J., Blom, K. J., Conduit, R., Rehm, I., & De Foe, A. (2019). Public speaking anxiety decreases within repeated virtual reality training sessions. *PloS one*, 14(5), e0216288. <https://doi.org/10.1371/journal.pone.0216288>
- Tanveer, M. I., Lin, E., & Hoque, M. (2015). *Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking* Proceedings of the 20th International Conference on Intelligent User Interfaces, Atlanta, Georgia, USA. <https://doi.org/10.1145/2678025.2701386>
- Tardy, C. H., Thompson, W. R., & Allen, M. T. (1989). Cardiovascular Responses during Speech: Does Social Support Mediate the Effects of Talking on Blood Pressure? *Journal of Language and Social Psychology*, 8(3-4), 271-285. <https://doi.org/10.1177/0261927X8983007>
- Tashev, I. J., Wang, Z. Q., Godin, K., & Ieee. (2017). *Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks* 2017 Information Theory and Applications Workshop. <https://doi.org/10.1109/ITA.2017.8023477>
- Tavi, L. (2019). Classifying females' stressed and neutral voices using acoustic–phonetic analysis of vowels: an exploratory investigation with emergency calls. *International journal of speech technology*, 22(3), 511-520. <https://doi.org/10.1007/s10772-018-09574-6>
- Tecles, F., Fuentes-Rubio, M., Tvarijonaviciute, A., Martínez-Subiela, S., Fatjó, J., & Cerón, J. J. (2014). Assessment of stress associated with an oral public speech in veterinary students by

- salivary biomarkers. *Journal of veterinary medical education*, 41(1), 37-43.
<https://doi.org/10.3138/jvme.0513-073R1>
- Teoh, A. N., & Hilmert, C. (2018). Social support as a comfort or an encouragement: A systematic review on the contrasting effects of social support on cardiovascular reactivity. *British Journal of Health Psychology*, 23(4), 1040-1065. <https://doi.org/10.1111/bjhp.12337>
- The EndNote Team. (2013). *EndNote*. In (Version EndNote X9) [64 bit]. Clarivate.
- Thomas, S. A., Lynch, J. J., Friedmann, E., Suginochara, M., Hall, P. S., & Peterson, C. (1984). Blood pressure and heart rate changes in children when they read aloud in school. *Public Health Rep*, 99(1), 77-84. <https://stacks.cdc.gov/view/cdc/66054>
- Thorsteinsson, E. B., & James, J. E. (1999). A Meta-analysis of the effects of experimental manipulations of social support during laboratory stress. *Psychology & Health*, 14(5), 869-886. <https://doi.org/10.1080/08870449908407353>
- Ticknor, B. (2018). Using Virtual Reality to Treat Offenders: An Examination. *International Journal of Criminal Justice Sciences*, 13(2), 316-325. <https://doi.org/110.5281/zenodo.2654383>
- Ticknor, B. (2019). Virtual Reality and Correctional Rehabilitation: A Game Changer. *Criminal Justice and Behavior*, 46(9), 1319-1336. <https://doi.org/10.1177/0093854819842588>
- Titze, I. R., & Martin, D. W. (1998). Principles of voice production. In: Acoustical Society of America. <https://doi.org/10.1121/1.424266>
- Toastmasters International. (2024). *The Toastmasters International*. <https://www.toastmasters.org/>
- Tolkmitt, F. J., & Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3), 302-313. <https://doi.org/10.1037/0096-1523.12.3.302>
- Tom Bäckström, O. R., Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, & Daniel Ramos. (2022). *Introduction to Speech Processing: 2nd Edition*. <https://doi.org/10.5281/zenodo.6821775>
- Tomba, K., Dumoulin, J., Mugellini, E., Abou Khaled, O., & Hawila, S. (2018, 2018). Stress detection through speech analysis. <https://doi.org/10.5220/0006855805600564>

- Truong, K. P., & van Leeuwen, D. A. (2007). Automatic Discrimination Between Laughter And Speech. *Speech communication*, 49(2), 144-158.
<https://doi.org/10.1016/j.specom.2007.01.001>
- Truong, K. P., Van Leeuwen, D. A., & De Jong, F. M. (2012). Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech communication*, 54(9), 1049-1063. <https://doi.org/10.1016/j.specom.2012.04.006>
- Tucker, B., Barton, K., Burger, A., Drye, J., Hunsicker, C., Mendes, A., & LeHew, M. (2019). Exploring public speaking. <https://oer.galileo.usg.edu/communication-textbooks/1/>
- Uchino, B. N., & Garvey, T. S. (1997). The Availability of Social Support Reduces Cardiovascular Reactivity to Acute Psychological Stress. *Journal of Behavioral Medicine*, 20(1), 15-27.
<https://doi.org/10.1023/A:1025583012283>
- Ursin, H., & Eriksen, H. R. (2004). The cognitive activation theory of stress. *Psychoneuroendocrinology*, 29(5), 567-592. [https://doi.org/https://doi.org/10.1016/S0306-4530\(03\)00091-X](https://doi.org/https://doi.org/10.1016/S0306-4530(03)00091-X)
- Ursin, H., & Olff, M. (1993). Psychobiology of coping and defence strategies. *Neuropsychobiology*, 28(1-2), 66-71. <https://doi.org/10.1159/000119002>
- Vallade, J. I., Kaufmann, R., Frisby, B. N., & Martin, J. C. (2021). Technology acceptance model: investigating students' intentions toward adoption of immersive 360 videos for public speaking rehearsals. *Communication Education*, 70(2), 127-145.
<https://doi.org/10.1080/03634523.2020.1791351>
- Valmaggia, L. R., Latif, L., Kempton, M. J., & Rus-Calafell, M. (2016). Virtual reality in the psychological treatment for mental health problems: An systematic review of recent evidence. *Psychiatry Research*, 236, 189-195. <https://doi.org/10.1016/j.psychres.2016.01.015>
- Van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, 14, 62-80. <https://doi.org/10.1016/j.edurev.2015.02.002>

- Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance [Review]. *Frontiers in psychology*, 9. <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.01994>
- VanderPlas, J. (2016). Python Data Science Handbook. In *Python Data Science Handbook*. O'Reilly Media, Inc. https://jakevdp.github.io/PythonDataScienceHandbook/?utm_source=chatgpt.com
- Wallbott, H. G., & Scherer, K. R. (1986). How Universal And Specific Is Emotional Experience? Evidence From 27 Countries On Five Continents. *Information (International Social Science Council)*, 25(4), 763-795. <https://doi.org/10.1177/053901886025004001>
- Walters, L., & Walters, L. (1993). *Secrets of successful speakers: how you can motivate, captivate, and persuade*. McGraw-Hill
- Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech Emotion Recognition Using Fourier Parameters. *IEEE Transactions on Affective Computing*, 6(1), 69-75. <https://doi.org/10.1109/TAFFC.2015.2392101>
- Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, 98(2), 219. <https://doi.org/10.1037/0033-2909.98.2.219>
- Westenberg, P. M., Bokhorst, C. L., Miers, A. C., Sumter, S. R., Kallen, V. L., van Pelt, J., & Blöte, A. W. (2009). A Prepared Speech In Front Of A Pre-Recorded Audience: Subjective, Physiological, And Neuroendocrine Responses To The Leiden Public Speaking Task. *Biological Psychology*, 82(2), 116-124. <https://doi.org/10.1016/j.biopsycho.2009.06.005>
- Whately, R. (1857). *Elements of rhetoric*. James Munroe and Company
- Wiederhold, B. K., & Wiederhold, M. D. (2005). *Virtual reality therapy for anxiety disorders: Advances in evaluation and treatment*. American Psychological Association. <https://doi.org/10.1037/10858-000>
- Wiemers, U. S., Schultheiss, O. C., & Wolf, O. T. (2015). Public speaking in front of an unreceptive audience increases implicit power motivation and its endocrine arousal signature. *Hormones and behavior*, 71, 69-74. <https://doi.org/10.1016/j.yhbeh.2015.04.007>

- Wirtz, P. H., Ehlert, U., Kottwitz, M. U., La Marca, R., & Semmer, N. K. (2013). Occupational Role Stress is Associated With Higher Cortisol Reactivity to Acute Stress. *Journal of Occupational Health Psychology, 18*(2), 121-131. <https://doi.org/10.1037/a0031802>
- Wohlgenannt, I., Simons, A., & Stieglitz, S. (2020). Virtual reality. *Business & Information Systems Engineering, 62*, 455-461. <https://doi.org/10.1007/s12599-020-00658-9>
- Won, E., & Kim, Y.-K. (2016). Stress, the autonomic nervous system, and the immune-kynurenine pathway in the etiology of depression. *Current neuropharmacology, 14*(7), 665-673. <https://doi.org/10.2174/1570159x14666151208113006>
- Woollen, B. (1995). *Vocal tone and experienced stress: A clinical study of vocal timbre*. New School for Social Research. https://books.google.com/books/about/Vocal_Tone_and_Experienced_Stress.html?id=pJDGGwAACAAJ
- Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelbogen, R., & Scherer, S. (2015). *Multimodal Public Speaking Performance Assessment* Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA. <https://doi.org/10.1145/2818346.2820762>
- Xia, Y., & Xu, F. (2022). Study on music emotion recognition based on the machine learning model clustering algorithm. *Mathematical Problems in Engineering, 2022*. <https://doi.org/10.1155/2022/9256586>
- Xie, T., Li, Y., & Tang, Y. (2023). Effects of Using Immersive Virtual Reality for Science Education on Learning Outcomes: A Randomized Controlled Pilot Study. *IEEE Transactions on Learning Technologies, 16*(6), 1045-1056. <https://doi.org/10.1109/TLT.2023.3263587>
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. <https://doi.org/10.1002/cne.920180503>
- Yi-Lin, L., & Gang, W. (2005, 18-21 Aug. 2005). Speech emotion recognition based on HMM and SVM. 2005 International Conference on Machine Learning and Cybernetics. <https://doi.org/10.1109/ICMLC.2005.1527805>

- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., & Povey, D. (2002). The HTK book. *Cambridge university engineering department*, 3(175), 12. <https://www.danielpovey.com/files/htkbook.pdf>
- Yu-Chih, S. (2008). The Toastmasters approach: An innovative way to teach public speaking to EFL learners in Taiwan. *RELC Journal*, 39(1), 113-130.
<https://doi.org/10.1177/0033688208091143>
- Zhai, J., & Barreto, A. (2008). Stress detection in computer users through non-invasive monitoring of physiological signals. *Blood*, 5(0). <https://doi.org/10.1109/IEMBS.2006.259421>
- Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on speech and audio processing*, 9(3), 201-216.
<https://doi.org/10.1109/89.905995>

Appendices

Appendix A

Neutral Positive (NP) condition script

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class ThreeMinPositiveS : MonoBehaviour
{
    Animator anim;
    AudioSource audienceclip;
    public AudioClip[] clip;
    bool M_clap;
    bool M_sitidle1;
    bool M_clap_0;
    bool F_sitidle1_0;
    bool M_clap_1;
    bool F_sitidle1_1;
    bool F_wave;

    // Start is called before the first frame update
    void Start()
    {
        StartCoroutine(posi());
        anim = GetComponent<Animator>();
        audienceclip = GetComponent<AudioSource>();
    }

    IEnumerator posi()
    {
        //neutral behaviour (a little noisy audio, animation - 60 sec.)
        //positive behaviour (6, 9, 12 audios, 3 animations - 120sec)
        //total 171sec

        print("positive behaviour time to start");
        yield return new WaitForSecondsRealtime(2);

        audienceclip.clip = clip[0];
        audienceclip.Play();
        audienceclip.volume = 0.25f;
        yield return new WaitForSecondsRealtime(58);

        anim.SetBool("M_clap", true);
        print("First clap animation 1 set");
    }
}

```

```

audienceclip.clip = clip[1];
audienceclip.Play();
audienceclip.volume = 0.75f;
print("12 sec. applause 1 playing..");
yield return new WaitForSecondsRealtime(12);
anim.SetBool("M_clap", false);

anim.SetBool("M_sitidle1", true);
print("M_sitidle position 1 set");
yield return new WaitForSecondsRealtime(25);
anim.SetBool("M_sitidle1", false);

anim.SetBool("M_clap_0", true);
print("M_clap2_0 animation 2 set");
audienceclip.clip = clip[1];
audienceclip.Play();
audienceclip.volume = 0.75f;
print("12 sec. applause 2 playing.....");
yield return new WaitForSecondsRealtime(12);
anim.SetBool("M_clap_0", false);

anim.SetBool("F_sitidle1_0", true);
print("F_sitidle1_0 sitting position 2 set");
yield return new WaitForSecondsRealtime(25);
anim.SetBool("F_sitidle1_0", false);

anim.SetBool("M_clap_1", true);
print("M_clap3 animation 3 set");
audienceclip.clip = clip[1];
audienceclip.Play();
audienceclip.volume = 0.75f;
print("12 sec. applause 3 playing.....");
yield return new WaitForSecondsRealtime(12);
anim.SetBool("M_clap_1", false);

anim.SetBool("F_sitidle1_1", true);
print("F_sitidle1_0 sitting position 2 set");
yield return new WaitForSecondsRealtime(4);
anim.SetBool("F_sitidle1_1", false);

}
}

```

Neutral-Negative (NN) condition script

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class negativescript : MonoBehaviour

```

```

{

    Animator anim;
    AudioSource audienceclip;
    public AudioClip[] clip;
    bool M_sitidle2;
    bool M_sit_phone1;
    bool M_telloff;
    bool M_sitidle1_0;
    bool M_sitidle2_0;
    bool F_sitidle1_1;

    void Start()
    {
        StartCoroutine(nega());
        anim = GetComponent<Animator>();
        audienceclip = GetComponent<AudioSource>();
    }

    IEnumerator nega()
    {
        //neutral behaviour (a little noisy audio, animation - 60 sec.)
        //positive behaviour (6, 9, 12 audios, 3 animations - 120sec)
        //total 171sec

        print("negative behaviour time to start");
        yield return new WaitForSecondsRealtime(2);

        audienceclip.clip = clip[0];
        audienceclip.Play();
        audienceclip.volume = 0.45f;
        yield return new WaitForSecondsRealtime(57);

        audienceclip.clip = clip[1];
        audienceclip.Play();
        audienceclip.loop = true;
        audienceclip.volume = 0.37f;

        //-ve(1)
        anim.SetBool("M_sit_phone1", true);
        print("M_sit_phone1 set");
        yield return new WaitForSecondsRealtime(20);
        anim.SetBool("M_sit_phone1", false);

        //-ve(2)
        anim.SetBool("M_sit_phone2", true);
        print("M_sit_phone2 set");
        audienceclip.clip = clip[2];
    }
}

```

```

audienceclip.Play();
audienceclip.volume = 0.27f;
yield return new WaitForSecondsRealtime(30);
audienceclip.Stop();
anim.SetBool("M_sit_phone2", false);

// -ve(3): not attentive, sitting animation 2
anim.SetBool("M_sit_phone1", true);
print("M_sit_phone1 set");
audienceclip.clip = clip[1];
audienceclip.Play();
audienceclip.volume = 0.37f;
yield return new WaitForSecondsRealtime(20);
audienceclip.Stop();
anim.SetBool("M_sit_phone1", false);

// -ve(4)
anim.SetBool("M_telloff", true);
print("telloff set");
audienceclip.clip = clip[3];
audienceclip.Play();
audienceclip.volume = 0.55f;
print("crowd angry booing playing");
yield return new WaitForSecondsRealtime(26);
audienceclip.Stop();
anim.SetBool("M_telloff", false);

anim.SetBool("M_sitidle1_1", true);
print("M_sitidle1_1 set");
yield return new WaitForSecondsRealtime(2);
anim.SetBool("M_sitidle1_1", false);

}
}

```

Appendix B *Poster, and Email to JCU staff for recruiting participants and Social Media Post for Experiment 1*

Poster for Experiment 1



**JAMES COOK
UNIVERSITY**
SINGAPORE

You are invited to take part in a research project about the effectiveness of a developed computational algorithm model for detecting stress and improving public speaking performance.

[Ethics Approval No: H8289]

RESEARCH PARTICIPANTS NEEDED!

A Virtual Reality Approach to Improving Public Speaking Skills

Principal Investigator: Arushi

Research Supervisors: Dr. Roberto Dillon & Dr. Ai Ni Teoh

You May Qualify If You Are

- Between 18 - 40 years
- Able to understand and Speak English
- Have no history of epileptic seizures or motion sickness

Duration

- 30 minutes

To Register Scan QR Code



Participation Involves

- Giving a 5-minute speech in virtual reality
- Measurement of heart rate and blood pressure, recording of speech audio
- Filling in two questionnaires

Location

- Psychology Lab (B2-10), James Cook University Singapore, 149 Sims Drive, Singapore 387380

Incentive

- SGD 25 voucher (Grab e-voucher or NTUC FairPrice voucher)

For more info, contact Arushi at - email: arushi@my.jcu.edu.au

Email to JCU lecturers and Staff for E1

Dear Lecturers/ Senior lecturers,

I hope this email finds you well.

My name is Arushi and I'm currently pursuing PhD under supervision of Dr Roberto Dillon and Dr Ai Ni Teoh at JCU Singapore.

I'm in the process of recruiting participants for my research study titled "A VR approach to improving public speaking skills".

The present study seeks to validate the stress arousal capabilities of a developed Virtual Reality (VR) scenario. Participants will give a **5-minute speech in virtual reality**. Blood pressure, heart rate and speech audio will be recorded. There will be questionnaires for participants to complete before and after the speech. The whole experiment will take approximately **30 minutes**. All responses provided by the participants will be kept confidential.

The participants will be given **SGD 10 Grab e- voucher or SGD 10 NTUC FairPrice voucher** as an appreciation for taking part in the study.

I will be really grateful:

- you could help me show the attached poster to students during your class (the poster includes study info and a QR code to register) or
- email the poster to your students or
- allow me to distribute the poster just before or end of the class (if there are physical classes) or
- allow me to advertise it in class for 2-3min online, (if classes happening via Zoom or any other virtual platform), if possible

Whichever you think is convenient.

I very much look forward to your kind support.

Best Regards,

Arushi

Master of Information Technology (majoring in Business Informatics)

PhD (Information Technology) Candidate

Post on social media for E1

Hello everyone!

My name is Arushi, and I'm recruiting participants for my doctoral research study titled "A virtual reality approach to improving public speaking skills".

The present study seeks to evaluate the developed virtual reality application for its capabilities to detect stress from voice and provide feedback in real-time. You are eligible to participate if you are between the ages of 18 and 40; understand and know how to speak English; and have no history of epileptic seizures and motion sickness.

Participants will give a 5-minute speech in virtual reality. Blood pressure, heart rate and speech audio will be recorded. There will be two questionnaires. The total experiment duration is 30 minutes. Your responses will be de-identified and kept confidential.

Participants can choose to receive SGD \$10 Grab e-voucher or SGD \$10 NTUC FairPrice voucher.

If you choose to take part in this research, kindly register your availability via the following link (<https://lnkd.in/gjQxfSZY>) or scanning the QR code from poster attached.

Feel free to contact me if you have any questions. I look forward to having you as a participant in our research.

Thank you!

Sincerely,

Arushi

Ph.D. Candidate (Information Technology)

James Cook University Singapore

Email: arushi@my.jcu.edu.au

Appendix C Scales and Demographic Questionnaires

Stress Arousal Checklist

Instructions: Please answer each question truthfully on how you are **currently** feeling / felt **during the task** by putting a tick in the most appropriate box. There are no right or wrong answers, and no trick questions.

		0. Not at all	1. A Little	2. Quite a bit	3. Very much
1	Calm				
2	Contented				
3	Comfortable				
4	Uneasy				
5	Worried				
6	Distressed				
7	Uptight				
8	Tense				
9	Relaxed				
10	Bothered				
11	Active				
12	Vigorous				
13	Lively				
14	Sleepy				
15	Drowsy				
16	Energetic				
17	Alert				
18	Tired				
19	Passive				
20	Aroused				

Posttest questionnaire: Please answer each question truthfully on how well it describes you. There are no right or wrong answers, and no trick

No.	Items	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
1	During the speech, I was confident							
2	I am interested in improving my public speaking skills							
3	I have attended classes for improving public speaking before							
4	I have good knowledge of public speaking skills							
5	I understand that public speaking is a valuable skill							
6	I understand the influence of public speaking in school, university and work environments							
7	Public speaking reflects good learning capability							
8	I was never taught about improving public speaking							
9	I usually have difficulty in giving a public speech							
10	I have tried virtual reality (VR) before							
11	I have tried VR for public speaking before							
12	During the speech, I was not expecting the sudden responses from the audiences							
13	I felt pleasant about the speech I gave							
14	During the speech, I felt threatened (threat: possible future damage)							
15	During the speech, I felt challenged (challenge: the potential to overcome and even profit)							
16	During the speech, I felt that I could control the situation							
17	During the speech, I felt that I was evaluated by the audience							
18	During the speech, I had sudden feelings of uncertainty							
19	During the speech, I had an intention to avoid embarrassment							
20	During the speech, I desired to finish the task as soon as possible							
21	During the speech, I desired to speak slowly several times							

I put forth effort on the speech task. (Circle one) TRUE, FALSE

I did or did not put forth effort on the task due to:

Experiment 1 Informed Consent Form

This administrative form
has been removed

Experiment 1 Information Sheet

PROJECT TITLE: “A Virtual Reality Approach to Improving Public Speaking Skills”

You are invited to take part in a research project about “A Virtual Reality Approach to Improving Public Speaking Skills”. The purpose of this research is to investigate the effectiveness of a developed computational algorithm model for detecting stress and improving public speaking performance. The research is being conducted by Arushi under the guidance of Dr Roberto Dillon and Dr Ai Ni Teoh will contribute to the degree in Doctor of Philosophy (Information Technology) at James Cook University. This research is funded by James Cook University Australia.

DESCRIPTION OF PROCEDURE

If you agree to be involved in the research, you will be invited to take part in the experiment. During the experiment you will speak in Virtual Reality. With your consent, your speech will be audio taped. Your blood pressure and heart rate will be measured before, during and after your speech. The speech should take total 6 minutes (3 minute). The whole experiment will take approximately 30 minutes and will be conducted in the Psychology Lab at James Cook University Singapore campus. There are two questionnaires that you will need to complete, which ask you about your psychological state before and during the speech and how you feel about the task.

Delivering a speech might have the potential to induce mild psychological distress. Participation is completely voluntary, and participants can withdraw from the study at any time without explanation or prejudice. Counselling services are available at JCU Singapore (Contact number:)

Your responses will be strictly anonymous. The data from the study will be used in research publications and theses. You will not be identified in any way in these publications.

(If you know of others that might be interested in this study, can you please pass on this information sheet to them so they may contact me to volunteer for the study.)

Principal Investigator: Arushi
College: James Cook University
Phone:
Email: Arushi@my.jcu.edu.au

Primary Supervisor: Associate Professor Roberto Dillon
College: James Cook University Singapore
Phone:
Email: Roberto.dillon@my.jcu.edu.au

Secondary Supervisor
Name: Dr Ai Ni Teoh
College: James Cook University Singapore
Phone:
Email: Aini.teoh@my.jcu.edu.au

If you have any concerns regarding the ethical conduct of the study, please contact:
Human Ethics, Research Office
James Cook University, Townsville, Queensland, 4811
Phone: (07) 4781 5011 (ethics@jcu.edu.au)

Appendix D *Devices Used in Experiment 1 and Experiment 2*

I. OMRON Automatic Blood Pressure Monitor HEM-7121



II. Polar Heart rate blood pressure monitor



III. OMEN HP Gaming Laptop



IV. iPhone 7SE for audio voice recordings



V. Virtual Reality HTC Vive



Appendix E *Ethics Approval for Experiment 1*

This administrative form
has been removed

Appendix F *Feature Extraction Code for all the Voice Samples for RAVDESS and VRPS database*

Sample Code snippets are shown for one of the voice samples for all features extracted below:

RAVDESS Database Voice Samples Feature Extraction Code specific to Section 4.1.2

Amplitude Envelope calculation

```
#calculate the amplitude envelope
```

```
FRAME_SIZE = 1024
```

```
HOP_LENGTH = 512
```

```
def amplitude_envelope (signal, frame_size, hop_length):
```

```
    amplitude_envelope = []
```

```
    #calculate AE for each frame
```

```
    for i in range(0, len(signal), hop_length):
```

```
        current_frame_amplitude_envelope = max(signal[i:i+frame_size])
```

```
        amplitude_envelope.append(current_frame_amplitude_envelope)
```

```
    return np.array(amplitude_envelope)
```

```
ae_Actor1_stressed = amplitude_envelope(Actor1_stressed, FRAME_SIZE,
```

```
HOP_LENGTH)
```

```
len (ae_Actor1_stressed)
```

```
print(ae_Actor1_stressed)
```

```
ae_Actor1_stressed.tofile('ae_Actor1_stressed.csv', sep = ',')
```

RMS calculation

```
rms_Actor18_confident = librosa.feature.rms(Actor18_confident, FRAME_SIZE,
```

```
HOP_LENGTH)
```

```
rms_Actor18_confident.tofile('rms_Actor18_confident.csv', sep = ',')
```

```
rms_Actor18_stressed = librosa.feature.rms(Actor18_stressed, FRAME_SIZE,
```

```
HOP_LENGTH)
```

```
rms_Actor18_stressed.tofile('rms_Actor18_stressed.csv', sep = ',')
```

ZCR calculation

```

zcr_Actor23_confident = librosa.feature.zero_crossing_rate(Actor23_confident,
FRAME_SIZE, HOP_LENGTH)
zcr_Actor23_confident.tofile('zcr_Actor23_confident.csv', sep = ',')
zcr_Actor23_stressed = librosa.feature.zero_crossing_rate(Actor23_stressed, FRAME_SIZE,
HOP_LENGTH)
zcr_Actor23_stressed.tofile('zcr_Actor23_stressed.csv', sep = ',')

```

SC calculation

```

sc_Actor1_confident = librosa.feature.spectral_centroid(Actor1_confident, sr=sr,
n_fft=FRAME_SIZE, hop_length=HOP_LENGTH)
sc_Actor1_confident.tofile('sc_Actor1_confident.csv', sep = ',')
sc_Actor1_stressed = librosa.feature.spectral_centroid(Actor1_stressed, sr=sr,
n_fft=FRAME_SIZE, hop_length=HOP_LENGTH)
sc_Actor1_stressed.tofile('sc_Actor1_stressed.csv', sep = ',')

```

SB calculation

```

bw_Actor1_confident= librosa.feature.spectral_bandwidth( Actor1_confident, sr=sr,
n_fft=FRAME_SIZE, hop_length=HOP_LENGTH)
bw_Actor1_confident.tofile('bw_Actor1_confident.csv', sep = ',')

```

MFCCs calculation

```

mfccs_Actor1_confident= librosa.feature.mfcc( Actor1_confident, sr=sr,
n_fft=FRAME_SIZE, hop_length=HOP_LENGTH)
mfccs_Actor1_confident.tofile('mfccs_Actor1_confident.csv', sep = ',')
mfccs_Actor1_stressed= librosa.feature.mfcc( Actor1_stressed, sr=sr, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH)
mfccs_Actor1_stressed.tofile('mfccs_Actor1_stressed.csv', sep = ',')

```

For the VRPS database, the same feature extraction methods and Librosa functions were used as demonstrated above for the RAVDESS database, including RMS, ZCR, SC, SB, SR and MFCCs. However, the frame size and hop length parameters were adjusted to frame size = 24,576 samples, hop length = 8,192 samples. Only the AE and MFCCs extraction methods differed slightly in implementation, as shown on next page for VRPS.

VRPS Database Voice Sample Feature Extraction Code specific to Section 4.2.2Amplitude Envelope

```

# Load the audio file
audio_file_path = "C:/temp/E1/NN/P9NN.wav"
y, sr = librosa.load(audio_file_path)

# Set the desired frame size and hop length
frame_size = 24576 # Size of each analysis frame
hop_length = 8192 # Number of samples between successive frames

# Calculate the amplitude envelope using librosa 's amplitude_to_db function
amplitude_envelope_P9NN = np.abs(librosa.amplitude_to_db(librosa.stft(y,
n_fft=frame_size, hop_length=hop_length), ref=np.max))

# amplitude_envelope is a 2D array where rows represent frequency bins and columns
represent time frames.

# If you want to get the envelope for a specific frequency bin, you can select the
corresponding row.

# For example, to get the envelope for the first frequency bin (index 0):
#envelope_at_first_bin = amplitude_envelope[0, :]
# Now, envelope_at_first_bin contains the AE for the first frequency bin.
# You can also plot the amplitude envelope:
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.imshow(amplitude_envelope_P9NN, aspect='auto', origin='lower', cmap='viridis')
plt.colorbar(label='Amplitude (dB)')
plt.xlabel('Time Frames')
plt.ylabel('Frequency Bins')
plt.title('Amplitude Envelope')
plt.show()

```

```
# Convert the amplitude envelope to a pandas DataFrame
df = pd.DataFrame(amplitude_envelope_P9NN.T) # Transpose the array to have time
frames in rows and frequency bins in columns
```

```
csv_file_path = "C:/temp/E1/Features/AE/AE_P9NN.csv"
df.to_csv(csv_file_path, index=False)
```

Cosine Similarity Score for AE comparison

```
# Resample the longer amplitude envelope to match the length of the shorter one
min._length = min(amplitude_envelope_P5NP.shape[1],
amplitude_envelope_P5NN.shape[1])
amplitude_envelope_P5NP = amplitude_envelope_P5NP[:, :min_length]
amplitude_envelope_P5NN = amplitude_envelope_P5NN[:, :min_length]

distance_score = distance.cosine(amplitude_envelope_P5NP.flatten(),
amplitude_envelope_P5NN.flatten())
print("Cosine Similarity Score between the amplitude envelopes:", 1 - distance_score)
```

MFCCs calculation

Each of the MFCCs were separately extracted for comparison for all actors:

```
import librosa
import librosa.display
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os

FRAME_SIZE = 24576 #24576*2
#HOP_LENGTH = 4096 #8192*0.5
#HOP_LENGTH = 24576
HOP_LENGTH = 8192

P56NP, sr = librosa.load('C:\\temp\\E1\\P56 NP on off.wav')
```

```

P56NPmfcc= librosa.feature.mfcc(P56NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P56NPmfcc = P56NPmfcc.transpose()
P56NPmfcc = P56NPmfcc[:,8:40]

P60NP, sr = librosa.load('C:\\temp\\E1\\P60 NP on off.wav')
P60NPmfcc= librosa.feature.mfcc(P60NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P60NPmfcc = P60NPmfcc.transpose()
P60NPmfcc = P60NPmfcc[:,8:40]

""""

P53NP, sr = librosa.load('C:\\temp\\E1\\P53 NP on off.wav')
P53NPmfcc= librosa.feature.mfcc(P53NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P53NPmfcc = P53NPmfcc.transpose()
P53NPmfcc = P53NPmfcc[:,8:40]

P42NP, sr = librosa.load('C:\\temp\\E1\\P42 NP on off.wav')
P42NPmfcc= librosa.feature.mfcc(P42NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P42NPmfcc = P42NPmfcc.transpose()
P42NPmfcc = P42NPmfcc[:,8:40]

P51NP, sr = librosa.load('C:\\temp\\E1\\P51 NP on off.wav')
P51NPmfcc= librosa.feature.mfcc(P51NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P51NPmfcc = P51NPmfcc.transpose()
P51NPmfcc = P51NPmfcc[:,8:40]

P49NP, sr = librosa.load('C:\\temp\\E1\\P49 NP on off.wav')
P49NPmfcc= librosa.feature.mfcc(P49NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P49NPmfcc = P49NPmfcc.transpose()

```

```
P49NPmfcc = P49NPmfcc[:,8:40]
```

```
P44NP, sr = librosa.load('C:\\temp\\E1\\P44 NP on off.wav')
P44NPmfcc= librosa.feature.mfcc(P44NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P44NPmfcc = P44NPmfcc.transpose()
P44NPmfcc = P44NPmfcc[:,8:40]
```

```
P46NP, sr = librosa.load('C:\\temp\\E1\\P46 NP on off.wav')
P46NPmfcc= librosa.feature.mfcc(P46NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P46NPmfcc = P46NPmfcc.transpose()
P46NPmfcc = P46NPmfcc[:,8:40]
```

```
P45NP, sr = librosa.load('C:\\temp\\E1\\P45 NP on off.wav')
P45NPmfcc= librosa.feature.mfcc(P45NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P45NPmfcc = P45NPmfcc.transpose()
P45NPmfcc = P45NPmfcc[:,8:40]
```

```
P40NP, sr = librosa.load('C:\\temp\\E1\\P40 NP on off.wav')
P40NPmfcc= librosa.feature.mfcc(P40NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P40NPmfcc = P40NPmfcc.transpose()
P40NPmfcc = P40NPmfcc[:,8:40]
```

```
"""
```

```
"""
```

```
P39NP, sr = librosa.load('C:\\temp\\E1\\P39 NP on off.wav')
P39NPmfcc= librosa.feature.mfcc(P39NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P39NPmfcc = P39NPmfcc.transpose()
P39NPmfcc = P39NPmfcc[:,8:40]
```

```
P27NP, sr = librosa.load('C:\\temp\\E1\\P27-NP-ON-oFF.wav')
P27NPmfcc= librosa.feature.mfcc(P27NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P27NPmfcc = P27NPmfcc.transpose()
P27NPmfcc = P27NPmfcc[:,8:40]
```

```
P28NP, sr = librosa.load('C:\\temp\\E1\\P28 Np on off.wav')
P28NPmfcc= librosa.feature.mfcc(P28NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P28NPmfcc = P28NPmfcc.transpose()
P28NPmfcc = P28NPmfcc[:,8:40]
```

```
P37NP, sr = librosa.load('C:\\temp\\E1\\P37 NP on off.wav')
P37NPmfcc= librosa.feature.mfcc(P37NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P37NPmfcc = P37NPmfcc.transpose()
P37NPmfcc = P37NPmfcc[:,8:40]
```

```
P30NP, sr = librosa.load('C:\\temp\\E1\\P30 Np on off.wav')
P30NPmfcc= librosa.feature.mfcc(P30NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P30NPmfcc = P30NPmfcc.transpose()
P30NPmfcc = P30NPmfcc[:,8:40]
```

```
P33NP, sr = librosa.load('C:\\temp\\E1\\P33 Np on off.wav')
P33NPmfcc= librosa.feature.mfcc(P33NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P33NPmfcc = P33NPmfcc.transpose()
P33NPmfcc = P33NPmfcc[:,8:40]
```

```
P32NP, sr = librosa.load('C:\\temp\\E1\\P32 Np on off.wav')
P32NPmfcc= librosa.feature.mfcc(P32NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P32NPmfcc = P32NPmfcc.transpose()
```

```

P32NPmfcc = P32NPmfcc[:,8:40]

P3NP, sr = librosa.load('C:\\temp\\E1\\P3NP.wav')
P3NPmfcc= librosa.feature.mfcc(P3NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P3NPmfcc = P3NPmfcc.transpose()
P3NPmfcc = P3NPmfcc[:,8:40]

P5NP, sr = librosa.load('C:\\temp\\E1\\P5NP.wav')
P5NPmfcc= librosa.feature.mfcc(P5NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P5NPmfcc = P5NPmfcc.transpose()
P5NPmfcc = P5NPmfcc[:,8:40]

P6NP, sr = librosa.load('C:\\temp\\E1\\P6NP.wav')
P6NPmfcc= librosa.feature.mfcc(P6NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P6NPmfcc = P6NPmfcc.transpose()
P6NPmfcc = P6NPmfcc[:,8:40]
print (np.shape(P6NPmfcc))

P8NP, sr = librosa.load('C:\\temp\\E1\\P8NP.wav')
P8NPmfcc= librosa.feature.mfcc(P8NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P8NPmfcc = P8NPmfcc.transpose()
P8NPmfcc = P8NPmfcc[:,8:40]
"""
"""

P17NP, sr = librosa.load('C:\\temp\\E1\\P17NP.wav')
P17NPmfcc= librosa.feature.mfcc(P17NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P17NPmfcc = P17NPmfcc.transpose()
P17NPmfcc = P17NPmfcc[:,8:40]

```



```

P19NP, sr = librosa.load('C:\\temp\\E1\\P19-NP.wav')
P19NPmfcc= librosa.feature.mfcc(P19NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P19NPmfcc = P19NPmfcc.transpose()
P19NPmfcc = P19NPmfcc[:,8:40]

P22NP, sr = librosa.load('C:\\temp\\E1\\P22-NP-em.wav')
P22NPmfcc= librosa.feature.mfcc(P22NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P22NPmfcc = P22NPmfcc.transpose()
P22NPmfcc = P22NPmfcc[:,8:40]

P23NP, sr = librosa.load('C:\\temp\\E1\\P23-Np-on-off.wav')
P23NPmfcc= librosa.feature.mfcc(P23NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P23NPmfcc = P23NPmfcc.transpose()
P23NPmfcc = P23NPmfcc[:,8:40]

P14NP, sr = librosa.load('C:\\temp\\E1\\P14NP.wav')
P14NPmfcc= librosa.feature.mfcc(P14NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P14NPmfcc = P14NPmfcc.transpose()
P14NPmfcc = P14NPmfcc[:,8:40]

P13NP, sr = librosa.load('C:\\temp\\E1\\P13NP.wav')
P13NPmfcc= librosa.feature.mfcc(P13NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P13NPmfcc = P13NPmfcc.transpose()
P13NPmfcc = P13NPmfcc[:,8:40]

P15NP, sr = librosa.load('C:\\temp\\E1\\P15NP.wav')
P15NPmfcc= librosa.feature.mfcc(P15NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P15NPmfcc = P15NPmfcc.transpose()

```

```
P15NPmfcc = P15NPmfcc[:,8:40]
```

```
P12NP, sr = librosa.load('C:\\temp\\E1\\P12NP.wav')
P12NPmfcc= librosa.feature.mfcc(P12NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P12NPmfcc = P12NPmfcc.transpose()
P12NPmfcc = P12NPmfcc[:,8:40]
```

```
P11NP, sr = librosa.load('C:\\temp\\E1\\P11NP.wav')
P11NPmfcc=librosa.feature.mfcc(P11NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P11NPmfcc = P11NPmfcc.transpose()
P11NPmfcc = P11NPmfcc[:,8:40]
```

```
P26NP, sr = librosa.load('C:\\temp\\E1\\P26-NP-ON-oFF.wav')
P26NPmfcc= librosa.feature.mfcc(P26NP, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P26NPmfcc = P26NPmfcc.transpose()
P26NPmfcc = P26NPmfcc[:,8:40]
```

```
""""
```

For labelling the data, following code was run:

```
import numpy as np
P11NPtarget = []
P11NPtarget = np.array([0] * 419)
P11NPtarget = P11NPtarget[:, np.newaxis]
```

```
P3NPtarget = []
P3NPtarget = np.array([0] * 411)
P3NPtarget = P3NPtarget[:, np.newaxis]
```

```
P5NPtarget = []
P5NPtarget = np.array([0] * 433)
P5NPtarget = P5NPtarget[:, np.newaxis]
```

```

P6NPtarget = []
P6NPtarget = np.array([0] * 565)
P6NPtarget = P6NPtarget[:, np.newaxis]

P8NPtarget = []
P8NPtarget = np.array([0] * 388)
P8NPtarget = P8NPtarget[:, np.newaxis]

P37NPtarget = []
P37NPtarget = np.array([0] * 398)
P37NPtarget = P37NPtarget[:, np.newaxis]

P12NPtarget = []
P12NPtarget = np.array([0] * 514)
P12NPtarget = P12NPtarget[:, np.newaxis]

P13NPtarget = []
P13NPtarget = np.array([0] * 490)
P13NPtarget = P13NPtarget[:, np.newaxis]

P15NPtarget = []
P15NPtarget = np.array([0] * 427)
P15NPtarget = P15NPtarget[:, np.newaxis]

P17NPtarget = []
P17NPtarget = np.array([0] * 412)
P17NPtarget = P17NPtarget[:, np.newaxis]

P14NPtarget = []
P14NPtarget = np.array([0] * 446)
P14NPtarget = P14NPtarget[:, np.newaxis]

P19NPtarget = []

```

```
P19NPtarget = np.array([0] * 395)
P19NPtarget = P19NPtarget[:, np.newaxis]
```

```
P22NPtarget = []
P22NPtarget = np.array([0] * 470)
P22NPtarget = P22NPtarget[:, np.newaxis]
```

```
P23NPtarget = []
P23NPtarget = np.array([0] * 397)
P23NPtarget = P23NPtarget[:, np.newaxis]
```

```
P26NPtarget = []
P26NPtarget = np.array([0] * 479)
P26NPtarget = P26NPtarget[:, np.newaxis]
```

```
P27NPtarget = []
P27NPtarget = np.array([0] * 395)
P27NPtarget = P27NPtarget[:, np.newaxis]
```

```
P28NPtarget = []
P28NPtarget = np.array([0] * 436)
P28NPtarget = P28NPtarget[:, np.newaxis]
```

```
P30NPtarget = []
P30NPtarget = np.array([0] * 429)
P30NPtarget = P30NPtarget[:, np.newaxis]
```

```
P32NPtarget = []
P32NPtarget = np.array([1] * 406)
P32NPtarget = P32NPtarget[:, np.newaxis]
```

```
P33NPtarget = []
P33NPtarget = np.array([0] * 447)
P33NPtarget = P33NPtarget[:, np.newaxis]
```

```
P39NPtarget = []
P39NPtarget = np.array([0] * 407)
P39NPtarget = P39NPtarget[:, np.newaxis]
```

```
P40NPtarget = []
P40NPtarget = np.array([0] * 434)
P40NPtarget = P40NPtarget[:, np.newaxis]
```

```
P42NPtarget = []
P42NPtarget = np.array([0] * 415)
P42NPtarget = P42NPtarget[:, np.newaxis]
```

```
P44NPtarget = []
P44NPtarget = np.array([0] * 291)
P44NPtarget = P44NPtarget[:, np.newaxis]
```

```
P45NPtarget = []
P45NPtarget = np.array([0] * 385)
P45NPtarget = P45NPtarget[:, np.newaxis]
```

```
P46NPtarget = []
P46NPtarget = np.array([0] * 420)
P46NPtarget = P46NPtarget[:, np.newaxis]
```

```
P49NPtarget = []
P49NPtarget = np.array([0] * 432)
P49NPtarget = P49NPtarget[:, np.newaxis]
```

```
P51NPtarget = []
P51NPtarget = np.array([0] * 377)
P51NPtarget = P51NPtarget[:, np.newaxis]
```

```
P53NPtarget = []
```

```
P53NPtarget = np.array([0] * 414)
P53NPtarget = P53NPtarget[:, np.newaxis]
```

```
P56NPtarget = []
P56NPtarget = np.array([0] * 366)
P56NPtarget = P56NPtarget[:, np.newaxis]
```

```
P60NPtarget = []
P60NPtarget = np.array([0] * 419)
P60NPtarget = P60NPtarget[:, np.newaxis]
```

Same process was adopted for NN files.

```
# -*- coding: utf-8 -*-
"""
"""

import librosa
import librosa.display
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os

FRAME_SIZE = 24576 #24576*2
#HOP_LENGTH = 4096 #8192*0.5
#HOP_LENGTH = 24576
HOP_LENGTH = 8192

"""

P3NN, sr = librosa.load('C:\\temp\\E1\\P3NN.wav')
P3NNmfcc= librosa.feature.mfcc(P3NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P3NNmfcc = P3NNmfcc.transpose()
P3NNmfcc = P3NNmfcc[:,8:40]
```

```
P5NN, sr = librosa.load('C:\\temp\\E1\\P5NN.wav')
P5NNmfcc= librosa.feature.mfcc(P5NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P5NNmfcc = P5NNmfcc.transpose()
P5NNmfcc = P5NNmfcc[:,8:40]
```

```
P6NN, sr = librosa.load('C:\\temp\\E1\\P6NN.wav')
P6NNmfcc= librosa.feature.mfcc(P6NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P6NNmfcc = P6NNmfcc.transpose()
P6NNmfcc = P6NNmfcc[:,8:40]
```

```
P8NN, sr = librosa.load('C:\\temp\\E1\\P8NN.wav')
P8NNmfcc= librosa.feature.mfcc(P8NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P8NNmfcc = P8NNmfcc.transpose()
P8NNmfcc = P8NNmfcc[:,8:40]
```

```
P11NN, sr = librosa.load('C:\\temp\\E1\\P11NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P11NNmfcc= librosa.feature.mfcc(P11NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P11NNmfcc = P11NNmfcc.transpose()
P11NNmfcc = P11NNmfcc[:,8:40]
```

```
P14NN, sr = librosa.load('C:\\temp\\E1\\P14NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P14NNmfcc = P14NNmfcc.transpose()
P14NNmfcc = P14NNmfcc[:,8:40]
```

```
P13NN, sr = librosa.load('C:\\temp\\E1\\P13NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P13NNmfcc= librosa.feature.mfcc(P13NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P13NNmfcc = P13NNmfcc.transpose()
P13NNmfcc = P13NNmfcc[:,8:40]
```

```
P12NN, sr = librosa.load('C:\\temp\\E1\\P12NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P12NNmfcc= librosa.feature.mfcc(P12NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P12NNmfcc = P12NNmfcc.transpose()
P12NNmfcc = P12NNmfcc[:,8:40]
```

```
P15NN, sr = librosa.load('C:\\temp\\E1\\P15NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P15NNmfcc= librosa.feature.mfcc(P15NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P15NNmfcc = P15NNmfcc.transpose()
P15NNmfcc = P15NNmfcc[:,8:40]
```

```
P17NN, sr = librosa.load('C:\\temp\\E1\\P17NN.wav')
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P17NNmfcc= librosa.feature.mfcc(P17NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P17NNmfcc = P17NNmfcc.transpose()
P17NNmfcc = P17NNmfcc[:,8:40]
```

```
P19NN, sr = librosa.load('C:\\temp\\E1\\P19NN.wav')
```



```
#P14NNmfcc= librosa.feature.mfcc(P14NN, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=20)
P19NNmfcc= librosa.feature.mfcc(P19NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P19NNmfcc = P19NNmfcc.transpose()
P19NNmfcc = P19NNmfcc[:,8:40]
```

```
P22NN, sr = librosa.load('C:\\temp\\E1\\P22-NN-ONOF.wav')
P22NNmfcc= librosa.feature.mfcc(P22NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P22NNmfcc = P22NNmfcc.transpose()
P22NNmfcc = P22NNmfcc[:,8:40]
```

```
P23NN, sr = librosa.load('C:\\temp\\E1\\P23-NN-em.wav')
P23NNmfcc= librosa.feature.mfcc(P23NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P23NNmfcc = P23NNmfcc.transpose()
P23NNmfcc = P23NNmfcc[:,8:40]
```

```
P26NN, sr = librosa.load('C:\\temp\\E1\\P26-NN-emp.wav')
P26NNmfcc= librosa.feature.mfcc(P26NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P26NNmfcc = P26NNmfcc.transpose()
P26NNmfcc = P26NNmfcc[:,8:40]
```

```
P27NN, sr = librosa.load('C:\\temp\\E1\\P27-NN-emp.wav')
P27NNmfcc= librosa.feature.mfcc(P27NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P27NNmfcc = P27NNmfcc.transpose()
P27NNmfcc = P27NNmfcc[:,8:40]
```

```
P28NN, sr = librosa.load('C:\\temp\\E1\\P28 NN emp.wav')
P28NNmfcc= librosa.feature.mfcc(P28NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
```

```
P28NNmfcc = P28NNmfcc.transpose()
P28NNmfcc = P28NNmfcc[:,8:40]
```

```
P30NN, sr = librosa.load('C:\\temp\\E1\\P30 NN Emp.wav')
P30NNmfcc= librosa.feature.mfcc(P30NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P30NNmfcc = P30NNmfcc.transpose()
P30NNmfcc = P30NNmfcc[:,8:40]
```

```
P32NN, sr = librosa.load('C:\\temp\\E1\\P32 NN emp.wav')
P32NNmfcc= librosa.feature.mfcc(P32NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P32NNmfcc = P32NNmfcc.transpose()
P32NNmfcc = P32NNmfcc[:,8:40]
```

```
P33NN, sr = librosa.load('C:\\temp\\E1\\P33 NN Emp.wav')
P33NNmfcc= librosa.feature.mfcc(P33NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P33NNmfcc = P33NNmfcc.transpose()
P33NNmfcc = P33NNmfcc[:,8:40]
```

```
P37NN, sr = librosa.load('C:\\temp\\E1\\P37 NN emp.wav')
P37NNmfcc= librosa.feature.mfcc(P37NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P37NNmfcc = P37NNmfcc.transpose()
P37NNmfcc = P37NNmfcc[:,8:40]
```

```
P39NN, sr = librosa.load('C:\\temp\\E1\\P39 NN Emp.wav')
P39NNmfcc= librosa.feature.mfcc(P39NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P39NNmfcc = P39NNmfcc.transpose()
P39NNmfcc = P39NNmfcc[:,8:40]
```

```
P40NN, sr = librosa.load('C:\\temp\\E1\\P40 NN Emp.wav')
```

```
P40NNmfcc= librosa.feature.mfcc(P40NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P40NNmfcc = P40NNmfcc.transpose()
P40NNmfcc = P40NNmfcc[:,8:40]
```

```
P42NN, sr = librosa.load('C:\\temp\\E1\\P42 NN emp.wav')
P42NNmfcc= librosa.feature.mfcc(P42NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P42NNmfcc = P42NNmfcc.transpose()
P42NNmfcc = P42NNmfcc[:,8:40]
```

```
P44NN, sr = librosa.load('C:\\temp\\E1\\P44 NN emp.wav')
P44NNmfcc= librosa.feature.mfcc(P44NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P44NNmfcc = P44NNmfcc.transpose()
P44NNmfcc = P44NNmfcc[:,8:40]
```

```
P45NN, sr = librosa.load('C:\\temp\\E1\\P45 NN Emp.wav')
P45NNmfcc= librosa.feature.mfcc(P45NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P45NNmfcc = P45NNmfcc.transpose()
P45NNmfcc = P45NNmfcc[:,8:40]
```

```
P46NN, sr = librosa.load('C:\\temp\\E1\\P46 NN Emp.wav')
P46NNmfcc= librosa.feature.mfcc(P46NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P46NNmfcc = P46NNmfcc.transpose()
P46NNmfcc = P46NNmfcc[:,8:40]
```

""""

```
P51NN, sr = librosa.load('C:\\temp\\E1\\P51 NN Emp.wav')
P51NNmfcc= librosa.feature.mfcc(P51NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
```

```
P51NNmfcc = P51NNmfcc.transpose()
P51NNmfcc = P51NNmfcc[:,8:40]
```

```
P53NN, sr = librosa.load('C:\\temp\\E1\\P53 NN Emp.wav')
P53NNmfcc= librosa.feature.mfcc(P53NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P53NNmfcc = P53NNmfcc.transpose()
P53NNmfcc = P53NNmfcc[:,8:40]
```

```
P56NN, sr = librosa.load('C:\\temp\\E1\\P56 NN Emp.wav')
P56NNmfcc= librosa.feature.mfcc(P56NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P56NNmfcc = P56NNmfcc.transpose()
P56NNmfcc = P56NNmfcc[:,8:40]
```

```
P60NN, sr = librosa.load('C:\\temp\\E1\\P60 NN Emp.wav')
P60NNmfcc= librosa.feature.mfcc(P60NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P60NNmfcc = P60NNmfcc.transpose()
P60NNmfcc = P60NNmfcc[:,8:40]
```

```
P49NN, sr = librosa.load('C:\\temp\\E1\\P49 NN Emp.wav')
P49NNmfcc= librosa.feature.mfcc(P49NN, sr=8192,
n_fft=FRAME_SIZE,hop_length=HOP_LENGTH, n_mfcc=40)
P49NNmfcc = P49NNmfcc.transpose()
P49NNmfcc = P49NNmfcc[:,8:40]
```

Appendix G *SDM code*

```
@author: singh
"""

import tensorflow
from tensorflow import keras
from keras import layers
from keras import activations
from keras.models import load_model
import numpy as np
import NP
import NN
import NPtarget
import NNtarget
from keras.models import Sequential
from keras.layers import Dense
import pickle

# define the model
model = Sequential()
model.add(Dense(24, input_dim=32, activation='relu'))
model.add(Dense(12, activation='relu'))
model.add(Dense(6, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

# compile the model
model.compile(loss='binary_crossentropy', optimiser='adam', metrics=['accuracy'])
#model.compile(loss='mean_squared_error', optimiser='adam', metrics=['accuracy'])

# fit the model to the training data
X_train = np.r_[NN.P26NNmfcc, NN.P27NNmfcc, NN.P28NNmfcc,
NN.P30NNmfcc, NN.P32NNmfcc, NN.P33NNmfcc, NN.P37NNmfcc, NN.P39NNmfcc,
NN.P40NNmfcc, NN.P42NNmfcc, NN.P44NNmfcc, NN.P45NNmfcc, NN.P46NNmfcc,
NN.P49NNmfcc, NN.P51NNmfcc, NN.P53NNmfcc, NN.P56NNmfcc, NN.P60NNmfcc,
NP.P26NPmfcc, NP.P27NPmfcc, NP.P28NPmfcc, NP.P30NPmfcc, NP.P32NPmfcc,
```

```

NP.P33NPmfcc, NP.P37NPmfcc, NP.P39NPmfcc, NP.P40NPmfcc, NP.P42NPmfcc,
NP.P44NPmfcc, NP.P45NPmfcc, NP.P46NPmfcc, NP.P49NPmfcc, NP.P51NPmfcc,
NP.P53NPmfcc, NP.P56NPmfcc, NP.P60NPmfcc]

y_train = np.r_[NNtarget.P26NNtarget, NNtarget.P27NNtarget,
NNtarget.P28NNtarget, NNtarget.P30NNtarget, NNtarget.P32NNtarget,
NNtarget.P33NNtarget, NNtarget.P37NNtarget, NNtarget.P39NNtarget,
NNtarget.P40NNtarget, NNtarget.P42NNtarget, NNtarget.P44NNtarget,
NNtarget.P45NNtarget, NNtarget.P46NNtarget, NNtarget.P49NNtarget,
NNtarget.P51NNtarget, NNtarget.P53NNtarget, NNtarget.P56NNtarget,
NNtarget.P60NNtarget, NPtarget.P26NPtarget, NPtarget.P27NPtarget, NPtarget.P28NPtarget,
NPtarget.P30NPtarget, NPtarget.P32NPtarget, NPtarget.P33NPtarget, NPtarget.P37NPtarget,
NPtarget.P39NPtarget, NPtarget.P40NPtarget, NPtarget.P42NPtarget, NPtarget.P44NPtarget,
NPtarget.P45NPtarget, NPtarget.P46NPtarget, NPtarget.P49NPtarget, NPtarget.P51NPtarget,
NPtarget.P53NPtarget, NPtarget.P56NPtarget, NPtarget.P60NPtarget]

model.fit(X_train, y_train, epochs=100, batch_size=32)

#pickle.dump(model, open('kerasModel.pkl', 'wb'))
model.save_weights('model3.h5')
model.save("fullmodel3.h5")

```

NN target samples training

```

# -*- coding: utf-8 -*-
"""
Created on Tue. Jan 3 12:48:23 2023

@author: singh
"""
import numpy as np

"""
P3NNtarget = []

```

```
P3NNtarget = np.array([1] * 423)
P3NNtarget = P3NNtarget[:, np.newaxis]
```

```
P5NNtarget = []
P5NNtarget = np.array([1] * 819)
P5NNtarget = P5NNtarget[:, np.newaxis]
```

```
P6NNtarget = []
P6NNtarget = np.array([1] * 432)
P6NNtarget = P6NNtarget[:, np.newaxis]
```

```
P8NNtarget = []
P8NNtarget = np.array([1] * 389)
P8NNtarget = P8NNtarget[:, np.newaxis]
```

```
P11NNtarget = []
P11NNtarget = np.array([1] * 434)
P11NNtarget = P11NNtarget[:, np.newaxis]
```

```
P12NNtarget = []
P12NNtarget = np.array([1] * 514)
P12NNtarget = P12NNtarget[:, np.newaxis]
```

```
P13NNtarget = []
P13NNtarget = np.array([1] * 353)
P13NNtarget = P13NNtarget[:, np.newaxis]
```

```
P14NNtarget = []
P14NNtarget = np.array([1] * 422)
P14NNtarget = P14NNtarget[:, np.newaxis]
```

```
P15NNtarget = []
P15NNtarget = np.array([1] * 404)
P15NNtarget = P15NNtarget[:, np.newaxis]
```

```
P17NNtarget = []  
P17NNtarget = np.array([1] * 434)  
P17NNtarget = P17NNtarget[:, np.newaxis]
```

```
P19NNtarget = []  
P19NNtarget = np.array([1] * 407)  
P19NNtarget = P19NNtarget[:, np.newaxis]
```

```
P22NNtarget = []  
P22NNtarget = np.array([1] * 392)  
P22NNtarget = P22NNtarget[:, np.newaxis]
```

```
P23NNtarget = []  
P23NNtarget = np.array([1] * 408)  
P23NNtarget = P23NNtarget[:, np.newaxis]
```

```
P26NNtarget = []  
P26NNtarget = np.array([1] * 414)  
P26NNtarget = P26NNtarget[:, np.newaxis]
```

```
P27NNtarget = []  
P27NNtarget = np.array([1] * 388)  
P27NNtarget = P27NNtarget[:, np.newaxis]
```

```
P28NNtarget = []  
P28NNtarget = np.array([1] * 512)  
P28NNtarget = P28NNtarget[:, np.newaxis]
```

```
P30NNtarget = []  
P30NNtarget = np.array([1] * 427)
```



```
P30NNtarget = P30NNtarget[:, np.newaxis]
```

```
P32NNtarget = []
```

```
P32NNtarget = np.array([1] * 388)
```

```
P32NNtarget = P32NNtarget[:, np.newaxis]
```

```
P33NNtarget = []
```

```
P33NNtarget = np.array([1] * 420)
```

```
P33NNtarget = P33NNtarget[:, np.newaxis]
```

```
P37NNtarget = []
```

```
P37NNtarget = np.array([1] * 428)
```

```
P37NNtarget = P37NNtarget[:, np.newaxis]
```

```
P39NNtarget = []
```

```
P39NNtarget = np.array([1] * 390)
```

```
P39NNtarget = P39NNtarget[:, np.newaxis]
```

```
P40NNtarget = []
```

```
P40NNtarget = np.array([1] * 440)
```

```
P40NNtarget = P40NNtarget[:, np.newaxis]
```

```
P42NNtarget = []
```

```
P42NNtarget = np.array([1] * 372)
```

```
P42NNtarget = P42NNtarget[:, np.newaxis]
```

```
P44NNtarget = []
```

```
P44NNtarget = np.array([1] * 415)
```

```
P44NNtarget = P44NNtarget[:, np.newaxis]
```

```
P45NNtarget = []
```

```
P45NNtarget = np.array([1] * 396)
```

```
P45NNtarget = P45NNtarget[:, np.newaxis]
```

```
P46NNtarget = []
P46NNtarget = np.array([1] * 409)
P46NNtarget = P46NNtarget[:, np.newaxis]
```

```
"""
```

```
P49NNtarget = []
P49NNtarget = np.array([1] * 569)
P49NNtarget = P49NNtarget[:, np.newaxis]
```

```
P51NNtarget = []
P51NNtarget = np.array([1] * 422)
P51NNtarget = P51NNtarget[:, np.newaxis]
```

```
P53NNtarget = []
P53NNtarget = np.array([1] * 345)
P53NNtarget = P53NNtarget[:, np.newaxis]
```

```
P56NNtarget = []
P56NNtarget = np.array([1] * 378)
P56NNtarget = P56NNtarget[:, np.newaxis]
```

```
P60NNtarget = []
P60NNtarget = np.array([1] * 398)
P60NNtarget = P60NNtarget[:, np.newaxis]
```

Appendix H *Count Down Timer C# Code*

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using UnityEngine.UI;

public class Timer : MonoBehaviour
{
    public float timeValue = 90;
    public Text timerText;

    // Start is called before the first frame update
    void Start()
    {

    }

    // Update is called once per frame
    void Update()
    {
        if (timeValue > 0)
        {
            timeValue -= Time.deltaTime;
        }

        else
        {
            timeValue = 0;
        }

        DisplayTime(timeValue);
    }

    void DisplayTime (float timeToDisplay)
    {

```

```

        if (timeToDisplay < 0)
        {
            timeToDisplay = 0;
        }

        float minutes = Mathf.FloorToInt(timeToDisplay / 60);
        float seconds = Mathf.FloorToInt(timeToDisplay % 60);

        timerText.text = string.Format("{0:00}:{1:00}", minutes, seconds);
    }
}

```

To begin with the timer code, `System.Collections` and `System.Collections.Generic` are namespaces that contain classes and interfaces for working with collections of objects, such as lists and arrays. `UnityEngine` is the core namespace for Unity, containing fundamental classes for working with game objects, components and more. `UnityEngine.UI` is specifically for working with UI components in Unity, such as `Text`, `Button` and others. The `Timer` class is derived from `MonoBehaviour`, the base class from which every Unity script inherits. This class allows the script to be attached to game objects in Unity and provides fundamental game object behaviours.

The `timeValue` is a float variable that stores the amount of time in seconds, initialised to 90. The `timerText` is a reference to a `Text` UI element in the Unity scene. The `Start()` method is a Unity-specific method called before the first frame update, often used to initialise variables or game states. In this script, `Start()` is currently empty. The `Update()` method is called once per frame. If `timeValue > 0`, it decreases `timeValue` by the time elapsed since the last frame (`Time.deltaTime`). If `timeValue` reaches 0, it's set to zero to stop the timer. The `DisplayTime(timeValue)` method is called to update the UI text with the current time. This

method formats and displays the time in the "MM" format on the UI. The timerText.text property then updates the Text component in the UI with the formatted time string.

Appendix I *Real-time testing Code of the SDM*

```
import librosa
import numpy as np
import librosa.display
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import time
from keras.models import load_model

#access the audio wav file directory
dir_path = r'C:\\Users\\singh\\PythonModelDev\\MLP'

def loadfilenames():
    filelist = []
    for path in os.listdir(dir_path):
        filepath = os.path.join(dir_path, path)
        if os.path.isfile(filepath) and filepath.endswith('.wav'):
            filelist.append(filepath)
    filelist.sort(key=lambda x:os.path.getmtime(x), reverse=True)
    return filelist

#in the beginning raw data is not processed
processedfile = None

#3 sec. identify stress
FRAME_SIZE = 24576
HOP_LENGTH = 8192
```

```

print('loading model.....')
model = load_model('C:/temp/fullmodel3.h5')
print ('model loaded.....')

while True:
    files = loadfilenames()
    #all data which comes here will go through processing, the initial condition is always true
    if processedfile != files[0]:
        processedfile = files[0]
        head, tail = os.path.split(processedfile)

        print ('filename:', tail)
        file, sr = librosa.load(processedfile, 8192)
        filemfcc = librosa.feature.mfcc(file, sr=8192, n_fft=FRAME_SIZE,
hop_length=HOP_LENGTH, n_mfcc=40)
        filemfcc = filemfcc.transpose()
        filemfcc = filemfcc[:,8:40]

        predictions = model.predict(filemfcc)
        print(predictions)

        threshold = 0.5
        count = np.sum(predictions > threshold)

        print (count)
        f = open("score.txt", "w")
        f.write(str(count))
        f.close()

    else:
        time.sleep(2)
        print('latest file not updated')

```

The live implementation of the SDM in this script begins by importing various libraries, including librosa, a Python library for analysing and processing audio signals. librosa provides functions for extracting features from audio files, such as MFCC. Additional libraries imported include numpy for numerical computations, matplotlib for creating visual graphs, pandas for data manipulation and analysis, os for interacting with the operating system (e.g., reading and writing to the filesystem), time for handling real-time operations, and keras.models.load_model for loading a pretrained Keras model.

The script accesses and processes audio files, likely .wav format, stored in a specified directory (dir_path). The loadfilenames() function lists all the files in this directory, checks for new .wav files, and stores them in filelist, which is then sorted by modification time, with the most recent file first. The variable processedfile, initially set to None, will later store the path of the most recently processed audio file. FRAME_SIZE and HOP_LENGTH are parameters used for audio feature extraction, determining the number of samples per frame and the number of samples between successive frames. The load_model function loads a pretrained SDM (fullmodel3.h5) from a specified file path.

After loading the SDM, the script enters an infinite loop that continuously checks for new audio files using loadfilenames(). If the most recent file has not yet been processed, it is loaded and processed. librosa loads the audio file at a sample rate of 8192 Hz, and librosa.feature.mfcc() extracts MFCC features from the audio data. These features are then passed to the loaded ML model to generate predictions. numpy is then used to count how many of the model's predictions exceed a specified threshold (0.5 in this case). The prediction results are written to a score.txt file. If the latest file has already been processed, the script waits for 2 seconds before checking for new files and repeating the process.

Appendix J *Real-time Audio Recording Using Scipy*

```

# Recording properties
SAMPLE_RATE = 8192
SECONDS = 10

# Channels
MONO = 1
STEREO = 2
time = 60

# Command to get all devices listed: py -m sounddevice
# Device you want to record

REC.default.device = 'Microphone Array (Intel® Smart Sound Technology (Intel® SST)),
Windows DirectSound'
print(f'Recording for {SECONDS} seconds')

counter = 0
while counter < time:

# Starts recording

    recording = REC.rec( int(SECONDS * SAMPLE_RATE), samplerate = SAMPLE_RATE,
channels = MONO)
    REC.wait() # Waits for recording to finish

# Writes recorded data in to the wave file

    filename = 'recording' + str(counter) + ".wav"
    print(f'Just recorded file {filename}')
    #"C:\Users\singh\PythonModelDev\MLP\recording.wav"
    write(filename, SAMPLE_RATE, recording)
    counter = counter + 1

```


This live audio recording script begins by importing necessary libraries, including `sounddevice` for recording and playing back audio in Python, and `scipy`, specifically its `wavfile` module, for writing recorded audio data to `.wav` files.

The recording properties are defined with a sample rate of 8192 Hz and a duration of 10 seconds for each audio clip. Constants are set for mono and stereo channels, though mono (single-channel audio) is used in this script. The recording device is specified as a particular microphone connected to the system.

The script then enters a loop, set to run for 60 seconds, where it repeatedly records 10-second audio segments. For each iteration, the recorded audio is stored in an array, and once recording is complete, the data is written to a `.wav` file, named uniquely using a counter. The loop continues until the total recording time reaches 60 seconds, incrementing the counter after each recording to ensure all files are saved with distinct names. The script efficiently automates the process of recording multiple short audio clips and saving them to disc in a structured manner.

Appendix K *Score Adjustment C# Script*

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using UnityEngine.UI;
using System;
using System.IO;

public class Score : MonoBehaviour
{
    public static int score, oldscore;
    public Text ScoreText;
    int step;
    int n;
    int c;
    // Start is called before the first frame update
    void Start()
    {
        Reset();
        InvokeRepeating("CheckFile", 1f, 10f);
    }
    private void Reset()
    {
        score = 0;
        oldscore = 0;
        step = 10; //10;
        c = 10; //5
        n = 5;
    }

    // Update is called once per frame

```

```

void CheckFile()
{
    int newscore;
    string tScore =
File.ReadAllText(@"C:\Users\singh\PythonModelDev\MLP\score.txt");
    Debug.Log("this is score from score file");
    Debug.Log(tScore);
    newscore = (int)float.Parse(tScore);
    //int.TryParse(tScore, out newscore);
    Debug.Log("this is newscore");
    Debug.Log(newscore);
    if (newscore > n)
        score -= step;
    else
        score += step;
    if (newscore >= oldscore && newscore > n)
        score -= c;
    else if (newscore <= oldscore && newscore <= n)
        score += c;
    score += (int) newscore;
    //set boundaries
    if (score >= 100)
        score = 100;
    else if (score <= -100)
        score = -100;
    ScoreText.text = score.ToString();
    //Debug.Log(score);

    PlayerPrefs.SetInt("score", score);
    oldscore = newscore;
}
}

```

The score adjustment Unity script is designed to manage and update a score based on input from an external text file and set the boundaries and range of the score. The script begins by importing necessary namespaces, including those for handling collections, working with Unity's game objects and UI components, and reading data from files. The Score class, which inherits from MonoBehaviour, includes several key variables: score and oldscore to track the current and previous scores, and a Text UI element reference, ScoreText, to display the score in the game. The script initialises these variables in the Reset method, where it sets score and oldscore to zero and assigns values to the step, n and c variables, which are later used to adjust the score.

Upon starting, the script calls the Reset method and sets up a recurring invocation of the CheckFile method using InvokeRepeating, which begins one second after the script starts and repeats every 10 seconds. The CheckFile method is the core of the script, where it reads a score from a specified text file; which is basically the prediction values from the SDM, parses it into an integer (newscore), and uses it to update the game score. The method applies several conditions: if the new score exceeds a threshold (n), the step value is either added or subtracted from the score. Additional adjustments are made based on the relationship between the current and previous scores. The final score is constrained within a range of -100 to 100, ensuring it doesn't exceed these boundaries. This updated score is displayed on the UI via the ScoreText component and is saved using Unity's PlayerPrefs, which persists the score across game sessions.

Appendix L *SF Score Adjustment C# Code*

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using UnityEngine.UI;
using System;
using System.IO;

public class SocialSupportScore : MonoBehaviour
{
    public static int score, oldscore;
    public Text ScoreText;
    //public Text TextScore;
    int step;
    int n;
    int c;
    int help_factor;

    // Start is called before the first frame update
    void Start()
    {
        Reset();
        InvokeRepeating("CheckFile", 1f, 10f);
    }

    private void Reset()
    {
        score = 0;
        oldscore = 0;
        step = 10; //10;
        c = 5;
        n = 5;
    }

```

```

        help_factor = 2;
    }

    // Update is called once per frame
    void CheckFile()
    {
        int newscore;
        string tScore
File.ReadAllText(@"C:\Users\singh\PythonModelDev\MLP\score.txt");

        Debug.Log("this is score from score file");
        Debug.Log(tScore);

        //newscore = int.Parse(tScore);
        newscore = (int)float.Parse(tScore);

        Debug.Log("this is newscore");
        Debug.Log(newscore);

        if (newscore >= n)
            score -= step;
        else
            score += step + help_factor;

        if (newscore >= oldscore)
            score -= c;
        else if (newscore <= oldscore)
            score += c + help_factor;

        // score += (int) newscore;
        //set boundaries
        if (score >= 100)
            score = 100;
        else if (score <= -100)

```

```

        score = -100;

        ScoreText.text = score.ToString();
        //TextScore.text = score.ToString();
        //Debug.Log(score);

        PlayerPrefs.SetInt("score", score);
        oldscore = newscore;
    }
}

```

This SF score adjustment C# script is part of a scoring system for a "Supportive Feedback" mode. The script, named `SocialSupportScore`, dynamically adjusts the player's score based on real-time data read from the Score Adjustment File which is based on the predictions of stress from the SDM. The scoring mechanism for this mode is influenced by several variables, including step, c, n, and specifically by `help_factor`. The help factor variable was introduced to allow flexibility and responsiveness in how the score is calculated for this supportive mode. The score is updated at regular intervals using the `CheckFile` method, which reads the score value from a specified file path, compares it with previous values, and adjusts the current score accordingly. The resulting score is then reflected in the SF's UI through a text element, providing immediate feedback to the player. The script ensures the score persists between game sessions by saving it with Unity's `PlayerPrefs`. The dynamic and adaptive scoring system enhances the player's experience by responding to real-time data and updating the game's feedback mechanisms accordingly for this social supportive mode which is directed to help the user with stress reduction.

Appendix M *SF and CF Modes Scripts**SS mode Script*

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class SocialSupportive : MonoBehaviour
{
    static int newscore;
    public Animator anim;
    AudioSource audienceclip;
    public AudioClip[] clip;
    public GameObject SoundManager;

    // Start is called before the first frame update
    void Start()
    {
        newscore = 0;
        anim = GetComponent<Animator>();
        audienceclip = GetComponent<AudioSource>();
    }
    // Update is called once per frame
    void Update()
    {
        int t1 = 0;
        int t2 = -79;
        int t3 = -80;
        int t4 = -100;
        newscore = PlayerPrefs.GetInt("score");
        if (anim != null)
        {
            //SoundManager.GetComponent<AudioScript>().PlayAudio(1);

```



```

if (newscore >= t1)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_sitidle1"))
    {
        anim.SetTrigger("M_sitidle1");
        print("M-sitidle1 attentive animation 1 set");
        SoundManager.GetComponent<AudioScript>().PlayAudio(1);
        audienceclip.clip = clip[1];
        audienceclip.Play();
    }
}

if (newscore < 0 && newscore >= t2)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("F_sitidle1"))
    {
        anim.SetTrigger("F_sitidle1");
        print("F_sitidle1 attentive position 2 set");
        SoundManager.GetComponent<AudioScript>().PlayAudio(2);
        audienceclip.clip = clip[2];
        audienceclip.Play();
    }
}

if (newscore <= t3 && newscore >= -100)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_clap"))
    {
        anim.SetTrigger("M_clap");
        print("M_clap set");

        SoundManager.GetComponent<AudioScript>().PlayAudio(0);
        audienceclip.clip = clip[0];
        audienceclip.Play();
    }
}

```

```

    }
}
}
}
}

```

CF Mode Script

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class ScoreBehaviourScript : MonoBehaviour
{
    static int newscore;
    public Animator anim;
    AudioSource audienceclip;
    public AudioClip[] clip;
    public GameObject SoundManager;
    // Start is called before the first frame update
    void Start()
    {
        newscore = 0;
        anim = GetComponent<Animator>();
        audienceclip = GetComponent<AudioSource>();

    }
    // Update is called once per frame
    void Update()
    {
        int t1 = 100;
        int t2 = 65;
    }
}

```

```

int t3 = 0;
int t4 = -75;
int t5 = -80;
int t6 = -100;
newscore = PlayerPrefs.GetInt("score");
if (anim != null)
{
    //SoundManager.GetComponent<AudioScript>().PlayAudio(1);

    if (newscore <= t1 && newscore >= t2)
    {

        if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_clap"))
        {
            //newscore += 10;
            anim.SetTrigger("M_clap");
            print("First clap animation 1 set");

            SoundManager.GetComponent<AudioScript>().PlayAudio(0);
            audienceclip.clip = clip[0];
            audienceclip.Play();

            //print("12 sec. applause 1 playing..");
            //Debug.Log(newscore);

        }
    }

    else if (newscore < t2 && newscore >= t3)
    {
        if (!anim.GetCurrentAnimatorStateInfo(0).IsName("F_sitidle1"))
        {
            newscore -= 10;

```

```

        anim.SetTrigger("F_sitidle1");
        print("F_sitidle position 1 set");
        //yield return new WaitForSecondsRealtime(25);
        //anim.SetTrigger("F_sitidle1");
        SoundManager.GetComponent<AudioScript>().PlayAudio(1);

        audienceclip.clip = clip[1];
        audienceclip.Play();

    }

}

if (newscore < t3 && newscore >= t4)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_sit_phone1"))
    {
        //-ve(1)
        anim.SetTrigger("M_sit_phone1");
        print("M_sit_phone1 set");

        SoundManager.GetComponent<AudioScript>().PlayAudio(2);

        audienceclip.clip = clip[4];
        audienceclip.Play();
        audienceclip.volume = 0.27f;
        //yield return new WaitForSecondsRealtime(30);
        audienceclip.Stop();
        //anim.SetTrigger("M_sit_phone1");
    }
}

if (newscore < t4 && newscore >= t5)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_sit_phone2"))

```

```

{
    //-ve(2): not attentive, sitting animation 2
    anim.SetTrigger("M_sit_phone2");
    print("M_sit_phone2 set");
    SoundManager.GetComponent<AudioScript>().PlayAudio(3);
    audienceclip.clip = clip[3];
    audienceclip.Play();
    audienceclip.volume = 0.37f;
    //yield return new WaitForSecondsRealtime(20);
    audienceclip.Stop();
    //anim.SetTrigger("M_sit_phone2");
}
}
if (newscore<=t6)
{
    if (!anim.GetCurrentAnimatorStateInfo(0).IsName("M_telloff"))
    {
        //-ve(3)
        anim.SetTrigger("M_telloff");
        print("telloff set");
        SoundManager.GetComponent<AudioScript>().PlayAudio(4);
        audienceclip.clip = clip[3];
        audienceclip.Play();
        audienceclip.volume = 0.55f;
        //print("crowd angry booing playing");
        //yield return new WaitForSecondsRealtime(26);
        audienceclip.Stop();
        //anim.SetTrigger("M_telloff");
        //anim.SetTrigger("M_telloff");
    }
}
}
}
}
}

```

Appendix N *Experiment 2 Poster, Email to Lecturers and Social Media Post*

Experiment 2 Poster



**JAMES COOK
UNIVERSITY
SINGAPORE**

You are invited to take part in a research project about the effectiveness of a developed computational algorithm model for detecting stress and improving public speaking performance.

[Ethics Approval No: H8939]

RESEARCH PARTICIPANTS NEEDED!

A Virtual Reality Approach to Improving Public Speaking Skills

Principal Investigator: Arushi

Research Supervisors: Dr. Roberto Dillon & Dr. Ai Ni Teoh

You May Qualify If You Are

- Between 18 - 40 years
- Able to understand and Speak English
- Have no history of epileptic seizures or motion sickness

Duration

- 30 minutes

To Register Scan QR Code



Participation Involves

- Giving a 5-minute speech in virtual reality
- Measurement of heart rate and blood pressure, recording of speech audio
- Filling in two questionnaires

Location

- Psychology Lab (B2-10), James Cook University Singapore, 149 Sims Drive, Singapore 387380

Incentive

- SGD 10 voucher (Grab e-voucher or NTUC FairPrice voucher)

For more info, contact Arushi at

email: arushi@my.jcu.edu.au

Experiment 2 Email to JCU lecturers requesting to showcase Poster

Dear Professor,

I hope this email finds you well.

My name is Arushi and I'm currently pursuing PhD under supervision of Dr Roberto Dillon and Dr Ai Ni Teoh at JCU Singapore.

I'm in the process of recruiting participants for my research experiment titled "A virtual reality approach to improving public speaking skills".

The present study seeks to validate the stress detection capabilities and effectiveness of social support offered in the developed Virtual Reality (VR) application. Participants will give a 5-minute speech in virtual reality. Blood pressure, heart rate and speech audio will be recorded. There will be questionnaires for participants to complete before and after the speech. The whole experiment will take approximately 30 minutes. All responses provided by the participants will be kept confidential.

The participants will be given **SGD 10 Grab e- voucher or SGD 10 NTUC FairPrice voucher** as an appreciation for taking part in the study.

I will be really grateful:

1. you could help me show the attached poster to students during your class (the poster includes study info and a QR code to register) or
2. email the poster to your students or
3. share the attached poster with your class students through announcements in the blackboard

Whichever you think is convenient.

I very much look forward to your kind support.

Best Regards,

Arushi

PhD (Information Technology) Candidate

Experiment 2 social media post message

Hello everyone!

I'm Arushi and I'm recruiting participants for my doctoral research study titled "A virtual reality approach to improving public speaking skills".

The present study seeks to evaluate the developed virtual reality application for its capabilities to detect stress from voice and provide feedback in real-time. You are eligible to participate if you are between the ages of 18 and 40; understand and know how to speak English; and have no history of epileptic seizures and motion sickness.

Participants will give a 5-minute speech in virtual reality. Blood pressure, heart rate and speech audio will be recorded. There will be two questionnaires. The total experiment duration is 30 minutes. Your responses will be de-identified and kept confidential.

Participants can choose to receive SGD \$10 Grab e-voucher or SGD \$10 NTUC FairPrice voucher.

If you choose to take part in this research, kindly register your availability via the following link (<https://lnkd.in/gjQxfSZY>) or scanning the QR code from poster attached.

Feel free to contact me via email at arushi@my.jcu.edu.au if you have any questions. I look forward to having you as a participant in our research.

Thank you!

Kind Regards

Arushi

Experiment 2 Consent Form

This administrative form
has been removed

Experiment 2 information sheet

PROJECT TITLE: “A Virtual Reality Approach to Improving Public Speaking Skills” (Ensure this title is in lay language – no discipline specific jargon)

You are invited to take part in a research project about “A Virtual Reality Approach to Improving Public Speaking Skills”. The purpose of this research is to investigate the effectiveness of a developed computational algorithm model for detecting stress and improving public speaking performance using virtual reality. The study is being conducted by Arushi will contribute to the degree in Doctor of Philosophy (Information Technology) at James Cook University. This research is funded by James Cook University Australia.

If you agree to be involved in the experiment, you will be invited to take part in the experiment. During the experiment you will speak in Virtual Reality. With your consent, your speech will be audio taped. And your blood pressure and heart rate will be measured before, during and after your speech. The speech should take approximately 5 minutes. The whole experiment will take approximately 30 minutes and will be conducted in the Psychology Lab at James Cook University Singapore campus. There are two questionnaires that you will need to complete, which ask you about your psychological state before and during the speech and how you feel about the task. Within the Virtual Reality environment, virtual audience will provide you feedback in real-time.

As a participant, you will receive SGD10 voucher (either grab e-voucher or NTUC Fair payment for participation).

Motion sickness should not be a reason of concern in our simulation as the avatars will remains static and you will remain seated during the speech task. Furthermore, if your participation in this study has caused you concerns, anxiety or any distress, counselling services are available at:
One Psychology Community: <https://singaporepsychologicalsociety.org/events/onepsychcommunity>
JCU Singapore Psychology Clinic: Contact number: + 65 6709 3762

Taking part in this study is completely voluntary and you can stop taking part in the study at any time without explanation or prejudice.

Your responses will be de-identified and kept confidential. The data from the study will be used in research publications and theses. You will not be identified in any way in these publications.

If you have any questions about the study, please contact – Arushi, A/Prof Roberto Dillon or Dr Ai Ni Teoh

Principal Investigator: Arushi College: James Cook University Singapore Phone: Email: Arushi@jcu.edu.au	Supervisor: A/Prof Roberto Dillon College: James Cook University Singapore Phone: + Email: Roberto.dillon@jcu.edu.au
	Secondary Supervisor Name: Dr Ai Ni Teoh College: James Cook University Singapore Phone: + Email: Aini.teoh@jcu.edu.au

If you have any concerns regarding the ethical conduct of the study, please contact:
Human Ethics, Research Office
James Cook University, Townsville, Queensland, 4811
Phone: (07) 4781 5011 (ethics@jcu.edu.au)

Demographic questionnaire for experiment 2

- 1) Please indicate your age: _____
- 2) sex: Male/ Female/ Indeterminate/Intersex/Unspecified/ Other
- 3) please indicate your country of birth: _____
- 4) please indicate your first language: _____
- 5) Please select the category that describes your occupation:
 - a. student
 - b. full-time/ self-employed
 - c. unemployed
 - d. others: _____
- 6) If you selected “student” in the previous question, please state the number of months left till the completion your course of study:
____(months)
- 7) I have a history of epileptic seizures and motion sickness
 - a. yes. b no

Appendix O *Ethics Approval for Experiment 2*

This administrative form
has been removed