

The Sound of Stress

New Insights into Speech-based Assessment and
Monitoring of Stress and Mental Health

Mitchel Kappen

Supervisor: prof. dr. Marie-Anne
Vanderhasselt

September 14th 2023



A dissertation submitted to Ghent University in partial fulfillment of the requirements for the degree of Doctor in Health Sciences

 FACULTY OF MEDICINE
AND HEALTH SCIENCES

 GHEPLab

 GHENT
UNIVERSITY

The Sound of Stress: New Insights into Speech-based Assessment and Monitoring of Stress and Mental Health

Mitchel Kappen

Supervisor: Prof Dr. Marie-Anne Vanderhasselt

*A dissertation submitted to Ghent University in partial fulfillment of the requirements for the
degree of Doctor in Health Sciences*

Academic year: 2022 - 2023

Doctoral guidance committee

Prof. Dr. Marie-Anne Vanderhasselt

Department of Head and Skin, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Ernst Koster

Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Gilbert Lemmens

Department of Head and Skin, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Kristof Hoorelbeke

Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Nilesh Madhu

Department of Electronics and Information Systems, Faculty of Engineering and Architecture, Ghent University, Ghent, Belgium

Doctoral exam committee

Prof. Dr. Patrick Calders (chair)

Department of Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Vincent Martin

Laboratoire Bordelais de Recherche en Informatique, Bordeaux Neurocampus, University of Bordeaux, Bordeaux, France

Prof. Dr. Jonas Everaert

Center of Research on Psychological Disorders and Somatic Diseases, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands

Prof. Dr. Martine Van Puyvelde

Department of Brain, Body, and Cognition, Faculty of Psychological and Educational Sciences, Free University of Brussels, Brussels, Belgium

Prof. Dr. Geert Crombez

Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

Prof. Dr. Miet De Letter

Department of Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

Dr. Matias Pulópolos Tripiana (secretary)

Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

ACKNOWLEDGMENTS

The acknowledgments are only present in the printed version of this dissertation.

“Keep your eyes on the stars,

And your feet on the ground.”

Theodore Roosevelt

SUMMARY

Stress is a universal phenomenon with significant implications for various health problems. Therefore, developing reliable and accessible methods to assess and monitor stress and mental health is crucial. Recently, speech has been proposed as a stress measure due to its production being affected by multiple stress-related physiological processes. The main objective of this dissertation is to identify the potential of speech as a measure of stress by addressing limitations in the existing literature and conducting a series of targeted studies.

In **Chapter 2**, we conducted network analyses on speech features pre- and post-stress induction, as well as on individual changes in each speech feature. The pre- and post-networks showed harmonics-to-noise ratio (HNR) as a central component, while the change network revealed jitter as the only direct connection between speech features and changes in negative affect. This study highlights the complex relationships between speech parameters and stress, setting the stage for confirmatory investigations.

Chapter 3 employed a cognitively challenging task with blocks containing either neutral or negative feedback to investigate stress effects on speech features. Results showed a significant increase in Fundamental Frequency (F0) and HNR, and a significant decrease in shimmer during the negative feedback condition. These results contribute to the understanding of stress effects on specific acoustic speech features in a well-controlled but ecologically-valid stress setting. This study is a solid step toward the generalization of these findings to real-life settings.

In **Chapter 4**, we developed the Ghent Semi-spontaneous Speech Paradigm (GSSP) to obtain speech samples closer to free speech while maintaining experimental control. The GSSP is a picture description task. We validated the GSSP by comparing it to both read-out-loud and everyday speech samples, showing that speech from the GSSP is acoustically similar to everyday speech samples and distinct from read-out-loud speech. Specifically, we propose that the GSSP should be the advised method of collecting speech in experimental settings, generating results that would be directly implementable in real-world scenarios.

In **Chapter 5** we exposed participants to different stress induction paradigms (Cyberball and MIST) to assess heterogeneity from a stressor perspective. We found changes in F0, speech rate, and jitter during the MIST paradigm, which elicited additional self-reported stress and physiological stress responses, but not during the Cyberball paradigm, which primarily affected self-reported negative affect. The results indicate that observed speech features are robust in semi-guided speech (as compared to previous studies using read-out-loud speech) and sensitive to stressors eliciting additional physiological stress responses, rather than solely increases in negative affect. This highlights the promise of speech as a tool for measuring stress in everyday settings, considering its affordability, non-intrusiveness, and ease of collection.

In **Chapter 6**, we discuss the potential impact of speech as a biosignal or biomarker in precision psychiatry, emphasizing its accessibility and affordability. We propose that speech, as a marker for stress as a transdiagnostic risk factor, could be the missing link in fine-tuning

systems for high-risk monitoring and just-in-time interventions (JITAls) if implemented securely and appropriately. Practical and ethical implications are also addressed.

In summary, this dissertation investigated the potential of speech analysis as a stress biomarker, providing a foundation for future research in this area. We identified specific speech features and investigated how they interact under stress (Chapter 2), validated these features using read-out-loud speech and negative social feedback (Chapter 3), developed a new methodology to collect naturalistic speech samples (Chapter 4), validated the speech features in semi-freely spoken speech and in different stressor paradigms (Chapter 5), and presented a context of future implementations for speech as a biomarker for stress and precision psychiatry (Chapter 6). Our findings support the potential of speech analysis as a non-invasive, affordable, and easily accessible tool for detecting and monitoring acute stress and hold promise for improving mental health care and overall well-being through novel monitoring methods and interventions.

SAMENVATTING

Stress is een universeel fenomeen met aanzienlijke gevolgen voor verschillende gezondheidsproblemen. Daarom is het cruciaal om betrouwbare en toegankelijke methoden te ontwikkelen om stress en geestelijke gezondheid te beoordelen en te monitoren. Onlangs is spraak voorgesteld als een stressmaat vanwege de betrokkenheid van meerdere stressgerelateerde fysiologische processen in de productie ervan. Het belangrijkste doel van dit proefschrift is het potentieel van spraak als stressmaat te identificeren door beperkingen in de bestaande literatuur aan te pakken en een reeks gerichte onderzoeken uit te voeren.

In **Hoofdstuk 2** voerden we netwerkanalyses uit op spraakkenmerken voor en na stressinductie, evenals op individuele veranderingen in elk spraakkenmerk. De voor- en na-netwerken toonden harmonics-to-noise ratio (HNR) als een centraal onderdeel, terwijl het veranderingsnetwerk jitter onthulde als de enige directe verbinding tussen spraakkenmerken en veranderingen in negatieve affectiviteit. Deze studie benadrukt de complexe relaties tussen spraakparameters en stress en vormt het uitgangspunt voor bevestigende onderzoeken.

Hoofdstuk 3 maakte gebruik van een cognitief uitdagende taak met blokken met neutrale of negatieve feedback om stress-effecten op spraakkenmerken te onderzoeken. Resultaten toonden een significante toename van de Fundamental Frequency (F0) en HNR, en een significante afname van shimmer tijdens de negatieve feedback conditie. Deze resultaten dragen bij aan het begrip van stress-effecten op specifieke akoestische spraakkenmerken in

een goed gecontroleerde maar ecologisch valide stressomgeving. Deze studie vormt een solide stap in de richting van de generalisatie van deze bevindingen naar real-life situaties.

In **Hoofdstuk 4** ontwikkelden we het Ghent Semi-spontaneous Speech Paradigm (GSSP) om spraakopnames die dichter bij vrije spraak liggen te verkrijgen, terwijl we experimentele controle behouden. Het GSSP is een taak waarin afbeeldingen luidop omschreven worden. We valideerden het GSSP door het te vergelijken met zowel voorlees- als alledaagse spraakopnames, waarbij we aantonden dat spraak uit het GSSP akoestisch vergelijkbaar is met alledaagse spraakopnames en verschilt van voorlees-spraak. We stellen specifiek voor dat het GSSP de geadviseerde methode zou moeten zijn voor het verzamelen van spraak in experimentele settings, om resultaten die direct implementeerbaar zijn in real-world scenario's te genereren.

In **Hoofdstuk 5** stelden we deelnemers bloot aan verschillende stressinductieparadigma's (Cyberball en MIST) om heterogeniteit vanuit het perspectief van stressoren te beoordelen. We vonden veranderingen in F0, spraaksnelheid en jitter tijdens het MIST-paradigma, dat bijkomende zelfgerapporteerde stress en fysiologische stressreacties opriep, maar niet tijdens het Cyberball-paradigma, dat voornamelijk invloed had op zelfgerapporteerde negatieve affectiviteit. De resultaten geven aan dat waargenomen spraakkenmerken robuust zijn in semi-begeleide spraak (in vergelijking met eerdere onderzoeken met voorlees-spraak) en gevoelig zijn voor stressoren die extra fysiologische stressreacties oproepen, in plaats van uitsluitend toenames van negatieve affectiviteit. Dit benadrukt de potentie van spraak als hulpmiddel voor het meten van stress in alledaagse situaties, gezien de betaalbaarheid, lage intrusiviteit en het gemak van verzameling.

In **Hoofdstuk 6** bespreken we de mogelijke impact van spraak als biosignaal of biomarker in precisiepsychiatrie, waarbij we de nadruk leggen op toegankelijkheid en betaalbaarheid. We stellen voor dat spraak, als marker voor stress als een transdiagnostische risicofactor, de ontbrekende schakel zou kunnen zijn in het verfijnen van systemen voor monitoring van hoge risico patiëntengroepen en interventies op het juiste moment (JITAls) indien op een veilige en geschikte manier geïmplementeerd. Praktische en ethische implicaties worden ook besproken.

Samenvattend heeft dit proefschrift het potentieel van spraakanalyse als stressbiomarker onderzocht en een basis gelegd voor toekomstig onderzoek op dit gebied. We identificeerden specifieke spraakkenmerken en onderzochten hoe ze zich onder stress onderling verhouden (hoofdstuk 2), valideerden deze kenmerken met behulp van voorlees-spraak en negatieve sociale feedback (hoofdstuk 3), ontwikkelden een nieuwe methodologie om natuurlijke spraakopnames te verzamelen (hoofdstuk 4), valideerden de spraakkenmerken in semi-vrij gesproken spraak en in verschillende stressorparadigma's (hoofdstuk 5) en presenteerden een context voor toekomstige implementaties voor spraak als biomarker voor stress en precisiepsychiatry (hoofdstuk 6). Onze bevindingen ondersteunen het potentieel van spraakanalyse als een niet-invasieve, betaalbare en gemakkelijk toegankelijke tool voor het detecteren en monitoren van acute stress en de potentie voor het verbeteren van de geestelijke gezondheidszorg en het algemeen welzijn door middel van nieuwe monitoringmethoden en interventies.

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
ANS	Autonomic Nervous System
BIC	Bayesian Information Criterion
CGN	Corpus Gesproken Nederlands
CI	Confidence Interval
DASS	Depression, Anxiety, and Stress Scale
DSM	Diagnostics and Statistical Manual of mental disorders
EBIC	Extended Bayesian Information Criterion
ECG	Electrocardiogram
EDA	Electrodermal Activity
EMA	Ecological Momentary Assessment
EMM	Estimated Marginal Mean
F0	Fundamental frequency
F1	Formant 1
F2	Formant 2
F1/2	Formant 1/2 ratio
FDR	False Discovery Rate
GGM	Gaussian Graphical Model
gLASSO	graphical Least Absolute Shrinkage and Selection Operator
GLMM	Generalized Linear Mixed Model

GSSP	Ghent Semi-spontaneous Speech Paradigm
HNR	Harmonics-to-Noise Ratio
HPA	Hypothalamic-Pituitary-Adrenal
HRV	Heart Rate Variability
IBI	Interbeat Interval
JIT (chapter 2)	Jitter
JITAI	Just-In-Time Adaptive Intervention
LLD	Low-Level Descriptor
LMM	Linear Mixed Model
M	Mean
M (chapter 4)	Marloes
MIST	Montreal Imaging Stress Task
MVSPS	Mean Voiced Segments Per Second
MVSL	Mean Voiced Segment Length
NA	Negative Affect
P (chapter 4)	PiSCES
PANAS	Positive And Negative Affect Schedule
POMS	Profile Of Mood States
PSS	Perceived Stress Scale
R (chapter 4)	Radboud
RRS	Ruminative Response Scale
SAM	Self-Assessment Manikin
SCL	Skin Conductance Level

SCR	Skin Conductance Responses
SCRR	Skin Conductance Response Rate
SD	Standard Deviation
SEM	Standard Error of the Mean
SET	Social Evaluative Threat
SNS	Sympathetic Nervous System
t-SNE	t-Distributed Stochastic Neighbor Embedding
TSST	Trier Social Stress Test
VAD	Voice Activity Detection
VAS	Visual Analog Scale
VO (chapter 2)	Mean Voiced Segment Length (MVSL)
VU-AMS	VU Ambulatory Monitoring System

TABLE OF CONTENTS

General Introduction.....	1
1.1. Stress.....	1
1.2. Components of Stress.....	5
1.2.1. Physiological Component.....	5
1.2.2. Behavioral Component.....	6
1.2.3. Psychological Component.....	6
1.3. Stress Measurement.....	7
1.3.1. Psychological methods.....	7
1.3.2. Behavioral methods.....	8
1.3.3. Physiological methods.....	9
1.3.4. Multimethod approach.....	10
1.4. Speech as a potential method for stress measurement.....	13
1.4.1. Practical implications.....	14
1.4.2. Definition of speech features and existing research on their link to stress.....	15
1.4.2.1. Fundamental Frequency (F0).....	16
1.4.2.2. Jitter.....	17
1.4.2.3. Shimmer.....	18
1.4.2.4. Harmonics-to-noise Ratio (HNR).....	19
1.4.2.5. Speech rate.....	20
1.4.3. Limitations of current literature.....	21
1.5. Objectives.....	23
1.6. References.....	26
Speech as an indicator for psychosocial stress: A network analytic approach.....	37
2.1. Abstract.....	38
2.2. Introduction.....	39
2.3. Methods.....	45
2.3.1. Apparatus and Procedure.....	45
2.3.2. Stress induction procedure.....	47
2.3.2.1. Trial.....	48
2.3.3. Self-report measurement – Negative Affect; NA.....	49
2.3.4. Extraction of speech parameters.....	49
2.3.5. Statistical Analyses.....	50
2.3.6. Data preparation and network estimation.....	51

2.3.7. Network visualization.....	52
2.3.8. Network comparison.....	53
2.3.9. Evaluation of the stability and accuracy of the models.....	53
2.4. Results.....	54
2.4.1. Manipulation Check.....	54
2.4.2. Impact of Stress on the Interrelations Between Speech Parameters (aim 1).....	56
2.4.3. Modeling the Unique Associations Between Stress Reactivity and Change in Speech Parameters (aim 2).....	60
2.5. Discussion.....	62
2.6. Conclusions.....	67
2.7. References.....	69
2.8. Supplemental Materials.....	75
2.8.1. Acknowledgments.....	75
2.8.2. Funding.....	75
2.8.3. Open practices statement.....	75
2.8.4. Supplemental materials.....	75
Acoustic speech features in social comparison: how stress impacts the way you sound.....	92
3.1. Abstract.....	93
3.2. Introduction.....	94
3.3. Methods.....	98
3.3.1. Participants.....	98
3.3.2. Apparatus and procedure.....	99
3.3.2.1. Read-out-loud text “Marloes”.....	99
3.3.2.2. On-site experimental session.....	99
3.3.2.3. Trial.....	100
3.3.2.4. Stress induction.....	102
3.3.3. Extraction of speech features.....	103
3.3.4. Data-Analysis.....	104
3.4. Results.....	106
3.4.1. Manipulation check.....	106
3.4.1.1. Self-reports.....	106
3.4.1.2. Physiological activity.....	107
3.4.2. Speech feature analysis.....	109
3.4.2.1. Harmonics-to-Noise Ratio (HNR).....	109
3.4.2.2. Shimmer.....	110
3.4.2.3. Fundamental Frequency (F0).....	110

3.4.2.4. Jitter, voiced segments per second, mean voiced segment length.....	111
3.5. Discussion.....	112
3.6 References.....	117
3.7. Supplemental Materials.....	123
3.7.1. Acknowledgments.....	123
3.7.2. Funding.....	123
3.7.3. Supplemental Materials.....	123
Ecologically Valid Speech Collection in Behavioral Research: The Ghent Semi-spontaneous Speech Paradigm (GSSP).....	130
4.1. Abstract.....	131
4.2. Introduction.....	132
4.3. Methods.....	136
4.3.1. Web app and procedure.....	137
4.3.1.1. Read-out-loud text “Marloes”.....	139
4.3.1.2. GSSP picture description speech.....	140
4.3.1.3. Drinking break.....	141
4.3.2. Participants.....	141
4.3.3. Data processing.....	142
4.3.3.1. Audio data processing.....	142
4.3.3.2. Acoustic Speech parameter extraction.....	143
4.3.3.3. External dataset “Corpus Gesproken Nederlands”.....	145
4.4. Results.....	146
4.4.1. Arousal & valence scores.....	146
4.4.2. Speech feature analysis.....	147
4.4.2.1. Speech duration.....	147
4.4.2.2. OpenSMILE acoustics.....	148
4.4.2.2.1. Temporal features.....	148
4.4.2.2.2. Frequency-related features.....	150
4.4.2.2.3. Amplitude-related features.....	151
4.4.2.2.4. Jitter and shimmer inconsistencies.....	152
4.4.3. ECAPA-TDNN projections.....	154
4.4.4. CGN validation.....	156
4.5. Discussion.....	159
4.6. References.....	163
4.7. Supplemental Materials.....	170
4.7.1. Acknowledgments.....	170
4.7.2. Funding.....	170

4.7.3. Supplemental Materials.....	170
4.7.3.S1. Web Application Details.....	172
S1.1. Welcome Page.....	172
S1.2. Introduction Page.....	172
S1.3. Instruction Page.....	174
S1.4. Rest Block.....	177
S1.5. “Marloes” Text.....	177
S1.6. GSSP Web App Image Subsets.....	178
4.7.3.S2. Speech Data Parsing.....	180
4.7.3.S4. OpenSMILE Sampling Rate Inconsistency.....	183
4.7.3.S5. OpenSMILE Delta Visualizations.....	185
4.7.3.S6. ECAPA-TDNN & GeMAPS Distribution Plots.....	186
4.7.3.S7. Logistic Regression Weight Coefficients.....	187
4.7.3.S8. Effect Size Shimmer & Jitter.....	189
S8.1. Shimmer.....	189
S8.2. Jitter.....	190
4.7.3.S9. Factor Analysis.....	191
1.0. Pisces Dataset.....	193
1.1. Valence.....	193
1.1.1. Cronbach’s Alpha.....	193
1.1.2. CFA.....	193
Fit and visualize.....	193
1.1.3. CFA Visualization.....	195
1.1.4. Distributions.....	195
Visualizations.....	195
1.2. Arousal.....	196
1.2.1. Cronbach’s Alpha.....	196
1.2.2. CFA.....	196
Fit and visualize.....	197
1.2.3. CFA Visualization.....	198
1.2.4. Distributions.....	199
Visualizations.....	199
2.1.1. Cronbach’s Alpha.....	199
2.1.2. CFA.....	200
Fit and visualize.....	200
2.1.3. CFA Visualization.....	202
2.1.4. Distributions.....	202

Visualizations.....	202
2.2.1. Cronbach's Alpha.....	202
2.2.2. CFA.....	203
Fit and visualize.....	203
2.2.3. CFA Visualization.....	205
2.2.4. Distributions.....	205
Visualizations.....	205
Acoustic and Prosodic Speech Features Reflect Physiological Stress but not Isolated Negative Affect: A Multi-Paradigm Study on Psychosocial Stressors.....	206
5.1. Abstract.....	207
5.2. Introduction.....	207
5.2.1. Research Objectives & Hypotheses.....	211
5.3. Methods.....	213
5.3.1. Participants.....	213
5.3.1. Procedure.....	214
5.3.1.1. On-site experimental session.....	214
5.3.1.2. Cyberball - Day 1.....	215
5.3.1.3. MIST - Day 2.....	216
5.3.2. Data Collection.....	217
5.3.2.1. Speech Data.....	217
5.3.2.2. Self-report Data.....	218
5.3.2.3. Physiological Data.....	218
5.3.3. Data Analysis.....	219
5.3.3.1. Physiological Data.....	219
5.3.3.2. Extraction of Speech Features.....	219
5.3.3.3. Statistical Analysis.....	220
5.4. Results.....	222
5.4.1. Physiological.....	222
5.4.1.1. Skin Conductance Response Rate (SCRR).....	222
5.4.2. Self-reports.....	223
5.4.2.1. Negative Affect.....	223
5.4.2.2. Stress.....	223
5.4.3. Speech.....	223
5.4.3.1. Fundamental Frequency (F0).....	223
5.4.3.2. Voiced segments per second (MVSPS).....	223
5.4.3.3. Voiced segment length (MVL).....	224
5.4.3.4. Harmonics-to-noise ratio (HNR).....	224

5.4.3.5. Shimmer.....	224
5.4.3.6. Jitter.....	224
5.5. Discussion.....	226
5.6. Conclusions.....	230
5.7. References.....	232
5.8. Supplemental Materials.....	237
5.8.1. Acknowledgments.....	237
5.8.2. Funding.....	237
5.8.3. Supplemental Materials.....	238
5.8.3.1. Exclusion Criteria.....	239
5.8.3.2. Complete study flowchart.....	240
5.8.3.3. Speech data collection screenshots.....	241
Cyberball.....	242
MIST.....	243
5.8.3.4. Self-reports.....	244
5.8.3.4.1. Positive activating affect.....	244
5.8.3.4.2. Positive soothing affect.....	245
5.8.3.5. ECG/HRV data.....	246
5.8.3.6. Software and packages used.....	248
5.8.3.6.1. R.....	248
5.8.3.6.2. Python.....	248
5.8.3.7. Full models & Anova results.....	248
5.8.3.7.1. Skin Conductance Response Rate (SCRR).....	249
5.8.3.7.2. Negative Affect.....	250
5.8.3.7.3. Self-Reported Stress.....	251
5.8.3.7.4. Fundamental Frequency (F0).....	252
5.8.3.7.5. Voiced segments per second.....	253
5.8.3.7.6. Voiced segment length.....	254
5.8.3.7.7. Harmonics-to-noise ratio (HNR).....	255
5.8.3.7.8. Shimmer.....	256
5.8.3.7.9. Jitter.....	257
Speech as a Promising Biosignal in Precision Psychiatry.....	258
6.1. Abstract.....	259
6.2. Introduction.....	260
6.3. Speech Contains Critical Psychosocial Information.....	261
6.4. Stress and Speech.....	263
6.5. Real-world Validation and Application.....	264

6.6. Ethical and Legal Issues.....	265
6.7. Conclusion.....	265
6.8. References.....	267
6.9. Supplemental Materials.....	272
6.9.1. Funding.....	272
General Discussion.....	273
7.1. General overview of the findings.....	276
7.2. Theoretical implications.....	281
7.2.1. Combined patterns of speech features.....	281
7.2.2. Current heterogeneity.....	281
7.2.2.1. Speech styles.....	283
7.2.2.2. Diversity in stressors.....	286
7.3. Practical and clinical implications.....	287
7.4. Continuous monitoring.....	289
7.5. Limitations.....	292
7.5.1. Limitations of speech research in previous work.....	292
7.5.2. Limitations of this dissertation.....	293
7.6. General suggestions for future research.....	295
7.6.1. Language-based markers.....	297
7.7. General Conclusions.....	299
7.8. References.....	301
Personal Contributions.....	315
Curriculum Vitae.....	317

General Introduction

1.1. Stress

Stress is a common phenomenon, experienced by all on a frequent basis. Stress is our body's psychological, physiological, and behavioral response to any kind of threat or demand. It can be triggered by a wide variety of factors that elicit stress, such as financial concerns, child care, interpersonal dynamics, or major life events, which we refer to as *stressors*. Nowadays, in our fast-paced and ever-changing society, stress has become an increasingly relevant topic, especially since it plays a significant role in both our physical and mental health.

While stress is a normal and adaptive part of life that can help us cope with challenges and motivate us to achieve our goals, it is important to note when stress becomes detrimental. Specifically, when stress becomes more chronic, it can lead to a wide range of health problems including cardiovascular disease, coronary heart disease, anxiety disorders, depression, autoimmune disease, and neurodegenerative disorders, among others (Bhushan et al., 2020;

Brosschot et al., 2017; Cohen et al., 2007; Juster et al., 2010; Kappen et al., 2023; Slavich & Irwin, 2014).

In this dissertation, we predominantly focus on psychosocial stressors, as they play a critical role in contributing to stress-induced (mental) health complications for several reasons (Epel et al., 2018; Kogler et al., 2015). As humans are inherently social beings, social interactions and the need to belong are essential aspects of our lives (Baumeister & Leary, 1995). Psychosocial stressors can disrupt these fundamental human needs, significantly impacting an individual's well-being. Additionally, these stressors are pervasive in everyday life, leading to a higher probability of chronic exposure compared to other types of stressors, such as physical (e.g., receiving electrical shocks) or cognitive stressors (e.g., reaction time tasks). This chronic exposure to psychosocial stressors further explains their dominant presence in stress-related diseases (Dupre et al., 2015; Melchior et al., 2007; Phelan et al., 1991; Tennant, 2001).

To better understand stress, many different models of stress have been developed. In this dissertation, the stress response will be explained from the transactional model of stress and coping by Lazarus and Folkman (1984). This model is chosen because 1) it highlights the dynamic nature of stress by acknowledging the ongoing, reciprocal relationship between the individual and their environment, making it highly relevant to real-life situations, and 2) it is applicable across various contexts, including health, work, and relationships, making it a versatile choice for exploring stress in different settings, and 3) it takes individual differences into accounts, allowing for a more personalized approach to understanding stress.

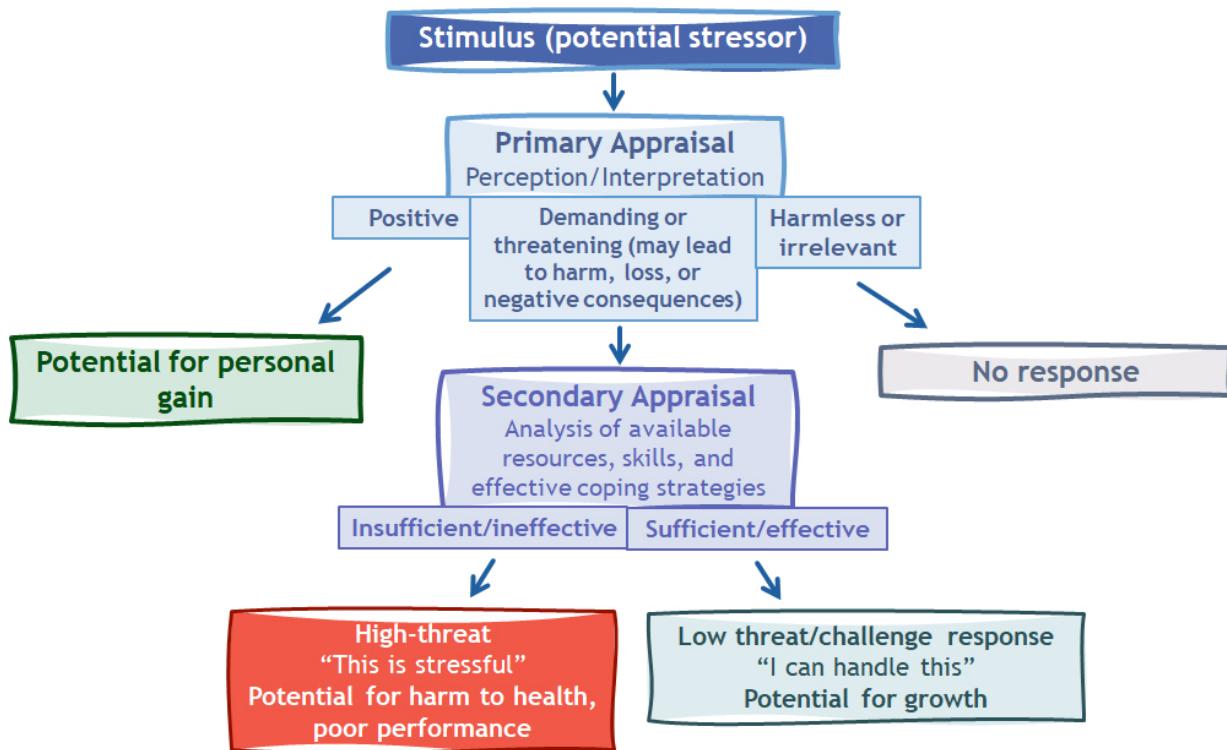
In the seminal transactional model of stress and coping by Lazarus and Folkman, stress is determined by a person's primary appraisal of a stimulus (Lazarus & Folkman, 1984). With this primary appraisal, an individual determines whether something could pose a potential threat to their well-being or not. Potential threats can come from many different parts of life and affect many different aspects of one's well-being, including but not limited to physical, emotional, social, cognitive, spiritual, behavioral, and financial well-being.

In addition to the primary appraisal, a secondary appraisal will be applied to any stimulus that poses a direct threat to the individual's well-being, irrespective of the nature of the threat. Primary and secondary appraisals are distinguished for conceptual purposes, with neither being more important than the other or that one occurs before the other (Carpenter, 2016). The secondary appraisal concentrates on how an individual can respond to and cope with the situation (Carpenter, 2016; Lazarus & Folkman, 1984). This evaluation involves assessing both internal and external resources, such as cognitive capacity, personal skills, social support networks, and available options to deal with the stressor (Carpenter, 2016; Lazarus & Folkman, 1984). Secondary appraisal can lead to different outcomes depending on the perceived availability and effectiveness of coping resources (Folkman & Moskowitz, 2004; Lazarus & Folkman, 1984). Coping strategies can be either problem-focused (aimed at changing the situation itself) or emotion-focused (aimed at changing the relation to the situation). The effectiveness of one's coping strategies depends on the context and the individual's ability to adapt to changing circumstances. If an individual believes they have adequate resources and effective coping strategies to deal with the stressor, they may experience less stress and negative emotions. On the other hand, if the person perceives their

coping resources as inadequate or ineffective, they may experience heightened stress and negative emotions (Folkman & Moskowitz, 2004; Lester et al., 1994).

Figure 1

Schematic illustration of the transactional model of stress by Lazarus & Folkman



Note. Image courtesy of Lydia G. Roos

Experienced stress is the result of an unfolding of multiple biological processes that comprise the stress response system (Brosschot et al., 2017; Cohen et al., 2007; Lazarus & Folkman, 1984). The stress response system is made up of multiple biological systems that work in a coordinated fashion to help an organism pick up and respond to environmental cues (stressors) in order to best respond (e.g., fight or flight) to reduce the potential for harm. This could include physical threats, but as we have evolved, this system has become sensitive in

responding to cues that confer not only physical danger but also social danger. Furthermore, it is important to note that the stress response will not exclusively be elicited by true and imminent dangers, but it can activate even before a stressor is faced. This allows us to prepare for an anticipated threat, which means the stress response system can still be activated regardless of whether an actual threat occurs.

1.2. Components of Stress

The multidimensional nature of stress can be decomposed into three main components: psychological, behavioral, and physiological, which will be shortly explained here after which we will discuss how each individual component is used in stress measurement. The physiological, behavioral, and psychological aspects of stress are interconnected, together forming a comprehensive system that helps individuals cope with and adapt to stressors. All responses, physiological, behavioral, and psychological, depend on a large network of brain systems involved in appraisal, emotion, and memory, including key forebrain areas such as the amygdala, prefrontal cortex, and hippocampus (Braun, 2011; Joëls & Baram, 2009; Myers et al., 2017; Sousa & Almeida, 2012).

1.2.1. Physiological Component

The physiological component of stress involves the activation of key pathways, such as the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system (SNS), as part of the autonomic nervous system (ANS) which mediate the body's adaptive functions in

response to stress. This includes the release of stress hormones (cortisol and adrenaline), increased heart rate, elevated blood pressure, and other bodily responses to help prepare the individual to confront or escape the stressor, often known as the "fight or flight" response (Cannon, 1939; Kyrou & Tsigos, 2009).

1.2.2. Behavioral Component

The behavioral component of stress encompasses the observable actions and reactions that individuals exhibit when encountering stressors. Responses can include emotional outbursts, such as anger or frustration, or changes in body language, such as tensing up or pacing. Individuals may also seek social support or display other actions that reflect their attempt to manage or adapt to the stressor. Crucially, these brain networks interact with the hypothalamus and brainstem, integrating behavioral responses with the physiological reactions mediated by the HPA axis and autonomic activity (Joëls & Baram, 2009; McKlveen et al., 2015; Myers et al., 2017; Ulrich-Lai & Herman, 2009). It's important to note that these responses can be automatic or deliberate, and their effectiveness in dealing with stress may vary.

1.2.3. Psychological Component

The psychological component involves cognitive processes, such as perception, appraisal, and emotion regulation, as well as the emotional experiences related to stress, such as anxiety, rumination, need to belong, or negative affect (Williams, 2007; Zwolinski, 2012). The psychological component influences how individuals perceive and react to stressors, which in turn affects their physiological and behavioral responses. However, there is a limited correlation

between physiological and psychological stress reactions, possibly due to the subjective nature (i.e., dependent on appraisal) of self-reported, experienced stress (Campbell & Ehlert, 2012; Zwolinski, 2012).

In this dissertation, we will mainly focus on the physiological and psychological aspects of stress to ensure the successful induction of stress in our studies, while speech lies at the intersection of the behavioral and physiological components. It is crucial to recognize that these different components are deeply interlinked, and are therefore hard to approach completely irrespective of each other.

1.3. Stress Measurement

Due to this multidimensional nature of stress, composed of three main components (i.e., psychological, behavioral, and physiological), stress measurement methods have been developed specifically targeting each of these components. In the next segments, we will outline each individually, shortly highlighting their pros and cons. Refer to Table 1 for a summarized overview of stress measurement techniques.

1.3.1. Psychological methods

These methods assess stress by asking individuals to report their stress levels through (self-report) questionnaires or interviews. Examples of such methods include the Perceived Stress Scale (PSS), Self-Assessment Manikin (SAM), Positive and Negative Affect Schedule (PANAS), and the Profile Of Mood States (POMS), which use items such as “*In the last*

week/month how often have you .. been upset because of something that happened unexpectedly? / felt nervous and stressed?" or visual analog scales (VAS) that visually depict one's feelings (Bradley & Lang, 1994; Cohen et al., 1994; McNair et al., 1971; Monroe, 2008; Rossi & Pourtois, 2012; Watson et al., 1988). The advantages of self-report methods include that they are relatively easy and inexpensive to administer, can provide information about an individual's subjective experience of stress, and can be used in a variety of settings, and will therefore likely remain an important feature in [multimethod approaches](#) (Abbas et al., 2021). Interview-based measures, on the other hand, are accurate yet costly and time-consuming (Cohen et al., 1997; Dohrenwend, 2006; G. S. Shields & Slavich, 2017). The limitations of assessing the subjective experience of stress are that they are by definition influenced by a multitude of systematic measurement errors, such as response biases (e.g., social desirability; Razavi, 2001; Welte & Russell, 1993). Moreover, these measures are mostly retrospective in nature and can be influenced by an individual's mood and motivation at the exact time of administration, making it susceptible to momentary changes and less representative of one's general state of being over a broader range of time (Monroe, 2008; Monroe & Slavich, 2016).

1.3.2. Behavioral methods

These methods assess stress by observing an individual's behavior, such as changes in task performance, body gestures, or facial expressions that occur in response to stress. The advantages of behavioral methods include that they can provide information about how stress affects an individual's behavior, and they can be used in real-world settings. Limitations include that these measures often require difficult remote setups in order to have high-quality

recordings (e.g., video), or for people to come into a laboratory (which is impractical and sometimes even not possible in cases of a pandemic, disease, or remote living). Besides, even though some behavioral bodily patterns (such as facial expressions and body gestures) are manifested in response to stress, they may also be subject to intentional or even partially conscious control (Hobfoll, 2004; Lin et al., 1985; Ryder & Chentsova-Dutton, 2012). Consequently, related recordings may also contain systematic errors when used to estimate the magnitude of the stress response (Giannakakis et al., 2022).

1.3.3. Physiological methods

These methods assess stress by measuring physiological responses that are associated with the stress response, such as changes in heart rate, blood pressure, cortisol levels, and electrodermal activity (Arza et al., 2019; Giannakakis et al., 2022). This dissertation mostly uses electrodermal activity and cardiac-related measures, such as skin conductance levels, skin conductance response rates, and event-related cardiac responses. Electrodermal activity (EDA) is focused on the conductance of one's skin, which is affected by the amount of sweat that is present. This is relevant, as sweat glands are predominantly innervated by the sympathetic chain of the ANS (Dawson et al., 2017; S. Shields et al., 1987). The EDA signal is a pure index of sympathetic activation (and arousal) and is built up of tonic (slow) and phasic (fast) components, from which we can extract several measures such as skin conductance level (SCL) and skin conductance responses (SCRs), each responding to different types of stimuli (Braithwaite et al., 2013). An electrocardiogram (ECG) is a waveform that represents cardiac activity (Sherwood, 2003). The ECG is composed of different components and is

closely related to the ANS (Klabunde, 2011; Sherwood, 2003). Variation in time between two consecutive heartbeats is described as heart rate variability (HRV) and changes therein are directly linked to changes in the ANS (Camm et al., 1996; Malik & Camm, 1990) and are also recently considered in an event-related manner, which allows for a more nuanced understanding of how the ANS reacts to individual stressors (Gunther Moor et al., 2010; van der Veen et al., 2019).

The advantages of physiological methods include that they are objective, can be used in real-world settings (e.g., cardiac and skin conductance), and can provide information about the body's stress response. The limitations include that they may not always be feasible to use in certain settings as they often require individual context to be accurately interpreted, and can be expensive, impractical, intrusive, and unreliable in real-world, daily life settings (Arza et al., 2019; Giannakakis et al., 2022; Slavich et al., 2019). Moreover, there is still lacking understanding concerning the consistency of some of these metrics, with for instance cortisol responses being strongly influenced by factors such as hormones associated with the menstrual cycle (Kirschbaum & Hellhammer, 1994), anticipatory appraisal (Gaab et al., 2005), and lack translatability to chronic metrics of stress (Lee et al., 2015).

1.3.4. Multimethod approach

Lastly, it should be noted that all of these different measures of stressor exposure, stress perceptions, and psychological, and physiological stress responses are only moderately associated with each other (Mauss et al., 2005). Especially when taking into account the different ways stress comes to show. For example, sometimes when you get stressed it affects

your mood (psychological) and sleep (physiological), other times you get sick more easily (immune), and other times you have a hard time focusing (cognitive). Therefore, focusing on only one method can give a severely limited view of one's stress levels. In addition, it is unclear which of these methods and measures is most accurate and closest related to health outcomes, as it differs between health outcomes, between individuals, and whether the focus is on short or long-term health outcomes (Cohen et al., 1997; Epel et al., 2018; Rehkopf et al., 2010). Considering the multidimensional nature of stress, and that it is being experienced on multiple levels such as social, psychological, and physiological, there have been developments in a multilevel approach to measuring stress (Arza et al., 2019; Epel et al., 2018; Monroe & Slavich, 2016). Despite a certain approach potentially yielding accurate stress indices, one should consider the trade-off between accuracy and intrusiveness for the individual. For example, the limitation of self-report data being momentary could be tackled by increasing the frequency of measurement, but this significantly increases the task load for individuals. Moreover, the limitation of physiological data requiring expensive specialized apparatus, trained professionals, and limiting freedom of movement due to wired electrodes are being tackled by the development of wearable devices (e.g., smartwatches), but this comes at a trade-off in signal quality and predictive capacity. Lastly, people from marginalized communities, such as low-income families, are often underrepresented in studies given the burden of many study designs and methodologies (e.g., physical lab presence), yielding insufficient knowledge and ability to measure stress in populations that need it most (L. T. Clark et al., 2019). Therefore, there is a persisting need for a novel method that is non-intrusive, remote-friendly, easy to collect, affordable, and yet accurate. Measuring stress through one's speech could be a promising approach to address some of the limitations and trade-offs mentioned earlier.

Indeed, due to the multimodal nature of stress, it is unlikely that one method alone can fulfill all requirements and achieve the highest accuracy. However, speech has the potential to both serve as a valuable indicator of daily stress and can act as a complementary measure in multimodal systems.

Table 1.

A schematic (concise) overview of different stress measurement techniques, their corresponding dimension of stress, and respective pros and cons.

Dimension	Measurement Method	Pros	Cons
Psychological	Self-report, Interview-based	Easy and inexpensive, Reflects subjective experience, Can be used in varied settings, Accurate (for interview-based)	Susceptible to response biases, Mostly retrospective, Influenced by mood and motivation at administration, Costly and time-consuming (for interview-based)
Behavioral	Observation of behaviors	Reflects real-world behavior	Requires remote setups or lab presence, Gestures and expressions can be controlled, Contains systematic errors for stress estimation
Physiological	Heart rate, blood pressure, cortisol levels, electrodermal activity	Objective, Can be used in real-world settings	Can be expensive, impractical, intrusive, Require context for interpretation, Unreliable in daily settings, Lacking understanding of consistency, Affected by various factors (e.g., menstrual cycle, anticipatory)appraisal

1.4. Speech as a potential method for stress measurement

Speech production is an intricate process involving multiple physiological systems and components of the body. The process begins with the conscious decision of what to say, considering factors such as word choice and tone of voice. The physical aspect of speech production, however, occurs more automatically, with the body modulating the tension of various muscles to expel air through the vocal folds and vocal tract, generating sound waves (Titze & Martin, 1998).

Voice production is described as the most complex of innately acquired human motor skills and processing relies on the collaboration of approximately 100 muscles innervated by multiple cranial and spinal nerves (Duffy, 2000), multiple subcortical and cortical brain regions (Carlson & Birkett, 2017; Jürgens, 2002), and cardiorespiratory processes (Câmara & Griessenauer, 2015; Monkhouse, 2005). As a result, speech is a psychophysiological process influenced by both external and internal stressors (Hansen & Patil, 2007; Van Puyvelde et al., 2018).

Van Puyvelde and colleagues further argue that vocal and stress responses share similar underlying cardiorespiratory processes governed by the ANS. The parasympathetic vagal system, which is critical for stress regulation (e.g., Berntson et al., n.d.; Thayer & Lane, 2000, 2009), is also involved in voice and speech coordination (e.g., Câmara & Griessenauer, 2015). This connection positions voice output as a psychophysiological response that is part of

the human integrative psychophysiological stress system (Thayer & Lane, 2000, 2009). This has resulted in an emerging field of identifying how speech is affected by stress, as well as stress detection from speech samples (Giddens et al., 2013; Slavich et al., 2019; Sondhi et al., 2015; Van Puyvelde et al., 2018; Zhou et al., 2001). The potential of speech as a promising avenue for stress research is evident, however, it remains a relatively young field with vastly heterogeneous results (Giddens et al., 2013; Van Puyvelde et al., 2018).

1.4.1. Practical implications

The detection of stress through speech recordings comes with several perks that have been identified and that have been noted as limitations in other methodologies (Slavich et al., 2019; Van Puyvelde et al., 2018). First, it is **affordable**. No specialized, expensive, apparatus are needed to collect or analyze speech samples. Any microphone will be able to record samples that contain valuable information, however, it should be noted that microphone and sample quality will affect the information density. This ties in with the advantage of this data being **easily accessible**. There is a ubiquity of high-quality microphones, with the increasing prevalence of microphones in everyday objects, that makes speech data more accessible and reliable. Moreover, these samples can be easily obtained from everyday sources like phone calls or meetings, thoroughly simplifying the data collection process. Ethical concerns should be taken into account, though, see Chapter 6 and Slavich et al. (2019). Stress measurements through speech would also have great potential in its **applicability in natural settings**. Due to the ability to collect this data and apply this method in real-world environments, it vastly increases the ecological validity of stress measurement. Moreover, speech data can be

collected without physical contact or the connection of electrodes, making it **non-intrusive**. This reduces discomfort and inconvenience for participants and may provide more ecologically valid indices of one's actual physical state as compared to controlled lab environments and highly intrusive physiological measurement methods. Speech samples are also highly information dense; containing a lot of information, even in relatively short samples. Therefore, a measure could be **swift**, and easily implemented on a **high-frequency** basis. This provides quick and efficient stress evaluations without the need for in-person contact and **no high demand for the participant**. Lastly, this enables us to use speech analysis for the **passive monitoring of stressors in daily life**, which could have significant health implications, without the day-to-day demand for participants and patients being too big.

1.4.2. Definition of speech features and existing research on their link to stress

Over the past several decades, researchers have sought to understand the impact of stress on speech and to identify specific speech features that change under stress. In this section, we will provide an overview of the speech features that have been explored in the literature (i.e., Fundamental Frequency, Jitter, Shimmer, Harmonics-to-noise Ratio, and Speech Rate), explaining their significance and how they are affected by stress. This specific set of speech features has been chosen due to multiple reasons. First, they have been described most commonly in the context of stress research, as summarized in reviews by Giddens and colleagues (2013) and Van Puyvelde and colleagues (2018). Second, these features are understandable and interpretable. Other features presented in the GeMAPS configuration of

OpenSMILE, a configuration specifically designed to contain features that are useful for emotion research, are harder to interpret and link back to physiological phenomena (Eyben et al., 2015). Lastly, we do generate other features in our paradigms and make these data publicly available for other researchers to test on, but as a matter of focus and interpretability, we exclusively describe and performed data analyses on the aforementioned set of speech features. Additionally, we will discuss the limitations of these findings, focusing on sample characteristics and differences in the stressors employed in the studies. A broader, overarching examination of the limitations in the literature will be presented in [section 1.4.3](#).

Before delving into specific speech features, it is important to understand the distinction between voiced and unvoiced speech, as these two types of sounds form the basis of human speech production. **Voiced speech** refers to the production of speech sounds where the vocal folds vibrate, creating bursts of air. Air from the lungs is modulated, which produces sound and sets the pitch of the voice. Examples of voiced speech include vowels and some consonants, such as ‘b’, ‘d’, and ‘g’. **Unvoiced speech**, on the other hand, involves speech sounds produced without the vibration of the vocal folds. Instead, unvoiced speech sounds are caused by constriction or closure of the vocal tract articulators, such as the tongue, lips, and the glottis, creating noise-like sounds. Examples of unvoiced speech sounds include consonants such as ‘t’, ‘k’, and ‘s’.

1.4.2.1. Fundamental Frequency (F0)

Fundamental Frequency (F0) refers to the frequency of the vocal fold vibrations, which determines the perceived pitch of the voice. When the vocal folds vibrate as air passes through

them, they generate a complex sound wave with multiple harmonics. The F0 corresponds to the rate at which these vibrations occur and forms the basis for the harmonic structure of the voice. Literature reviews concluded that F0 shows a consistent increase in the context of stress, especially when using well-controlled experimental stress induction procedures (Giddens et al., 2013; Van Puyvelde et al., 2018). For example, Sigmund collected read-out-loud speech samples from 31 students before and after a highly stressful oral exam and observed increases in F0 with stress (Sigmund, 2006). Increases in F0 are the most consistent changes in speech features observed and will play a key role in stress detection from speech samples.

1.4.2.2. Jitter

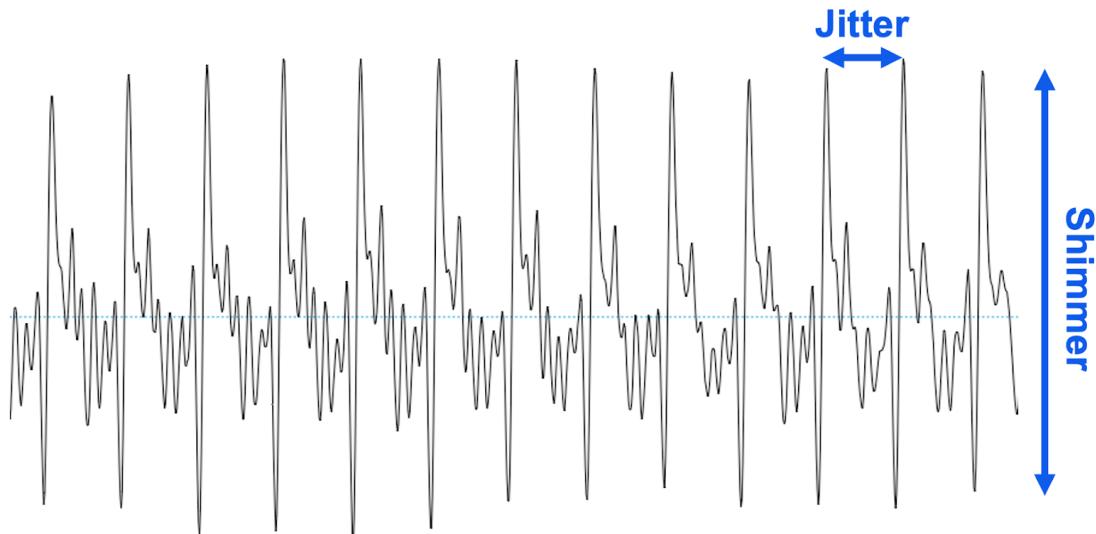
Jitter refers to the small, random fluctuations in the F0 of the voice over time (pitch variation). These variations occur naturally due to the inherent instability in the vocal fold vibrations during phonation. Jitter is often used as an acoustic measure of voice quality and can be indicative of vocal fold irregularities or vocal fatigue. In general, a higher level of jitter is associated with more rough or hoarse voice quality, while a lower level of jitter corresponds to a more stable and clear voice. In the context of stress, decreases in jitter are sometimes reported (Giddens et al., 2013; Mendoza & Carballo, 1998; Van Puyvelde et al., 2018). For example, in a study by Mendoza and Carballo, 82 students participated in three cognitively challenging tasks such as tongue-twisters and backward reading of the alphabet. In addition to increases in F0, they observed decreased jitter values (Mendoza & Carballo, 1998).

1.4.2.3. Shimmer

Shimmer refers to the fluctuations in the amplitude of the voice over time (loudness variation). It is a measure of the inconsistency in vocal fold vibrations, which can be caused by factors such as irregularities in vocal fold tension, mass, or subglottal pressure. Shimmer is often used as an acoustic measure of voice quality and can be indicative of vocal fold irregularities or vocal fatigue. Generally, a higher level of shimmer is associated with more breathy or weak voice quality, while a lower level of shimmer corresponds to a more stable and clear voice (J. Clark et al., 2007; Titze & Martin, 1998; Van Puyvelde et al., 2018). As such, jitter and shimmer are often described together. In the context of stress, decreases in shimmer are observed, although findings are mixed (Giddens et al., 2013; Van Puyvelde et al., 2018). Moreover, shimmer is often studied in the context of combined cognitive and emotional load, but research on the isolated effects of emotional load is limited (Van Puyvelde et al., 2018). For instance, in the aforementioned study by Mendoza and Carballo, decreases in shimmer were also observed during the cognitively challenging tasks (Mendoza & Carballo, 1998). In addition, Orlikoff has linked variation in cardiac amplitude directly to changes in shimmer, stating that it accounts for 11.8% of its variability (Orlikoff, 1990).

Figure 2

Image displaying the dimension along which variation occurs for the measures jitter (frequency variation) and shimmer (amplitude variation)



1.4.2.4. Harmonics-to-noise Ratio (HNR)

Harmonics-to-noise Ratio (HNR) is an acoustic measure used to quantify the balance between the harmonic and noise components in human speech. HNR is an indicator of voice quality, with higher values representing a clearer, more harmonic voice, and lower values suggesting the presence of more noise or breathiness, as also found in breaky, pathological, and whispering voices (J. Clark et al., 2007; Giddens et al., 2013; Van Puyvelde et al., 2018). HNR has been shown to change in people experiencing psychological stress (or when combined with cognitive load), yet results are mixed in their direction (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Mendoza & Carballo, 1998; Van Puyvelde et al., 2018). HNR does show consistent decreases in the context of physical stressors (any physical

event or stimulus that elicits stress), indicating that it is a potentially responsive feature to physical changes, but needs more research to gain clarity on its process, sensitivity, and direction (Giddens et al., 2013).

1.4.2.5. Speech rate

Speech rate, or tempo, is the speed at which a speaker produces words or syllables in a given time period, typically measured in words per minute or syllables per second. However, in order to compute either of these metrics, one would require the exact spoken words, which is a labor-intensive task considering for instance transcriptions. Therefore, a proxy for speech rate is often calculated; **Mean Voiced Segments per Second** (MVSPS). MVSPS serves as a proxy for syllable rate because it quantifies the density of voiced speech (see '*Voiced and Unvoiced speech*') production over time. Since syllables typically contain at least one voiced segment (i.e., a vowel or voiced consonant), MVSPS offers an estimate of the rate at which syllables are produced in a speech sample (J. Clark et al., 2007; Eyben et al., 2015; Titze & Martin, 1998).

It is an important parameter of speech prosody and can be influenced by factors such as linguistic context, cognitive load, and emotional state. In the context of stress, research has shown that speech rate often increases. However, results are inconclusive for isolated emotional load, as many former studies used different types of stressors such as speech samples from emergency phone calls, a Stroop task, and cold pressor tasks, and many were missing physiological and psychological reference measures to validate speakers' stress levels (Giddens et al., 2013; Rothkrantz et al., 2004; Scherer et al., 2002; Van Puyvelde et al., 2018).

For instance, Scherer and colleagues conducted multiple alterations of the same computerized task, distinguishing psychological and cognitive stressors. They observed an increased speech rate in the cognitive load condition, but not in the emotional load condition (Scherer et al., 2002). Despite the current inconsistencies, the potential of speech rate as a valuable indicator of stress should not be underestimated. With more targeted research and refined methodologies, analyzing speech rate could provide a powerful tool for detecting and assessing stress in various contexts and situations.

In addition to MVSPS, we will also discuss **Mean Voiced Segment Length** (MVSL) in the current dissertation. MVSL is the average duration of continuously voiced sounds, without any intervening unvoiced speech sounds or pauses. Therefore, MVSL can provide information about the overall rhythm and pacing of one's speech and can be influenced by factors like speaking style and emotional state. This measure has not frequently been reported upon, however, it could yield unique insights into the prosodic aspects of speech under stress (see: Eyben et al., 2010, 2015).

1.4.3. Limitations of current literature

Throughout the existing literature on the effects of stress on speech, there has been considerable heterogeneity in the observed results, which can be attributed to various limitations. This dissertation aims to address these limitations by conducting multiple studies, iteratively expanding the targeted limitations, which will help understand the interplay between speech parameters under stress and yield more reliable, valid, and generalizable results. Several limitations, as described in the literature (Giddens et al., 2013; Slavich et al., 2019; Van

Puyvelde et al., 2018), include **1)** small sample sizes; **2)** the use of vocal actors rather than participants experiencing genuine psychological stress, which affects the ecological validity of the findings; **3)** a lack of (high-quality) stress labels and verification of experienced stress (e.g., through physiological markers and self-reports); **4)** a lack of within-participant designs, therefore ignoring the influence of individual differences in stress reactivity on speech features; **5)** a lack of (active) control/neutral condition recordings to compare with stress conditions; **6)** A lack of (within-study) consistency in the testing environment and microphone (i.e., signal quality); **7)** using massive sets of speech features that include features lacking scientific basis; **8)** a lack of understanding of the complex interplay between speech parameters and how they are affected by stress, and **9)** a lack of diversity in stressors, generalizing results to a general stress term.

Despite these limitations in this relatively young field, some speech features have shown a clear physiological basis for their connection to stress. Moreover, interesting effects have been observed, yet due to the aforementioned limitations, require validation. To address these limitations, this dissertation presents multiple empirical studies. Each study targets points 1 through 7, but some points are targeted in different ways. In the Objectives, an overview of all chapters is presented, summarizing their setup and what they add to the literature.

1.5. Objectives

The main objective of this dissertation is straightforward; to identify the potential of speech as a measure of stress. By evaluating the physiological basis of speech production and assessing the similarities in systems activated during a stress response, it appears a promising avenue, yet current results are heterogeneous. In order to generate results and indices of speech that are valid, trustworthy, and generalizable in the context of stress, we conducted a series of studies that target these limitations. The results from this dissertation will serve as a solid basis for this novel method for other researchers to build upon.

All studies target limitations 1 through 7, by 1) using large, statistically-powered samples, 2) eliciting stress in actual, non-actor participants, 3) validating stress inductions using self-reports and/or physiological responses, 4) exclusively using within-participant designs to allow for individual differences, 5) including a neutral speech recording (chapter 2) or speech recordings from an active control condition (chapter 3 & 5), 6) ensuring consistency in the testing environment and using exclusively high-quality microphones, ensuring the quality of recordings and experimental setup, and 7) exclusively including speech features that have been described previously in the literature.

The main objective of **Chapter 2** is to investigate the unique associations between the selected speech parameters, from read-out-loud speech, before and after stress induction, and examine how changes in these parameters relate to changes in self-reported negative affect. The study aims to apply network modeling to explore the complex interplay between speech

parameters in the context of psychosocial stress in a well-controlled but ecologically valid setting, directly targeting limitation number 8.

The primary goal of **Chapter 3** is to investigate the effects of stress on the selected speech features in a controlled, within-subject psychosocial stress induction experiment. The study also uses read-out-loud speech, and its main addition is the use of an active control condition. This enables us to generate results directly related to the isolated, added effect of negative comparison to a cognitively tasking, active control task, further targeting limitation number 5.

Chapter 4's main objective is to develop and validate the Ghent Semi-spontaneous Speech Paradigm (GSSP), a new method for acquiring semi-guided (unscripted) speech data for affective-behavioral research in both experimental and real-world settings. The study evaluates the validity of the GSSP through an online task and acoustic analysis, with the aim of providing a valuable tool for capturing spontaneous speech. This tool moves the entire field forward, as it directly enables our main objective of reliable, valid, and generalizable results.

The primary aim of **Chapter 5** is to explore speech features in semi-guided speech following two distinct psychosocial stress paradigms (Cyberball and MIST) and their respective active control conditions. The study investigates whether observed speech features are robust in semi-guided speech and sensitive to stressors eliciting additional physiological stress responses, not solely increases in negative affect. This study directly targets limitation number 9, while considering limitations 4 and 5, by presenting a multi-day, multi-paradigm, within-participant design, yielding unique insights into the heterogeneity of results originating from the use of different stressors.

The main objective of **Chapter 6** is to discuss the potential of speech as a novel digital biosignal for predicting high-priority clinical outcomes and delivering tailored interventions. By reviewing existing tools for extracting health-relevant biosignals from smartphones by analyzing a person's voice and speech, the chapter emphasizes the importance of speech as a tool for measuring stress and predicting clinical outcomes in healthcare and research settings.

Together, these studies aim to revolutionize and advance the field of stress research by comprehensively examining the potential of speech as a reliable, valid, and non-intrusive measure of stress. By systematically addressing the identified limitations in the current literature, this dissertation seeks to provide a solid foundation for future research in stress detection through speech. By exploring the complex interplay between speech parameters, developing novel methods for obtaining semi-spontaneous speech, and investigating the robustness and sensitivity of speech features across diverse stressors, this body of work strives to establish speech as a valuable biomarker for stress. Ultimately, these studies contribute to the development of innovative tools for stress detection and prediction, which can be easily integrated into everyday technology such as smartphones, enhancing healthcare and research settings by offering accessible, affordable, and unobtrusive stress measurement solutions.

1.6. References

- Abbas, A., Schultebraucks, K., & Galatzer-Levy, I. R. (2021). Digital Measurement of Mental Health: Challenges, Promises, and Future Directions. *Psychiatric Annals*, 51(1), 14–20. <https://doi.org/10.3928/00485713-20201207-01>
- Arza, A., Garzón-Rey, J. M., Lázaro, J., Gil, E., Lopez-Anton, R., de la Camara, C., Laguna, P., Bailon, R., & Aguiló, J. (2019). Measuring acute stress response through physiological signals: Towards a quantitative assessment of stress. *Medical & Biological Engineering & Computing*, 57(1), 271–287. <https://doi.org/10.1007/s11517-018-1879-z>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (n.d.). *Autonomic Determinism: The Modes of Autonomic Control, the Doctrine of Autonomic Space, and the Laws of Autonomic Constraint*.
- Bhushan, D., Kotz, K., McCall, J., Wirtz, S., Gilgoff, R., Rishi Dube, S., Powers, C., Olson-Morgan, J., Galeste, M., Patterson, K., Harris, L., Mills, A., Bethell, C., & Burke Harris, N. (2020). *The Roadmap for Resilience: The California Surgeon General's Report on Adverse Childhood Experiences, Toxic Stress, and Health*. Office of the California Surgeon General. <https://doi.org/10.48019/PEAM8812>
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)

Braithwaite, J., Watson, D., Jones, R., & Rowe, M. A. (2013). Guide for Analysing Electrodermal Activity & Skin Conductance Responses for Psychological Experiments. *CTIT Technical Reports Series*.

<https://www.semanticscholar.org/paper/Guide-for-Analysing-Electrodermal-Activity-%26-Skin-Braithwaite-Watson/b99d1f004e4194ac6ef86a86bb0918a11152a01e>

Braun, K. (2011). The Prefrontal-Limbic System: Development, Neuroanatomy, Function, and Implications for Socioemotional Development. *Clinics in Perinatology*, 38(4), 685–702.

<https://doi.org/10.1016/j.clp.2011.08.013>

Brosschot, J. F., Verkuil, B., & Thayer, J. F. (2017). Exposed to events that never happen: Generalized unsafety, the default stress response, and prolonged autonomic activity. *Neuroscience & Biobehavioral Reviews*, 74, 287–296.

<https://doi.org/10.1016/j.neubiorev.2016.07.019>

Câmara, R., & Griessenauer, C. J. (2015). Chapter 27—Anatomy of the Vagus Nerve. In R. S. Tubbs, E. Rizk, M. M. Shoja, M. Loukas, N. Barbaro, & R. J. Spinner (Eds.), *Nerves and Nerve Injuries* (pp. 385–397). Academic Press.

<https://doi.org/10.1016/B978-0-12-410390-0.00028-7>

Camm, A. J., Malik, M., Bigger, J. T., Breithardt, G., Cerutti, S., Cohen, R. J., & Singer, D. H. (1996). Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation*, 93(5), 1043–1065.

<https://doi.org/10.1161/01.CIR.93.5.1043>

Campbell, J., & Ehlert, U. (2012). Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology*, 37(8),

1111–1134. <https://doi.org/10.1016/j.psyneuen.2011.12.010>

Cannon, W. B. (1939). *The wisdom of the body*, 2nd ed. Norton & Co.

Carlson, N. R., & Birkett, M. A. (2017). *Physiology of Behavior* (12th ed.). Pearson Education Limited.

Carpenter, R. (2016). A Review of Instruments on Cognitive Appraisal of Stress. *Archives of Psychiatric Nursing*, 30(2), 271–279. <https://doi.org/10.1016/j.apnu.2015.07.002>

Clark, J., Yallop, C., & Fletcher, J. (2007). *An Introduction to Phonetics and Phonology—3rd edition* (3rd ed.). Blackwell Publishing.

Clark, L. T., Watkins, L., Piña, I. L., Elmer, M., Akinboboye, O., Gorham, M., Jamerson, B., McCullough, C., Pierre, C., Polis, A. B., Puckrein, G., & Regnante, J. M. (2019). Increasing Diversity in Clinical Trials: Overcoming Critical Barriers. *Current Problems in Cardiology*, 44(5), 148–172. <https://doi.org/10.1016/j.cpcardiol.2018.11.002>

Cohen, S., Janicki-Deverts, D., & Miller, G. E. (2007). Psychological stress and disease. *Journal of the American Medical Association*, 298(14), 1685–1687.

<https://doi.org/10.1001/jama.298.14.1685>

Cohen, S., Kamarck, T., & Mermelstein, R. (1994). Perceived Stress Scale (PSS). *Measuring Stress: A Guide for Health and Social Scientists*.

https://doi.org/10.1007/978-1-4419-1005-9_773

Cohen, S., Kessler, R. C., & Gordon, L. U. (1997). *Measuring stress: A guide for health and social scientists*. Oxford University Press on Demand.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2017). The electrodermal system. In *Handbook of psychophysiology*, 4th ed (pp. 217–243). Cambridge University Press.

Dohrenwend, B. P. (2006). Inventorying Stressful Life Events as Risk Factors for

Psychopathology: Toward Resolution of the Problem of Intracategory Variability.
Psychological Bulletin, 132(3), 477–495. <https://doi.org/10.1037/0033-2909.132.3.477>

Duffy, J. R. (2000). Motor Speech Disorders: Clues to Neurologic Diagnosis. In C. H. Adler & J. E. Ahlskog (Eds.), *Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physician* (pp. 35–53). Humana Press.
https://doi.org/10.1007/978-1-59259-410-8_2

Dupre, M. E., George, L. K., Liu, G., & Peterson, E. D. (2015). Association Between Divorce and Risks for Acute Myocardial Infarction. *Circulation: Cardiovascular Quality and Outcomes*, 8(3), 244–251. <https://doi.org/10.1161/CIRCOUTCOMES.114.001291>
Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, 49(December 2017), 146–169.
<https://doi.org/10.1016/j.yfrne.2018.03.001>

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.
<https://doi.org/10.1109/TAFFC.2015.2457417>

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE - The Munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 1459–1462. <https://doi.org/10.1145/1873951.1874246>
Folkman, S., & Moskowitz, J. T. (2004). Coping: Pitfalls and Promise. *Annual Review of Psychology*, 55(1), 745–774. <https://doi.org/10.1146/annurev.psych.55.090902.141456>

- Gaab, J., Rohleder, N., Nater, U. M., & Ehlert, U. (2005). Psychological determinants of the cortisol stress response: The role of anticipatory cognitive appraisal. *Psychoneuroendocrinology*, 30(6), 599–610. <https://doi.org/10.1016/j.psyneuen.2005.02.001>
- Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, 13(1), 440–460. <https://doi.org/10.1109/TAFFC.2019.2927337>
- 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.e21-390.e29. <https://doi.org/10.1016/j.jvoice.2012.12.010>
- Godin, K. W., & Hansen, J. H. L. (2015). Physical task stress and speaker variability in voice quality. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015(1). <https://doi.org/10.1186/s13636-015-0072-7>
- Godin, K. W., Hasan, T., & Hansen, J. H. L. (2012). Glottal waveform analysis of physical task stress speech. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2(January 2012), 1646–1649.
- Gunther Moor, B., Crone, E. A., & van der Molen, M. W. (2010). The Heartbreak of Social Rejection: Heart Rate Deceleration in Response to Unexpected Peer Rejection. *Psychological Science*, 21(9), 1326–1333. <https://doi.org/10.1177/0956797610379236>
- 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. https://doi.org/10.1007/978-3-540-74200-5_6
- Hobfoll, S. E. (2004). *Stress, Culture, and Community: The Psychology and Philosophy of*

Stress. Springer Science & Business Media.

Joëls, M., & Baram, T. Z. (2009). The neuro-symphony of stress. *Nature Reviews Neuroscience*, 10(6), Article 6. <https://doi.org/10.1038/nrn2632>

Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience & Biobehavioral Reviews*, 26(2), 235–258. [https://doi.org/10.1016/S0149-7634\(01\)00068-9](https://doi.org/10.1016/S0149-7634(01)00068-9)

Juster, R.-P., McEwen, B. S., & Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, 35(1), 2–16. <https://doi.org/10.1016/j.neubiorev.2009.10.002>

Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. (2023). Speech as a promising biosignal in precision psychiatry. *Neuroscience & Biobehavioral Reviews*, 148, 105121. <https://doi.org/10.1016/j.neubiorev.2023.105121>

Kirschbaum, C., & Hellhammer, D. H. (1994). Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology*, 19(4), 313–333. <https://doi.org/10.10111/j.0269-8463.2004.00893.x>

Klabunde, R. (2011). *Cardiovascular Physiology Concepts*. Lippincott Williams & Wilkins.

Kogler, L., Mueller, V. I., Chang, A., Eickhoff, S. B., Peter, T., Gur, R. C., & Derntl, B. (2015). Psychosocial versus physiological stress—Meta-analyses on deactivations and activations of the neural correlates of stress reactions. *Neuroimage*, 119, 235–251. <https://doi.org/10.1016/j.neuroimage.2015.06.059>. Psychosocial

Kyrou, I., & Tsigos, C. (2009). Stress hormones: Physiological stress and regulation of metabolism. *Current Opinion in Pharmacology*, 9(6), 787–793. <https://doi.org/10.1016/j.coph.2009.08.007>

Lazarus, R. S., & Folkman, S. (1984). Stress, appraisal, and coping. In *Spirit and Capital in an*

Age of Inequality. Springer publishing company.

<https://doi.org/10.4324/9781315413532>

Lee, D. Y., Kim, E., & Choi, M. H. (2015). Technical and clinical aspects of cortisol as a biochemical marker of chronic stress. *BMB Reports*, 48(4), 209–216.

<https://doi.org/10.5483/BMBRep.2015.48.4.275>

Lester, N., Smart, L., & Baum, A. (1994). Measuring coping flexibility. *Psychology & Health*, 9(6), 409–424. <https://doi.org/10.1080/08870449408407468>

Lin, E. H., Carter, W. B., & Kleinman, A. M. (1985). An exploration of somatization among Asian refugees and immigrants in primary care. *American Journal of Public Health*, 75(9), 1080–1084. <https://doi.org/10.2105/AJPH.75.9.1080>

Malik, M., & Camm, A. J. (1990). Heart rate variability. *Clinical Cardiology*, 13(8), 570–576.
<https://doi.org/10.1002/clc.4960130811>

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The Tie That Binds? Coherence Among Emotion Experience, Behavior, and Physiology. *Emotion*, 5(2), 175–190. <https://doi.org/10.1037/1528-3542.5.2.175>

McKlveen, J. M., Myers, B., & Herman, J. P. (2015). The Medial Prefrontal Cortex: Coordinator of Autonomic, Neuroendocrine and Behavioural Responses to Stress. *Journal of Neuroendocrinology*, 27(6), 446–456. <https://doi.org/10.1111/jne.12272>

McNair, D. M., Lorr, M., & Droppleman, L. M. (1971). *Manual profile of mood states*.

Melchior, M., Caspi, A., Milne, B. J., Danese, A., Poulton, R., & Moffitt, T. E. (2007). Work stress precipitates depression and anxiety in young, working women and men. *Psychological Medicine*, 37(8), 1119–1129. <https://doi.org/10.1017/S0033291707000414>

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of

cognitive workload tasks. *Journal of Voice*, 12(3), 263–273.

[https://doi.org/10.1016/S0892-1997\(98\)80017-9](https://doi.org/10.1016/S0892-1997(98)80017-9)

Monkhouse, S. (2005). *Cranial Nerves: Functional Anatomy*. Cambridge University Press.

Monroe, S. M. (2008). Modern Approaches to Conceptualizing and Measuring Human Life Stress. *Annual Review of Clinical Psychology*, 4(1), 33–52.

<https://doi.org/10.1146/annurev.clinpsy.4.022007.141207>

Monroe, S. M., & Slavich, G. M. (2016). Psychological stressors: Overview. In *Stress: Concepts, cognition, emotion, and behavior* (pp. 109–115).

Myers, B., Scheimann, J. R., Franco-Villanueva, A., & Herman, J. P. (2017). Ascending mechanisms of stress integration: Implications for brainstem regulation of neuroendocrine and behavioral stress responses. *Neuroscience & Biobehavioral Reviews*, 74, 366–375. <https://doi.org/10.1016/j.neubiorev.2016.05.011>

Orlikoff, R. F. (1990). Vowel amplitude variation associated with the heart cycle. *Journal of the Acoustical Society of America*, 88(5), 2091–2098. <https://doi.org/10.1121/1.400106>

Phelan, J., Schwartz, J. E., Bromet, E. J., Dew, M. A., Parkinson, D. K., Schulberg, H. C., Dunn, L. O., Blane, H., & Curtis, E. C. (1991). Work stress, family stress and depression in professional and managerial employees. *Psychological Medicine*, 21(4), 999–1012.

<https://doi.org/10.1017/S0033291700029998>

Razavi, T. (2001). *Self-report measures: An overview of concerns and limitations of questionnaire use in occupational stress research*.

Rehkopf, D. H., Kuper, H., & Marmot, M. G. (2010). Discrepancy between objective and subjective measures of job stress and sickness absence. *Scandinavian Journal of Work, Environment & Health*, 36(6), 449–457.

- Rossi, V., & Pourtois, G. (2012). Transient state-dependent fluctuations in anxiety measured using STAI, POMS, PANAS or VAS: A comparative review. *Anxiety, Stress and Coping*, 25(6), 603–645. <https://doi.org/10.1080/10615806.2011.582948>
- Rothkrantz, L. J. M., Wiggers, P., Van Wees, J. W. A., & Van Vark, R. J. (2004). Voice stress analysis. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3206, 449–456. <https://doi.org/10.4135/9781452229300.n1969>
- Ryder, A. G., & Chentsova-Dutton, Y. E. (2012). Depression in Cultural Context: “Chinese Somatization,” Revisited. *Psychiatric Clinics of North America*, 35(1), 15–36. <https://doi.org/10.1016/j.psc.2011.11.006>
- Scherer, K. R., Grandjean, D., Johnstone, T., Klasmeyer, G., & Bänziger, T. (2002). Acoustic correlates of task load and stress. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2017–2020. <https://doi.org/10.21437/ICSLP.2002-554>
- Sherwood, L. (2003). *Human physiology: From cells to systems*. 4th ed. West Pub. Co.,. http://repository.vnu.edu.vn/handle/VNU_123/85384
- Shields, G. S., & Slavich, G. M. (2017). Lifetime stress exposure and health: A review of contemporary assessment methods and biological mechanisms. *Social and Personality Psychology Compass*, 11(8), 1–17. <https://doi.org/10.1111/spc3.12335>
- Shields, S., MacDowell, K., Fairchild, S., & Campbell, M. (1987). Is Mediation of Sweating Cholinergic, Adrenergic, or Both? A Comment on the Literature. *Psychophysiology*, 24, 312–319. <https://doi.org/10.1111/j.1469-8986.1987.tb00301.x>
- Sigmund, M. (2006). Introducing the Database ExamStress for Speech under Stress. *Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006*, 290–293. <https://doi.org/10.1109/NORSIG.2006.275258>

- Slavich, G. M., & Irwin, M. R. (2014). From Stress to Inflammation and Major Depressive Disorder: A Social Signal Transduction Theory of Depression. *Psychological Bulletin*, 140(3), 774–815. <https://doi.org/10.1037/a0035302>
- Slavich, G. M., Taylor, S., Picard, R. W., Slavich, G. M., Taylor, S., & Stress, R. W. P. (2019). *Stress measurement using speech: Recent advancements , validation issues , and ethical and privacy considerations*. 3890. <https://doi.org/10.1080/10253890.2019.1584180>
- Sondhi, S., Khan, M., Vijay, R., & K. Salhan, A. (2015). Vocal Indicators of Emotional Stress. *International Journal of Computer Applications*, 122(15), 38–43. <https://doi.org/10.5120/21780-5056>
- Sousa, N., & Almeida, O. F. X. (2012). Disconnection and reconnection: The morphological basis of (mal)adaptation to stress. *Trends in Neurosciences*, 35(12), 742–751. <https://doi.org/10.1016/j.tins.2012.08.006>
- Tennant, C. (2001). Work-related stress and depressive disorders. *Journal of Psychosomatic Research*, 51(5), 697–704. [https://doi.org/10.1016/S0022-3999\(01\)00255-0](https://doi.org/10.1016/S0022-3999(01)00255-0)
- Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 61(3), 201–216. [https://doi.org/10.1016/S0165-0327\(00\)00338-4](https://doi.org/10.1016/S0165-0327(00)00338-4)
- Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, 33(2), 81–88. <https://doi.org/10.1016/j.neubiorev.2008.08.004>
- Titze, I. R., & Martin, D. W. (1998). Principles of Voice Production. *The Journal of the Acoustical Society of America*, 104(3), 1148–1148. <https://doi.org/10.1121/1.424266>

Ulrich-Lai, Y. M., & Herman, J. P. (2009). Neural Regulation of Endocrine and Autonomic Stress Responses. *Nature Reviews. Neuroscience*, 10(6), 397–409.

<https://doi.org/10.1038/nrn2647>

van der Veen, F. M., Burdzina, A., & Langeslag, S. J. E. (2019). Don't you want me, baby? Cardiac and electrocortical concomitants of romantic interest and rejection. *Biological Psychology*, 146, 107707. <https://doi.org/10.1016/j.biopsych.2019.05.007>

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>

Welte, J. W., & Russell, M. (1993). Influence of Socially Desirable Responding in a Study of Stress and Substance Abuse. *Alcoholism: Clinical and Experimental Research*, 17(4), 758–761. <https://doi.org/10.1111/j.1530-0277.1993.tb00836.x>

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, 58(1), 425–452. <https://doi.org/10.1146/annurev.psych.58.110405.085641>

Zhou, G., Hansen, J. H. L., Member, S., & Kaiser, J. F. (2001). Nonlinear Feature Based Classification of Speech Under Stress. 9(3), 201–216.

Zwolinski, J. (2012). Psychological and Neuroendocrine Reactivity to Ostracism. *Aggressive Behavior*, 38(2), 108–125. <https://doi.org/10.1002/ab.21411>

Speech as an indicator for psychosocial stress: A network analytic approach

Mitchel Kappen^{123*}, Kristof Hoorelbeke³, Nilesh Madhu⁴, Kris Demuynck⁴, Marie-Anne Vanderhasselt¹²

1Department of Head and Skin, Ghent University, University Hospital Ghent (UZ Ghent), Department of Psychiatry and Medical Psychology, Ghent, Belgium

2Ghent Experimental Psychiatry (GHEP) Lab, Ghent University, Ghent, Belgium

3Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

4IDLab, Ghent University - imec, Ghent, Belgium

Published as:

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M. A. (2021). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 1-12.

2.1. Abstract

Recently, the possibilities of detecting psychosocial stress from speech have been discussed. Yet, there are mixed effects and a current lack of clarity in relations and directions for parameters derived from stressed speech. The aim of the current study is - in a controlled psychosocial stress induction experiment - to apply network modeling to (1) look into the unique associations between specific speech parameters, comparing speech networks containing Fundamental frequency (F0), Jitter, Mean Voiced Segment Length, and Harmonics-to-Noise Ratio (HNR) pre- and post-stress induction, and (2) examine how changes pre- versus post-stress induction (i.e., change network) in each of the parameters are related to changes in self-reported negative affect. Results show that the network of speech parameters is similar after versus before the stress induction, with a central role of HNR, which shows that the complex interplay and unique associations between each of the used speech parameters is not impacted by psychosocial stress (aim 1). Moreover, we found a change network (consisting of pre-post stress difference values) with changes in Jitter being positively related to changes in self-reported negative affect (aim 2). These findings illustrate - for the first time in a well-controlled but ecologically valid setting - the complex relations between different speech parameters in the context of psychosocial stress. Longitudinal and experimental studies are required to further investigate these relationships and to test whether the identified paths in the networks are indicative of causal relationships.

2.2. Introduction

Stress is an increasingly relevant topic in modern society, with a majority of people experiencing regular stress symptoms. Considering the broad range of physiological and psychological factors that are influenced by stress, a plethora of methods have been developed to assess individuals' stress levels. Currently, commonly used methods determine stress levels by self-report questionnaires on factors that are affected by stress (e.g., mood) or broader indicators of psychological well-being (Monroe, 2008). Besides self-reports, stress is also commonly assessed through the measurement of biological processes involved with stress exposure. The main advantage of measuring stress through biomarkers, as compared to interviews and self-report instruments, is that these measures are not subject to self-report biases. Furthermore, biomarkers allow continuous monitoring of stress levels. Many different biological markers of stress have been identified such as heart rate, blood pressure, cortisol, skin conductance, and many more (for an extensive overview, see: Fink, 2017; Shields & Slavich, 2017). Even though many of these methods are highly effective in determining one's stress levels, they are often costly, requiring the attachment of electrodes (e.g., Electrocardiography; ECG) or the extraction of a blood or saliva sample, and demanding since they generally require interaction with a physician or expert and specialized apparatus to collect the data. With the emerging market of wearables (e.g., smartwatches), it has become increasingly easy to collect continuous data of stress-related physiological markers such as heart rate, skin conductance, and skin temperature. Although the quality of these methods is constantly improving, it is not always evident to continuously collect this data (e.g., costs, privacy) besides often reported problems with regards to continuity of its accuracy (e.g., loss of

connection). Therefore, the need to further explore alternatives for stress measurements remains. Recently, speech analysis has been proposed as a possible physiological marker for stress, however, further research is required (Giddens et al., 2013; Slavich et al., 2019).

Speech production is a complex process that requires the involvement of many different parts of the body. To produce speech, one first considers what words to say, tone of voice, and many more conscious aspects. However, the practical part happens more automatically, which is the actual sound production. When producing speech, the body modulates the tension of numerous muscles to push air through the vocal folds and out the vocal tract to produce sound waves (Titze & Martin, 1998). Since stress increases both muscle tension and respiration rate, which in turn influence speech production, it has been proposed that stress should be detectable from the way speech sounds (Sondhi et al., 2015). A major advantage of stress detection from speech is the non-intrusive obtainability of speech data and the possibility of swift, cost-effective, and remote stress assessments. As such, speech is considered a promising psychophysiological measure for stress assessment.

However, relatively little is known with regards to how specific speech parameters interact in a context with or without stress, and how this interaction of speech parameters changes following a stressor. That is, speech research in the context of stress is still in its infancy and has mostly flourished at the fundamental level of parameter identification and development. Speech consists of many different parameters (i.e., characteristics), which are contingent on many factors, both conscious and automatic. Fundamental frequency, Harmonics-to-Noise Ratio, and Jitter are such speech parameters that have been found to change in stressed subjects (Giddens et al., 2013; Kreiman & Sadtis, 2011; Mendoza &

Carballo, 1998; Orlikoff, 1990; Orlikoff & Baken, 1989). Based on the available literature, 1) Fundamental frequency (F0) can be considered a key speech parameter in the context of different types of stressors. F0 refers to the frequency at which the vocal cords vibrate, and gives rise to the idea of the pitch of the voice. Research suggests a universal trend of increase in F0 in stressed subjects (Giddens et al., 2010, 2013; Godin & Hansen, 2008; Johannes et al., 2007; Koblick, 2004; Kreiman & Sidiis, 2011; Mendoza & Carballo, 1998; Rothkrantz et al., 2004; Williams & Stevens, 1972). Another widely used speech parameter is 2) Harmonics-to-Noise Ratio (HNR), which indicates one's vocal quality by measuring the additive noise in the speech signal during voiced periods (e.g., when uttering vowels). HNR has mainly been studied in physical stress tasks (e.g., workout), and has shown to decrease with increased physical task stress (Godin et al., 2012; Godin & Hansen, 2015; Koblick, 2004), but has shown mixed results in the context of cognitive load/psychological stress (e.g., tongue twister, reciting the alphabet backwards; Mendoza & Carballo, 1998). 3) Jitter refers to the frequency variation from cycle to cycle and has been found to reduce in the context of stress, however this trend has not shown to be universal (Giddens et al., 2013; Mendoza & Carballo, 1998). Moreover, 4) formants have been opted as promising features of speech in distinguishing stress from speech, more specifically, the shifting of formant 1 (F1) and formant 2 (F2) have shown to be decent indicators of psychological stress (Sigmund, 2012; Van Puyvelde et al., 2018). Formants are the primary resonances of the vocal tract and can shift due to numerous conscious and unconscious processes and are dependent on one's speech style (Shahin & Botros, 2001). There has, however, not been consensus on the effects of psychological stress on F1 and F2, which indicates it to be heavily influenced by individual trends rather than global trends valid for all speakers (Kirchhuebel, 2010; Sigmund, 2012).

Since both change in F1 and F2 play a role in stress, a ratio score could be computed that is reactive to changes in either formant; Formant 1:2 Ratio. Lastly, it has been suggested that with increased physical stress, breathing patterns and muscle tension impact different aspects of speech, such as inappropriate pause placements (Van Puyvelde et al., 2018). As a final feature, 5) Mean Voiced Segment Length can be used to gain insight into such speaking patterns as it is the mean length of the continuously voiced regions which is expected to decrease under stress.

Even though research is currently lacking, the combination of these speech parameters is highly promising for the detection and understanding of increased stress. However, it should be noted that each of these parameters reflect unique features of a complex speech production process. Therefore, these parameters are highly interrelated, where the unique interplay between each of these measures remains to be modeled. In particular, little is known regarding the complex interplay between speech parameters and how it is affected by stress (Giddens et al., 2013; Kreiman & Sidtis, 2011). Much of the recent work in stress detection from speech has been conducted in controlled, quiet lab settings or with vocal actors acting out a stressful monologue rather than truly experiencing psychological stress (Giddens et al., 2013), limiting the ecological validity of previous findings. Moreover, it has been suggested that the effect of stress on the formants, which are shifted as a consequence, and Jitter is heavily influenced by individual differences in stress reactivity (Giddens et al., 2013; Scherer, 1986). This is likely to also be the case for other speech parameters and could explain the mixed results observed in the literature. Considering the variety of environments, microphones with different qualities, and interindividual differences in stress expression in speech, a number of researchers have

expressed the need for high-quality studies, using real participants rather than voice actors, to compose large datasets of speech data with high-quality stress labels and recorded in a variety of contexts (Giddens et al., 2013; Slavich et al., 2019). Moreover, many researchers investigated stressful versus non-stressful events in their experimental designs without verifying whether participants truly experienced stress by using physiological markers or inquiries. Lastly, previous research has primarily focused on how the entirety of indicators is associated with stress, without highlighting the complex dynamics amongst the indicators and how each of the speech features are uniquely related to stress.

The current study uses a design that takes the above-mentioned shortcomings into consideration to establish a common ground from which new insights can be developed. Healthy individuals will be instructed to read out loud standardized texts both prior to and after exposure to a highly controlled psychosocial stressor. Psychosocial stressors are often described as one of the most powerful and ecologically valid stressors (Kirschbaum & Hellhammer, 1994). Psychosocial stress is induced in situations of social evaluation, social exclusion, and other situations in which social threat occurs (Dickerson & Kemeny, 2004). The need to be associated with others and to maintain a social-self are core psychological needs (Panksepp, 2003; Tossani, 2013). When one of these needs is threatened, for example when being negatively compared to others, social threat and thus stress is induced (Dickerson & Kemeny, 2004). Social evaluation induces an increased stress response which is expressed in increased electrodermal activity (i.e., skin conductance), subjective (experienced) stress, and negative affect (Dedovic et al., 2009; Dickerson & Kemeny, 2004).

Given that our literature review demonstrates mixed effects for parameters derived from stressed speech (and thus a lack of clarity in their relations and direction), and that the interrelation between each of these constructs in the context of stress (i.e., speech parameters, skin conductance levels, and self-reported mood) remains to be explored, we will make use of psychological network models (Borsboom & Cramer, 2013; Newman, 2010). Network methodology is an increasingly used technique to gain insight into complex relationships in a data-driven manner, allowing mapping how each of the constructs of interest is uniquely related to one another. As such, network models are well-suited to explore whether and how the complex interplay between each of the above-presented core speech parameters is impacted by stress. In addition, network analysis allows us to map how changes in speech due to experimental manipulation of stress relate to changes in negative affect. This study has two main aims: 1) We aim to model the impact of a psychosocial stressor (the Montreal Imaging Stress Task (MIST); Dedovic et al., 2005) on the unique associations between the speech parameters of interest (Fundamental frequency, Jitter, Harmonics-to-Noise Ratio, Formant 1:2 Ratio, and Mean Voiced Segment Length) before and after the stressor (aim 1); 2) we will model how stress induced change in the speech parameters relates to change in the negative affect ratings (measured with VAS) as these analyses will shed light on the unique associations between the change in speech features and negative affect following a psychosocial stressor (aim 2). Given the exploratory data-driven approach and undirected nature of the models, the obtained network models are likely to allow further hypothesis generation, which will be informative for future confirmatory studies.

2.3. Methods

A convenience sample of 148 students (M age = 26.7, SD age = 12.5, 51 women, 97 men) was recruited through flyers, social media, and University of Ghent mailing lists informing them on the duration of the experiment, the possibility to win a 25 euro gift card, and a link to www.vopexperiment.be where participants could plan their session through a youcanbookme synchronization. The study was conducted in accordance with the ethical guidelines of the Faculty of Psychology and Educational Sciences of Ghent University, and all participants gave written consent before participating.

2.3.1. Apparatus and Procedure

Participants were seated in one of two nearly identical rooms in front of a Huawei MediaPad M5 tablet. The task was written in Java using Android Studio. Before any instructions commenced, participants signed the informed consent form. Then, participants were instructed on the procedure, how the tablet and application worked (how to record, etc.), and the cover story (cf. infra) was repeated to minimize the likelihood of the actual purpose of the study being identified. Next, participants were given a smartwatch (Chill+ Band) to put on their dominant hand from which Electrodermal Activity was measured (EDA). In addition, 2 ECG electrodes were placed on the sternum and chest. The participants were informed on the purpose of each of these measures with the cover story of it being used to validate the smartwatch measures. Data quality was inspected shortly before the actual experiment started. Firstly, participants were requested to rate the VAS (pre-baseline, exclusively to familiarize

participants and is not used in analysis). Next, participants were instructed to rest for 5 minutes to ensure they were relaxed and minimize the impact of any events occurring before the experiments (e.g., rushing or nervousness). Following the resting phase, participants were instructed to read-out-loud a 5-sentence piece of text that was the same for all participants and an often-used text in Dutch speech therapy:

“Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij, dus dat duurde wel even. Nu wachten ze tot de trein eraan komt. Het is al vijf over drie, dus het duurt nog vier minuten. Er staan nog veel meer mensen te wachten. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.” From: van de Weijer and Slis (1991)

Participants were instructed that the recording of the speech was to train speech-to-text algorithms to hide the actual purpose but to ensure the accurate pronunciation of the text. Next, once again the VAS sliding scales were answered, providing a baseline measure for NA in a relatively unstressed state (i.e., following the first resting block). After that, the MIST (see Stress induction procedure header) commenced, starting with instructions and 2 minutes of practice trials during which no social comparison was made and without a trial time limit. After the testing phase of the MIST, another speech recording and VAS segment was conducted, which corresponds to the post-stress measurement. The end phase of the experiment consisted of another 5-minute resting block, to prevent participants from leaving the experiment in a stressed state, followed by another block of VAS questions. The experiment was concluded by conducting the Ruminative Response Scale (RRS) and the Depression, Anxiety, and Stress Scale (DASS) in order to get an estimation of the sample

characteristics. At the end the participants were debriefed and informed on the actual purpose of the study. Figure 1 shows a visual representation of the study procedure, exclusively containing the elements relevant for the current manuscript.

Figure 1

Visual representation of the study procedure



2.3.2. Stress induction procedure

In order to induce acute stress in our participants and investigate the effects of stress on networks of speech parameters, we used the Montreal Imaging Stress Task (MIST; Dedovic et al., 2005). This is a sequence of arithmetic questions designed as a stress induction task. To ensure a proper understanding of the task, participants could practice for 2 minutes. Trials consisted of mathematical tasks where the correct answer was situated between 0 and 9. Participants were instructed to answer these trials as quickly as possible using arrow buttons to select the right answer on a number wheel. After the practice block, the actual task started and was performed for 5 minutes. During this task, participants were shown a time limit per trial, which was set at 90% of their average time during the practice block. In addition, time limits were reduced by another 10% when they answered three consecutive trials correctly. Throughout the task, participants were presented a performance indicator showing their performance as compared to the ‘average participant’, which in reality was an unfeasible, fictional benchmark. Participants were instructed that they should not deviate from the average

performance too much, and that if they would the data would be unusable for the purpose of the study. Throughout the task, the experimenter was sitting across from the participant and taking notes. Since the participant is always performing worse than the ‘average participant’, and with the experimenter taking notes and the constant time pressure, stress is induced. As a cover story, participants were told that the study attempts to link biometric signals to quick arithmetic solving skills.

2.3.2.1. Trial

Trials showed a numbered wheel from 0 - 9 on which the participant could select the desired answer using arrow buttons and confirm when ready. Above the numbered wheel, an arithmetic task was presented. The top of the screen showed a red bar which was slowly disappearing, indicating the time left for that specific trial. Another bar was shown representing how well the participant was performing compared to others, which was always negative. After answering, participants were either shown a green overlay saying *correct* or a red overlay saying *incorrect*. If they ran out of time, a red overlay saying *timeout* was presented. The practice trials, which were offered at the beginning of the task, did not have a time limit and did not show a comparison to other participants. These were used to familiarize the participant with the task as well as getting a reference reaction time to calculate the trial time limits in the experimental phase. See supplemental material for screenshots.

2.3.3. Self-report measurement – Negative Affect; NA

To evaluate negative affect (NA), as an indicator of stress, self-reported mood was measured at 4 time points (baseline [T1], pre-stress [T2], post-stress [T3], post-recovery [T4]), by using the 3 NA items of a 7 item mood questionnaire (adopted from the Profile of Mood States (POMS); Rossi & Pourtois, 2012) presented on a sliding scale from 0-100 on different states (angry, tense, dejected). The answers to these VAS are used as a manipulation check for the stress induction procedure. More specifically, we used the items representing negative affect (angry, tense, dejected) allowing a compound score for NA (ranging from 0 – 100) where high scores reflect being in a more negative mood state.

2.3.4. Extraction of speech parameters

Speech parameters were extracted using OpenSmile 2.3.0 (Eyben et al., 2010) and the GeMAPS configuration (Eyben et al., 2015), a parameter set used in voice research and affective computing. Fundamental frequency is the central tendency of the frequency of vibration of the vocal folds during speech, and as such, is closely related to pitch, which is defined as our perception of Fundamental frequency. Jitter is the deviation in the F0 computed across consecutive time segments. Formant 1:2 Ratio is the ratio of the energy of the first formant (F1) to the energy of the second formant (F2). Harmonics-to-Noise Ratio (HNR) is the relation of energy in harmonic components to energy in noise-like components, and lastly, Mean Voiced Segment Length is the average length of continuously voiced regions ($F0 > 0$), thus sounds made while the vocal cords vibrate. For more detailed information on parameter

calculation and extraction procedure, we refer the reader to Eyben et al. (2010) and Eyben et al. (2015) and the Supplemental Material.

2.3.5. Statistical Analyses

The network analyses were conducted in R (for detailed version information of the statistical software and packages used, see supplemental materials). As part of the manipulation check, we fitted generalized linear mixed models (GLMMs) using the ‘lme4’ (Bates et al., 2014) and ‘car’ (Bates et al., 2014; Fox et al., 2012) packages. The sum of squares for the models was estimated using the type III approach, and the statistical significance level was set to $p < .05$. Follow-up tests with pairwise comparisons of the estimated marginal means (EMMs) were performed with the ‘emmeans’ R Package (Lenth, 2018).

We relied on Gaussian Graphical Models (GGMs), also referred to as regularized partial correlation networks, to model the impact of stress on the unique associations between the speech parameters of interest (Fundamental frequency, Jitter, Harmonics-to-Noise Ratio, Formant 1:2 Ratio, and Mean Voiced Segment Length), as well as the relation between change in speech parameters and change in NA throughout the stress induction procedure. For this purpose, we estimated three separate GGMs. In particular, we computed: (1) a network including each of the speech parameters of interest, assessed following a resting phase (referred to as *resting state network*), (2) a network including the speech parameters, assessed immediately following the stress induction procedure (referred to as *stress network*), and (3) a network including the change scores for each of the speech parameters and the compound measure for NA (referred to as *stress reactivity network*). Change in NA / speech parameters

was computed by subtracting the resting state measure from the post-stressor measure. As such, a positive value reflects an increase in NA / the speech parameters throughout the induction procedure.

2.3.6. Data preparation and network estimation.

To improve normality, all variables underwent nonparanormal transformation using the *huge* package (Zhao et al., 2012), after which the GGMs were estimated using the *qgraph* package (Epskamp et al., 2012). As the name suggests, GGMs or regularized partial correlation networks depict the unique associations (partial correlations) between each of the variables (referred to as “nodes”) included in the analyses. In network models, the unique associations between each of the nodes are referred to as “edges”. However, given that absence of an association between two constructs does not always result in a correlation coefficient of exactly zero, the need arises for a phase of regularization to prevent the inclusion of spurious associations. For this purpose, we relied on the Graphical Least Absolute Shrinkage and Selection Operator (gLASSO; Friedman et al., 2014), which shrinks small associations, likely reflecting spurious / false-positive findings, to zero (similar to multiple comparison corrections, for more information see Friedman et al., 2014 and Epskamp & Fried, 2018 for a tutorial on GGMs including this regularization technique). The model with the best fit was then selected using the Extended Bayesian Information Criterion with hyperparameter $\gamma = 0.5$. This hyperparameter setting errs on the side of parsimony, maximizing model specificity (Epskamp & Fried, 2018). As a result, the obtained network model is less likely to include false-positive associations (for a more detailed discussion of estimation of GGMs, including an extensive

tutorial, see Epskamp & Fried, 2018). To examine which nodes take a more central role in the model, we estimated node Strength centrality. Strength centrality is calculated as the sum of absolute edge weights connected to each node in the model (Costantini et al., 2015). As such, high scores on Strength centrality reflect that the node is more strongly connected. Finally, we used a node-wise regression approach to estimate node predictability, the proportion of variance of each node that is explained by its neighboring nodes (Haslbeck & Fried, 2017). For this purpose, we relied on the *mgm* package (Haslbeck & Waldorp, 2020).

2.3.7. Network visualization.

The network models were plotted with *qgraph*, using a modification of the Fruchterman-Reingold's algorithm (Fruchterman & Reingold, 1991). This algorithm aims to position nodes in the network based on their level of connectivity (but see Jones et al., 2018). Unique associations between nodes are represented by edges. The thickness of each of the edges reflects the strength of the association, whereas the color and type of line (full/dashed) reflects the valence of the edge (blue/full: positive association; red/dashed: negative association). The GGMs are undirected and as such allow no interpretation regarding the direction of effects. To facilitate visual comparison between the resting state and stress network, the layout of these two networks was constrained to be identical (using the average layout of both models). In addition, for these two networks, we plotted the thickness of each of the edges relative to the strongest edge observed over both models. Moreover, for each of the nodes, we present the proportion of explained variance by the neighboring nodes as a pie

chart in the outer ring of the node (node predictability). Strength centrality was standardized to facilitate interpretation.

2.3.8. Network comparison.

To compare the resting state network and stress network, we first correlated the two obtained adjacency matrices. Similarly, we examined how the estimates of Strength centrality obtained for each of the network models correlated, as well as node predictability. We then proceeded with permutation tests for network structure invariance, allowing to test whether the network structures significantly differed, and global strength invariance, testing potential differences between the resting state- and stress network in (overall) strength of connectivity (van Borkulo et al., 2017). For this purpose, we relied on the *NetworkComparisonTest* package (for dependent samples; van Borkulo et al., 2016).

2.3.9. Evaluation of the stability and accuracy of the models.

To evaluate the stability and accuracy of each of the obtained network models, we followed bootstrapping procedures set-out by Epskamp, Borsboom, and Fried (2018). In particular, using the *bootnet* package (Epskamp & Fried, 2015) we modeled sampling variability in edge weights (edge accuracy) and plotted significant differences in edge weights. Furthermore, we evaluated the stability of the indicator of node centrality, modeling the extent to which the order of Strength centrality remained stable in subsets of the data (cf. case-dropping subset bootstrap). To be considered stable, the corresponding correlation stability coefficient should be $\geq .25$ (Epskamp et al., 2018).

2.4. Results

Due to technical malfunctions, all ECG data was unusable and a part of the EDA has not been collected properly for some participants throughout the experiment ($n=32$). Therefore, the EDA data (together with self-reported mood data) that was collected accurately is used to validate the stress induction method, but will not be included in the network analyses ($n=148$).

2.4.1. Manipulation Check

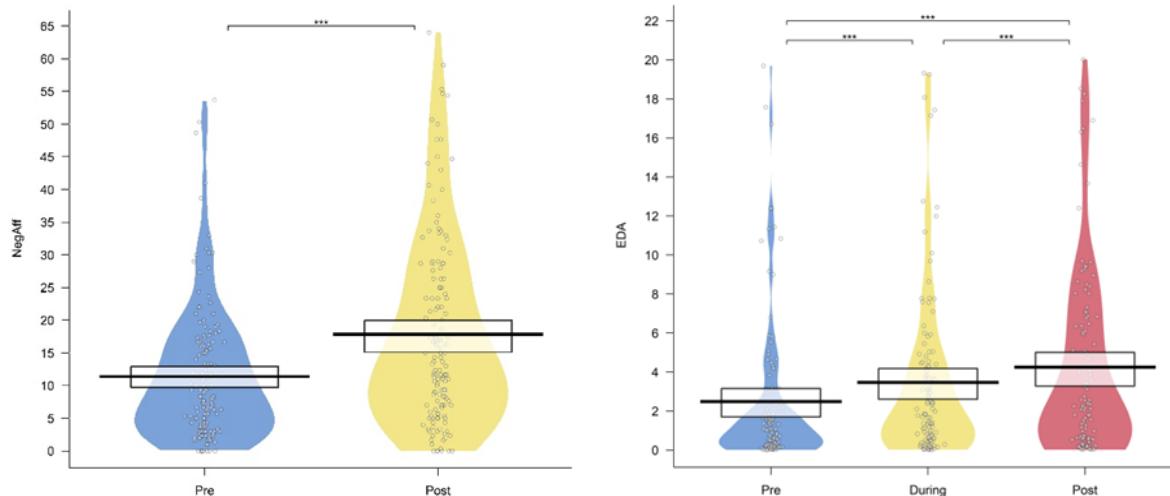
Before the main analysis, a manipulation check was conducted to verify whether the stress induction was successful by comparing both negative affect (NA) pre- and post-stress induction, and EDA pre-, during-, and post-stress induction. Given the non-normality of the EDA data, a series of (G)LMM (Generalized linear mixed models) were conducted to ensure the use of a statistical model that best fits the underlying distribution (e.g., normal, gamma). Based on the Akaike Information Criterion (AIC), EDA was best described by a gamma model with a log-link ($AIC = 802.5$).

Corresponding models were fit with only *time* (pre - post for negative affect (2 levels) and pre - during - post for EDA (3 levels)) as an independent variable and subject ID as random intercept. The LMM for negative affect showed a significant effect of time (see Figure 2a), $p < .001$ with post-stress scores showing significantly more negative affect than pre-stress, $b = 6.45$, $SE = .877$, $t = -7.35$, $p < .001$. Moreover, the GLMM for EDA also showed a significant effect of time (see Figure 2b), $\chi^2 = 247.59$, $p < .001$, with EDA increasing during the task versus prior to the task, $b = .621$, $SE = .029$, $z = -10.29$, $p < .001$, EDA after the task being higher than

during the task $b = .776$, $SE = .035$, $z = -5.53$, $p < .001$ and EDA after the task being higher than prior to the task, $b = .482$, $SE = .023$, $z = -15.513$, $p < .001$.

To further underline the stress induction effectiveness, we ran a Pearson correlation between the delta scores computed for EDA and negative affect. The delta scores were computed by subtracting the data from the resting state measure from the post-stressor measure for these variables, resulting in scores that indicate an increase after the stress induction when positive, and a decrease when negative. A significant correlation, $r(114) = .19$, $p = .04$, was found following the expected trend of negative interrelatedness, therefore supporting the stress-induction method.

Figure 2



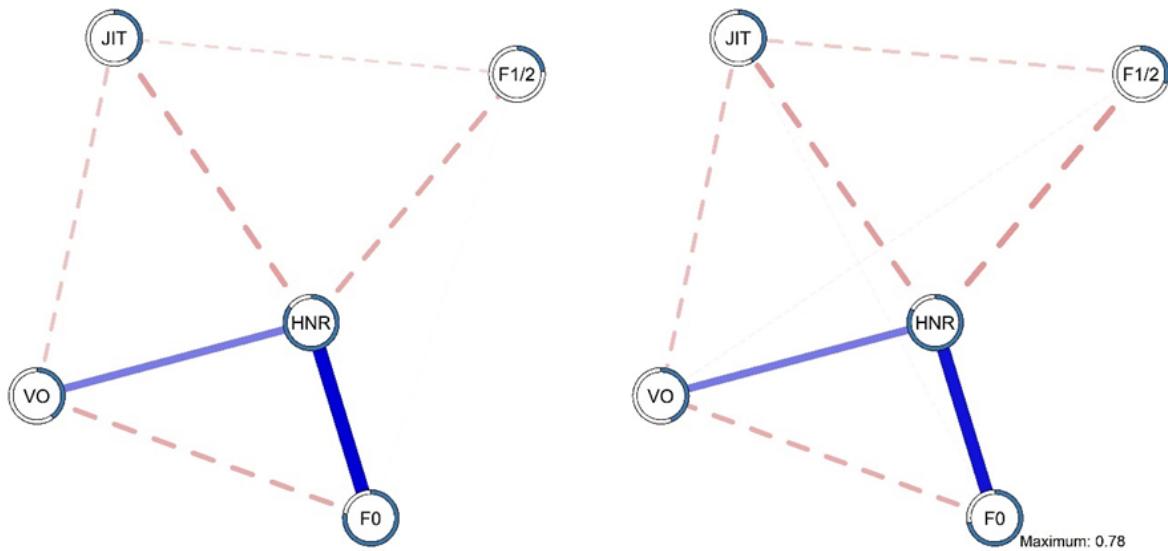
a; left) Negative Affect Pre- and Post-Stress Induction. b; right) EDA Pre-, During, and Post-Stress Induction

2.4.2. Impact of Stress on the Interrelations Between Speech Parameters (aim 1)

Our first aim was to model the impact of stress on the interrelations between the speech parameters of interest by comparing a pre-stressor network with a post-stressor network. These networks consist of nodes representing variables, connected by edges representing regularized partial correlations. As such, every edge (connection) between two nodes (variables) represents the sign (positive/negative) and the weight (strength) of the connection, depicting the unique associations between two nodes while controlling for all other nodes in the network (Epskamp et al., 2018). Figure 3a presents the unique associations between Fundamental frequency (F0), Jitter (JIT), Harmonics-to-Noise Ratio (HNR), Formant 1:2 Ratio (F1/2), and Mean Voiced Segment Length (VO) at rest ($n=148$). The strongest connection in the network occurs between HNR and F0 (.75). HNR is positively associated with F0 and VO. The latter two constructs are negatively associated with one another. In addition, VO and HNR are negatively associated with JIT. For HNR, an additional negative edge emerges with F1/2. Finally, we observe a negative association between JIT and F1/2.

Figure 3

Unique Associations Between the Voice Parameters pre- (a; Left) and Post-Stressor (b; Right)



Note. Edges in the models represent the unique associations between each of the nodes. Edge thickness reflects the strength of association, where strong associations are presented using thicker edges. Blue / Full edges represent positive associations, whereas red / dashed edges represent negative associations; the edge weights presented in the model can also be found in the edge weight matrix (Supplemental tables 6 & 7). Node predictability (R^2) is visualized as a pie chart around each node and can also be found in supplementary table 1.

The pattern of unique interrelations between the speech parameters of interest does not seem to be affected by the stress induction procedure. That is, the network model obtained based on the speech fragments that were collected immediately following the stressor (Figure 2b, n=148), is highly similar to the resting state network (Figure 3a). This is also reflected by the indicator of node centrality (Figure 4), which quantifies how well a node is directly connected to other nodes by adding up the strength of all connected edges to a node (Epskamp et al., 2018). In particular, in terms of node Strength, HNR is the most central node in each of the networks, followed by F0, VO, and JIT. F1/2 is the least connected node in the model. This is also reflected by the amount of explained variance for each of the nodes (i.e., node

predictability). In particular, node predictability of HNR was .85 and .83 in the resting state and stress network respectively, whereas only 24% and 31% of the variance in F1/2 was explained by the neighboring nodes in the resting state and stress network respectively (see table 1 for estimates of node predictability and supplemental material for, edge accuracy, edge differences, and centrality stability).

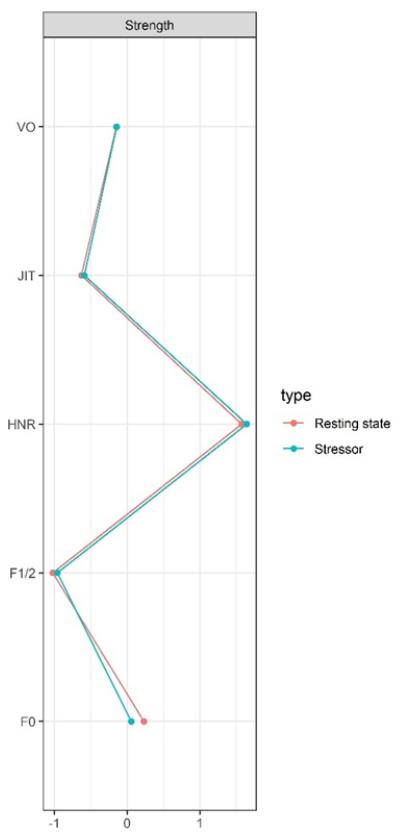
Table 1

Node predictability for Pre-Stressor network (aim 1), Post-Stressor network (aim 1), and Change Network (aim 2)

Node	R ² Pre-Stressor network	R ² Post-Stressor network	Change Network
F0	.78	.72	.16
HNR	.85	.83	.37
JIT	.41	.42	.35
VO	.40	.44	.19
F1/2	.24	.31	.12
NA			.02

Note. R² is explained variance. F0 is Fundamental Frequency. HNR is Harmonics-to-Noise Ratio. JIT is jitter. VO is Mean Voiced Segment Length. F ½ is Formant 1:2 Ratio. NA is Negative Affect.

Figure 4
Strength Centrality



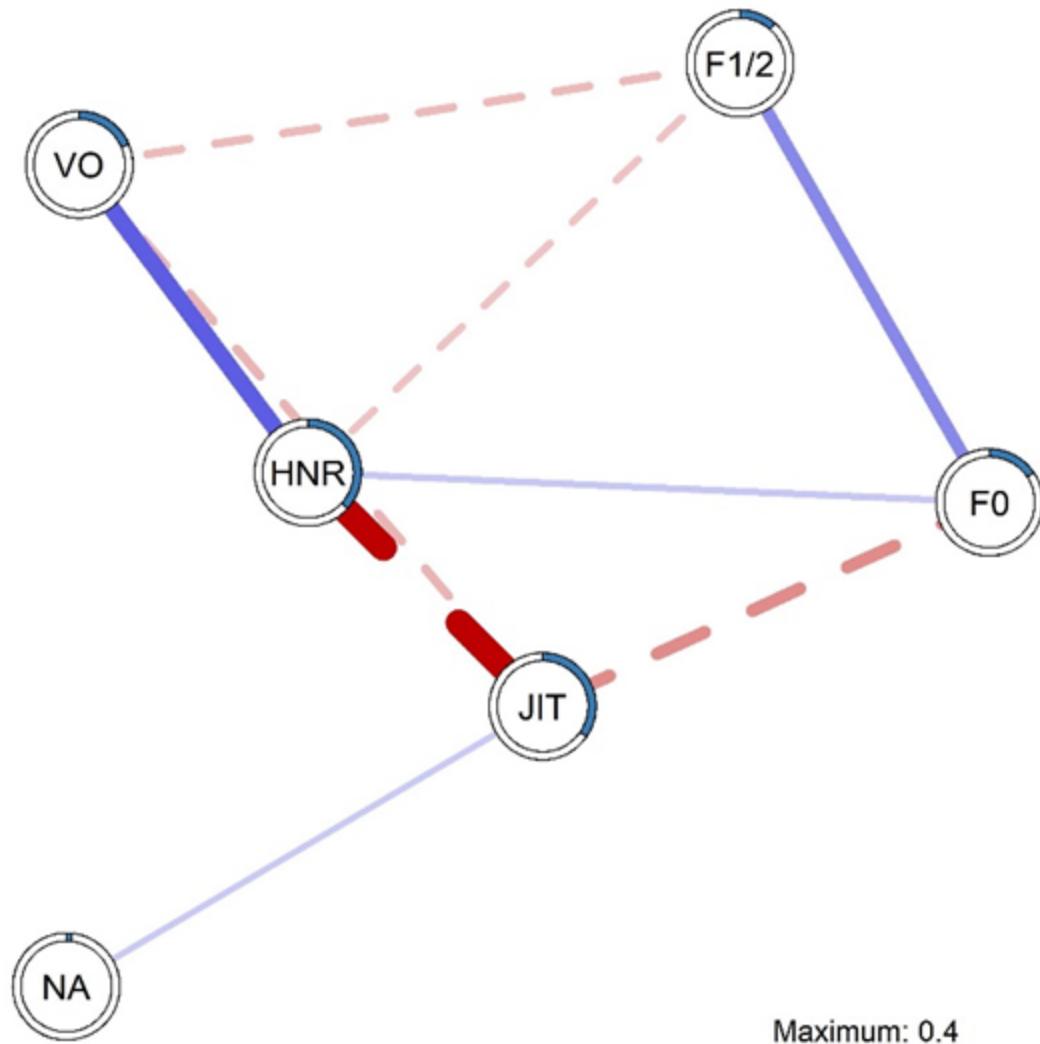
In line with the visual interpretation of the obtained network models, a statistical comparison of the models suggested strong overlap. That is, we observed a correlation of $r = .99$ between the adjacency matrices of both networks. Similarly, centrality Strength and node predictability for the resting state and post stressor networks each reached $r = .99$. Indeed, the Network Comparison Test suggested no significant differences in terms of overall network structure ($M = 0.06, p = .93$; network invariance test) or strength of connectivity (resting state network = 2.33; post stressor network = 2.44; $S = 0.11, p = .57$; global strength invariance test).

2.4.3. Modeling the Unique Associations Between Stress Reactivity and Change in Speech Parameters (aim 2)

Figure 5 (n=148) depicts the unique associations between change in negative affect (NA) and change in the speech parameters following the stress induction procedure with every edge (connection) between two nodes (variables) represents the sign (positive/negative) and the weight (strength) of the connection, depicting the unique associations between the two nodes while controlling for all other nodes in the network (Epskamp et al., 2018). Interestingly, change in JIT was the only speech parameter that was directly connected to change in NA. In particular, the experience of more negative affect throughout the stress induction procedure was directly related to increased JIT. All other speech parameters were only indirectly connected to change in NA through JIT. Change in JIT was negatively related to change in HNR and F0, and VO, which suggests that increases in JIT due to the stress induction procedure were related to decreases in HNR, F0 and VO. In addition, we observed positive associations between HNR and VO / F0, and F0 and F1/2. Finally we observed negative associations between F1/2 and VO / HNR. Based on node Strength, change in HNR and Jitter emerged as the most central nodes in the network, whereas change in NA was the least central node (see supplemental material for estimates of node predictability, edge accuracy, edge differences, and centrality stability).

Figure 5

Unique Associations Between Change in Negative Affect and Speech Parameters.



Note. Edges in the model represent the unique associations between each of the nodes. Edge thickness reflects the strength of association, where strong associations are presented using thicker edges. Blue / Full edges represent positive associations, whereas red / dashed edges represent negative associations; the edge weights presented in the model can also be found in the edge weight matrix (Supplemental table 8). Node predictability (R^2) is visualized as a pie chart around each node and can also be found in table 1.

2.5. Discussion

The two aims of the present study were to gain insight into (1) the unique associations between specific speech parameters (Fundamental frequency, Jitter, Harmonics-to-Noise Ratio, Voiced Segment Length, and Formant 1:2 Ratio) after as compared to before experiencing psychosocial stress, and (2) how change in these speech parameters was uniquely associated with change in self-reported NA following the stressor. We measured (change in) speech in the context of experimentally induced stress in a large sample of individuals selected from the community. The psychosocial stress induction was successful, as evidenced by higher skin conductance levels after the stress induction as compared to baseline, as well as increased negative affect following the stress induction. Moreover, we observed a significant positive association between these measures, indicating that the more skin conductance levels were increased following the psychosocial stress induction, the more negative mood was reported, providing support for the validity of the stress-induction method. As such, the comparison between the resting state and stress network allows us to test for changes in interrelations between the speech parameters after experiencing stress (aim 1).

First, network analyses were conducted on the selected speech parameters of interest at baseline, representing the unique associations between each of these parameters in a resting non-stressed state. This network shows Harmonics-to-Noise Ratio (HNR) as the most central node, being connected to all other speech parameters. The strongest connection that occurs is the positive connection between HNR and Fundamental frequency (F0), implying that less noise is present in higher pitched voices and vice versa, as has been reported by Ferrand

(2002). Furthermore, results show that several connections between most parameters are observed, which differ in their strength and orientation, indicating an interacting and cohesive network.

When comparing this baseline network with the post-stressor network, no differences between the interrelations of the different speech parameters were observed, suggesting that the relations between the parameters (selected as nodes in the current study) do not change after a stress induction procedure. More specifically, in both models, (1) HNR emerged as the most central node, (2) the strongest connection was observed between HNR and F0, and (3) all parameters were connected to at least 2 out of 4 other nodes in the network. The fact that the network of speech parameters was highly similar after versus prior to the stress induction procedure is an interesting and innovative observation, as it shows that the complex interplay between each of the above presented core speech parameters is not impacted by stress, and as such cannot be used as an indicator for stress. In particular, the unique interrelations remained stable in a stressed versus a non-stressed state.

In addition to comparing the pre- and post-stress networks of speech parameters, we composed an individual network model of change (delta) scores of each of the parameters and self-reported negative affect to gain insight into the unique relations between speech parameters and individual differences in stress reactivity (aim 2). We found that changes in Jitter (JIT), a fairly central speech parameter in the estimated network, were directly positively related to changes in self-reported negative affect, after controlling for the influence of other parameters in the network. Even though after the regularization procedure the strength of this association was relatively weak, this finding is important as it suggests a unique association

between speech and self-reported negative affect. Jitter quantifies the modulation of the periodicity of the voice signal and as such is related to the amplitude variation of the sound wave and is mainly affected by the lack of control of vibration of the cords (Teixeira et al., 2013). Increased jitter has been observed in pathological voices (Teixeira et al., 2013) and in physical task stress (Koblick, 2004), whereas decreased jitter is often discussed in the context of psychological stress (Giddens et al., 2013; Van Puyvelde et al., 2018). Yet, the literature is inconsistent, as both a decrease and an increase in Jitter have been observed with increased stress in different task designs (Giddens et al., 2013). However, Jitter has not been reported in relation to negative affect (Giddens et al., 2013). In early studies, it has been suggested that Jitter decreases in direct relation to stress levels (as described in Giddens et al., 2013; Van Puyvelde et al., 2018) and pointed out that Jitter might be a better indicator of stress than F0 (Hecker et al., 1968; Mendoza & Carballo, 1998). More recent studies have shown Jitter to be a crucial feature in the classification of stress and emotion (e.g. Li et al., 2007; Rothkrantz et al., 2004). However, Jitter has especially been highlighted in the field of speech pathology, being mainly affected by a lack of control over the vibration of the cords, which could explain its occurrence in psychosocial stress (Teixeira et al., 2013). As such, the unique association between the change in Jitter and the change in self-reported negative affect following a potent psychosocial stressor, while controlling for other effects and variables, opens a new avenue to the research field of speech parameters in the context of psychological stress.

Interestingly, even though the network model of the current study depicts a direct connection between change in self-reported negative affect and change in Jitter, Jitter is by itself strongly linked to several other speech parameters in the network model. A direct

connection with F0 was to be expected considering that Jitter represents the variations that occur in the fundamental frequency (F0). Moreover, especially a strong association between change in Jitter and change in HNR is observed, which together with Jitter form the most central nodes of the network. Prior studies have demonstrated that HNR is more sensitive to subtle differences in vocal function than is Jitter (Awan & Frenkel, 1994). Although direct connections between the other speech parameters and negative affect were expected, such as a positive unique association between negative affect and F0 (Giddens et al., 2013), our findings suggest that these parameters function through Jitter in their connections to changing mood in the context of psychosocial stress. It could be argued that, at least to some extent, these parameters function through HNR and its strong interplay with Jitter too.

To the best of our knowledge, this study is the first to examine the impact of psychosocial stress on the unique interrelations between key speech features, and how change in these parameters in the context of psychosocial stress relates to change in self-report measures for stress (i.e., negative affect). Our findings provide several implications for the measurement of speech in the context of psychosocial stress, as well as for the measurement of stress via speech features. That is, our findings point towards the stability of the network structure of speech features in the context of stress, and the role of Jitter as the only speech feature which showed a direct association with self-reported negative affect, suggesting the importance of Jitter in the context of stress assessment via speech. The present study used a standardized method of psychosocial stress induction in a highly controlled lab setting. The analysis has been conducted using an exploratory and data driven method which allows to

model complex interrelations in an intuitive manner. Therefore, the present study's main strength is the generation of trustworthy hypotheses.

Future studies using large sample sizes whilst maintaining a within-subject design in a controlled setting are absolutely warranted. On the other hand, considering the accessibility of high-quality microphones, combining frequent speech recordings with continuous smartwatch recordings of heart rate and skin conductance will generate more dynamic results that could withstand and overcome prior limitations of controlled lab settings and can uncover the stability and strength of the different relations. However, this is to be confirmed by basic experimental research investigating the complex relation between speech and stress in a well-controlled setting, which was the aim of the current study. Finding the key parameters of stressed speech and being able to use these to assess stress levels in a wide variety of settings, swiftly and cost-effectively, will enable us to monitor excessive stress levels and set up interventions where necessary.

Even though the current study has several strengths such as such as its innovative nature and ecological validity, some limitations should be discussed. Firstly, it should be noted that network models are merely descriptive rather than predictive. These networks are undirected and therefore do not allow any statements regarding the direction of the observed effects. This data driven, explorative analysis, is hypothesis generating as the identified paths in the networks might be indicative of causal relationships which should be tested in future prospective or experimental research. Secondly, as network models value each individual relation between the different parameters in an unguided manner, we were limited in the number of parameters that could be included in the model due to the sample size. The current

set of parameters was selected based on literature research and has brought forth a network of interesting relations. However, an expanded network would give more insight into the stability of these relations, as well as further explain the dynamics between speech parameters and negative affect. Thirdly, the required larger sample size to do so would also increase the strength of the network comparison (resting state vs. stressed) made. However, in this context both networks were highly similar. As such, the non-significant findings for the network comparison test are unlikely to be driven by a lack of power. Overall, Jitter seems to be a central node in the relation between speech and negative affect, which should therefore be further studied using confirmatory analyses. Fourthly, due to some technical setbacks, most of the collected data for ECG and EDA was not usable. This is especially unfortunate as this would give insight not only into the interplay between speech and self-reported negative affect but also into the relations with other indicators of objectively experienced stress (e.g., biomarkers).

2.6. Conclusions

Stress has long been a much-discussed topic, and as such many different methods for stress measurement have been proposed over the years. Recently, speech analysis has been proposed as a possible physiological marker for stress which can be measured in a remote and non-invasive matter. The current study deployed network analysis to investigate the unique associations between specific speech parameters prior to and following exposure to a psychosocial stressor (aim 1), and to model the unique associations between specific speech features and self-reported stress (i.e., experienced negative affect; aim 2). For this purpose, we

relied on a well-validated stress induction procedure in a controlled lab setting. The network of speech parameters was highly similar after versus before the stress induction, suggesting that the complex interplay between each of the used speech parameters was not impacted by stress. Interestingly, changes in Jitter were directly positively related to changes in self-reported negative affect, indicating that this speech feature may be of particular interest in the context of stress assessment. These findings warrant further investigation in the diagnostic value of speech features to monitor stress in daily life, which requires intensive time series data.

2.7. References

- Awan, S. N., & Frenkel, M. L. (1994). *Improvements in Estimating the Harmonics-to-Noise Ratio of the Voice*. 8(3), 255–262.
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, 9, 91–121.
<https://doi.org/10.1146/annurev-clinpsy-050212-185608>
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Möttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, 13–29.
<https://doi.org/10.1016/j.jrp.2014.07.003>
- Dedovic, K., Duchesne, A., Andrews, J., Engert, V., & Pruessner, J. C. (2009). The brain and the stress axis: The neural correlates of cortisol regulation in response to stress. *NeuroImage*, 47(3), 864–871. <https://doi.org/10.1016/j.neuroimage.2009.05.074>
- Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., & Pruessner, J. C. (2005). *The Montreal Imaging Stress Task: Using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain*. 30(5), 319–325.
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), 355–391. <https://doi.org/10.1037/0033-2909.130.3.355>

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195–212.

<https://doi.org/10.3758/s13428-017-0862-1>

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012).

qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4). <https://doi.org/10.18637/jss.v048.i04>

Epskamp, S., & Fried, E. I. (2015). bootnet: Bootstrap methods for various network estimation routines. *R-Package. Available at: Https://Rdrr.Io/Cran/Bootnet*.

Epskamp, S., & Fried, E. I. (2018). A Tutorial on Regularized Partial Correlation Networks. *Psychological Methods*, 23(4), 617–634. <https://doi.org/10.1037/met0000167>

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

<https://doi.org/10.1109/TAFFC.2015.2457417>

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE - The Munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 1459–1462. <https://doi.org/10.1145/1873951.1874246>

Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, 16(4), 480–487. [https://doi.org/10.1016/S0892-1997\(02\)00123-6](https://doi.org/10.1016/S0892-1997(02)00123-6)

Fink, G. (2017). Stress: Concepts, Definition and History☆. *Reference Module in Neuroscience and Biobehavioral Psychology, January*, 0–9.
<https://doi.org/10.1016/b978-0-12-809324-5.02208-2>

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ..., & Heiberger, R. (2012).

Package ‘car.’ *Vienna: R Foundation for Statistical Computing*.

Friedman, J., Hastie, T., & Tibshirani, R. (2014). glasso: Graphical lasso-estimation of Gaussian graphical models. *R Package Version, 1.*

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience, 21*(11), 1129–1164.

<https://doi.org/10.1002/spe.4380211102>

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.e21-390.e29.

<https://doi.org/10.1016/j.jvoice.2012.12.010>

Godin, K. W., Hasan, T., & Hansen, J. H. L. (2012). Glottal waveform analysis of physical task stress speech. *Thirteenth Annual Conference of the International Speech Communication Association, January*.

Haslbeck, J. M. B., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine, 47*(16), 2767–2776. <https://doi.org/10.1017/S0033291717001258>

Hecker, M. H. L., Stevens, K. N., von Bismarck, G., & Williams, C. E. (1968). Manifestations of Task-Induced Stress in the Acoustic Speech Signal. *The Journal of the Acoustical Society of America, 44*(4), 993–1001. <https://doi.org/10.1121/1.1911241>

Jones, P. J., Mair, P., & McNally, R. J. (2018). Visualizing Psychological Networks: A Tutorial in R. *Frontiers in Psychology, 9*. <https://doi.org/10.3389/fpsyg.2018.01742>

Kirchhuebel, C. (2010). The effects of Lombard speech on vowel formant measurements. *São Paulo School of Advanced Studies in Speech Dynamics SPSASSD 2010 Accepted Papers*, 38.

Kirschbaum, C., & Hellhammer, D. H. (1994). Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology*, 19(4), 313–333. [https://doi.org/10.1016/j.psyneuro.2004.00893.x](https://doi.org/10.1016/j.psyneuro.2004.00893)

Koblick, H. (2004). *Effects of Simultaneous Exercise and Speech Tasks on the Perception of Effort and Vocal Measures in Aerobic Instructors*. University of Central Florida, Orlando, Florida.

Kreiman, J., & Sidtis, D. (2011). Foundations of Voice Studies. In *Foundations of Voice Studies*. <https://doi.org/10.1002/9781444395068>

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*.

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*, IV-1081-IV-1084. <https://doi.org/10.1109/ICASSP.2007.367261>

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12(3), 263–273. [https://doi.org/10.1016/S0892-1997\(98\)80017-9](https://doi.org/10.1016/S0892-1997(98)80017-9)

Monroe, S. M. (2008). Modern Approaches to Conceptualizing and Measuring Human Life Stress. *Annual Review of Clinical Psychology*, 4(1), 33–52. <https://doi.org/10.1146/annurev.clinpsy.4.022007.141207>

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

- Orlikoff, R. F. (1990). Vowel amplitude variation associated with the heart cycle. *Journal of the Acoustical Society of America*, 88(5), 2091–2098. <https://doi.org/10.1121/1.400106>
- Orlikoff, R. F., & Baken, R. J. (1989). The Effect of the Heartbeat on Vocal Fundamental Frequency Perturbation. *Journal of Speech, Language, and Hearing Research*, 32(3), 576–582. <https://doi.org/10.1044/jshr.3203.576>
- Panksepp, J. (2003). Feeling the pain of social loss. *Science*, 302(5643), 237–239. <https://doi.org/10.1126/science.1091062>
- Rossi, V., & Pourtois, G. (2012). Transient state-dependent fluctuations in anxiety measured using STAI, POMS, PANAS or VAS: A comparative review. *Anxiety, Stress and Coping*, 25(6), 603–645. <https://doi.org/10.1080/10615806.2011.582948>
- Rothkrantz, L. J. M., Wiggers, P., Van Wees, J. W. A., & Van Vark, R. J. (2004). Voice stress analysis. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3206, 449–456. <https://doi.org/10.4135/9781452229300.n1969>
- Shahin, I., & Botros, N. (2001). Modeling and analyzing the vocal tract under normal and stressful talking conditions. *Proceedings. IEEE SoutheastCon 2001 (Cat. No.01CH37208)*, 213–220. <https://doi.org/10.1109/SECON.2001.923118>
- Shields, G. S., & Slavich, G. M. (2017). Lifetime stress exposure and health: A review of contemporary assessment methods and biological mechanisms. *Social and Personality Psychology Compass*, 11(8), 1–17. <https://doi.org/10.1111/spc3.12335>
- Sigmund, M. (2012). Influence of Psychological Stress on Formant Structure of Vowels. *Elektronika Ir Elektrotehnika*, 18(10), 45–48. <https://doi.org/10.5755/j01.eee.18.10.3059>

Slavich, G. M., Taylor, S., & Picard, (2019). *Stress measurement using speech: Recent advancements , validation issues , and ethical and privacy considerations*. 3890.

<https://doi.org/10.1080/10253890.2019.1584180>

Sondhi, S., Khan, M., Vijay, R., & K. Salhan, A. (2015). Vocal Indicators of Emotional Stress. *International Journal of Computer Applications*, 122(15), 38–43.

<https://doi.org/10.5120/21780-5056>

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, 9, 1112–1122.

<https://doi.org/10.1016/j.protcy.2013.12.124>

Titze, I. R., & Martin, D. W. (1998). *Principles of voice production*.

Tossani, E. (2013). The concept of mental pain. *Psychotherapy and Psychosomatics*, 82(2), 67–73. <https://doi.org/10.1159/000343003>

van Borkulo, C. D., Boschloo, L., Kossakowski, J., Tio, P., Schoevers, R., Borsboom, D., & Waldorp, L. (2017). *Comparing network structures on three aspects: A permutation test*.

<https://doi.org/10.13140/RG.2.2.29455.38569>

van Borkulo, C. D., Epskamp, S., & Millner, A. (2016). *Network Comparison Test: Statistical comparison of two networks based on three invariance measures. R Package*.

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Frontiers in Psychology*, 9.

<https://doi.org/10.3389/fpsyg.2018.01994>

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). *The huge Package for High-dimensional Undirected Graph Estimation in R*. 13(1), 1059–1062.

2.8. Supplemental Materials

2.8.1. Acknowledgments

We thank Vic Degraeve, Sean Deloddere, Ernest van Hoecke, and Marie Vanzieleghem for their help collecting the data and programming the experiment.

2.8.2. Funding

This research was supported by a grant for research at Ghent University (BOFSTA2017002501) and a grant from the King Baudouin Foundation (KBS 2018-J1130650–209563). KH is a Postdoctoral Fellow of the FWO (FWO.3EO.2018.0031.01).

2.8.3. Open practices statement

All codes, processed data, examples of raw data files, plots, and accompanying information are made openly available through <https://osf.io/7byzh/>

2.8.4. Supplemental materials

Supplemental Materials are also openly made available and can be found through <https://osf.io/7byzh/>

Package version info:

R version 4.0.2 (2020-06-22)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 10 x64 (build 18363)

Matrix products: default

locale:

```
[1] LC_COLLATE=English_Belgium.1252 LC_CTYPE=English_Belgium.1252 LC_MONETARY=English_Belgium.1252 LC_NUMERIC=C  
LC_TIME=English_Belgium.1252
```

attached base packages:

```
[1] stats graphics grDevices utils datasets methods base
```

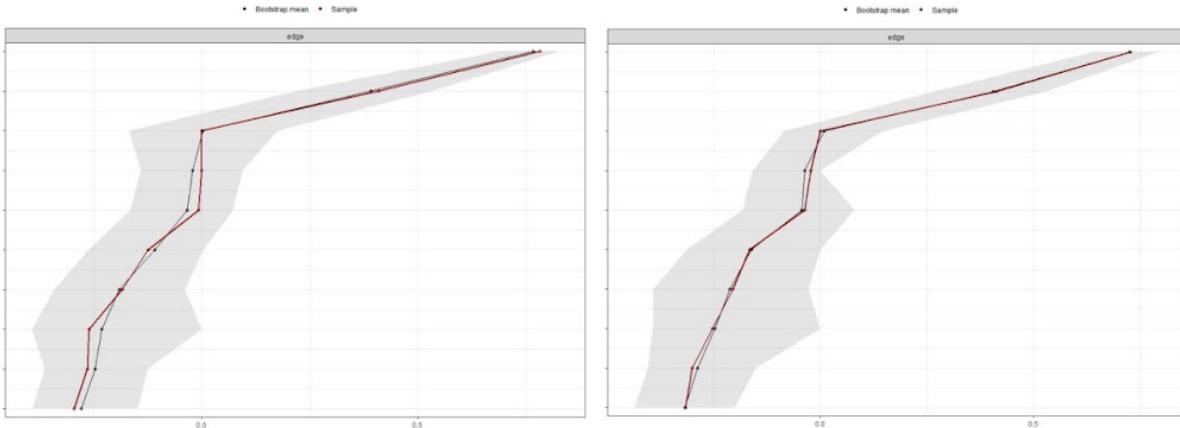
other attached packages:

```
[1] pastecs_1.3.21      qwraps2_0.5.2       reshape2_1.4.4      ggeffects_1.0.2     forcats_0.5.0       stringr_1.4.0       purrr_0.3.4  
[8] readr_1.4.0        tibble_3.0.3       tidyverse_1.3.0      gridExtra_2.3       ggsignif_0.6.0      MuMin_1.43.17      effects_4.2-0  
[15] fitdistrplus_1.1-1 survival_3.1-12    MASS_7.3-51.6       NetworkComparisonTest_2.2.1 emmeans_1.5.2-1      pander_0.6.3  
lmerTest_3.1-3  
[22] lme4_1.1-25       reshape_0.8.8      tidyR_1.1.2        skimr_2.1.2        huge_1.3.4.1       mgm_1.2-10        bootnet_1.4.3  
[29] qgraph_1.6.5       yarr_0.1.5        circlize_0.4.10    BayesFactor_0.9.12-4.2 Matrix_1.2-18      coda_0.19-3       jpeg_0.1-8.1  
[36] cowplot_1.1.0     ggpubr_0.4.0      ggplot2_3.3.2     dplyr_1.0.2        car_3.0-10        carData_3.0-4
```

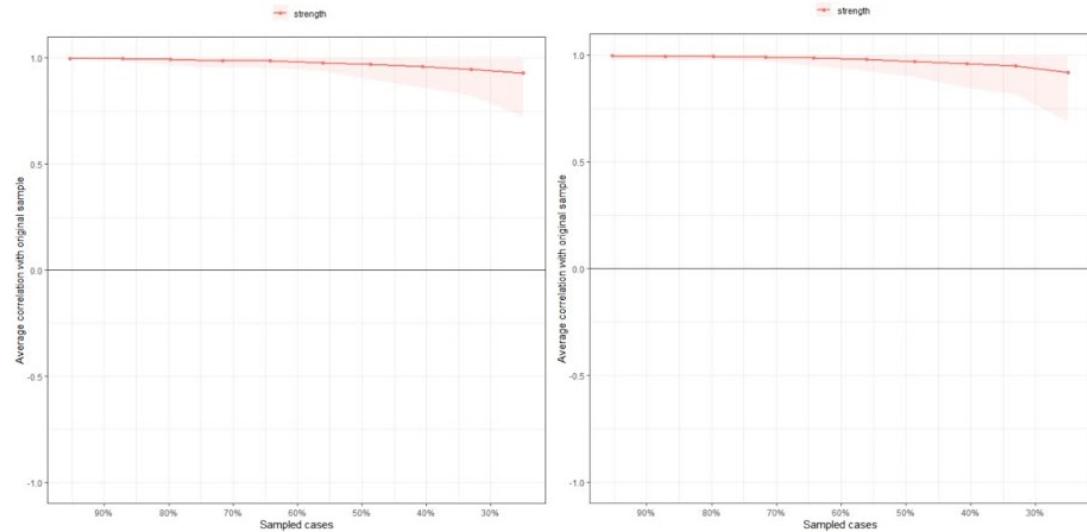
loaded via a namespace (and not attached):

```
[1] utf8_1.1.4        R.utils_2.10.1     tidyselect_1.1.0   htmlwidgets_1.5.2   grid_4.0.2       munsell_0.5.0      codetools_0.2-16   statmod_1.4.35  
withr_2.2.0         colorspace_1.4-1  
[11] NetworkToolbox_1.4.0 knitr_1.30      rstudioapi_0.11    stats4_4.0.2      labeling_0.3       repr_1.1.0        mnormt_2.0.2      farver_2.0.3  
vctrs_0.3.4         generics_0.0.2  
[21] xfun_0.18        R6_2.4.1        doParallel_1.0.16 smacof_2.1-1      assertthat_0.2.1   scales_1.1.1       nnet_7.3-14       gtable_0.3.0  
weights_1.0.1       rlang_0.4.7  
[31] MatrixModels_0.4-1 GlobalOptions_0.1.2 splines_4.0.2     rstatix_0.6.0     wordcloud_2.6     broom_0.7.2       checkmate_2.0.0   modelr_0.1.8  
abind_1.4-5         d3Network_0.5.2.1  
[41] backports_1.1.10 Hmisc_4.4-1      tools_4.0.2       psych_2.0.9      lavaan_0.6-7      ellipsis_0.3.1     RColorBrewer_1.1-2 polynom_1.4-0  
Rcpp_1.0.5          plyr_1.8.6  
[51] base64enc_0.1-3   rpart_4.1-15     pbapply_1.4-3     haven_2.3.1     cluster_2.1.0     fs_1.5.0         survey_4.0       magrittr_1.5  
data.table_1.13.2   openxlsx_4.2.2  
[61] reprex_0.3.0     tmvnsim_1.0-2    mvtnorm_1.1-1    matrixcalc_1.0-3  whisker_0.4      hms_0.5.3        xtable_1.8-4     pbkrtest_0.4-8.6  
rio_0.5.16          readxl_1.3.1  
[71] shape_1.4.5      compiler_4.0.2    ellipse_0.4.2    mice_3.11.0     crayon_1.3.4     minqa_1.2.4      R.oo_1.24.0     htmltools_0.5.0  
corpcor_1.6.9       Formula_1.2-4  
[81] lubridate_1.7.9   DBI_1.1.0      relaimpo_2.2-3   sjlabelled_1.1.7 dbplyr_1.4.4     boot_1.3-25     IsingSampler_0.2.1 IsingFit_0.3.1  
cli_2.0.2          heplots_1.3-5  
[91] mitools_2.4      R.methodsS3_1.8.1  gdata_2.18.0    parallel_4.0.2   insight_0.13.2   igraph_1.2.6     BDgraph_2.63    pkgconfig_2.0.3  
numDeriv_2016.8-1.1 foreign_0.8-80  
[101] xml2_1.3.2      foreach_1.5.1    pbivnorm_0.6.0   estimability_1.3  rvest_0.3.6     digest_0.6.25    cellranger_1.1.0 htmlTable_2.1.0  
curl_4.3           gtools_3.8.2  
[111] rjson_0.2.20    nloptr_1.2.2.2   lifecycle_0.2.0   nlme_3.1-148   glasso_1.11     jsonlite_1.7.1   fansi_0.4.1     pillar_1.4.6  
lattice_0.20-41    httr_1.4.2  
[121] plotrix_3.7-8   glue_1.4.2      networktools_1.2.3 zip_2.1.1      fdrtool_1.2.15   png_0.1-7       iterators_1.0.13 candisc_0.8-3  
glmnet_4.0-2       class_7.3-17  
[131] stringi_1.5.3   nnls_1.4       blob_1.2.1       latticeExtra_0.6-29 eigenmodel_1.11 e1071_1.7-4
```

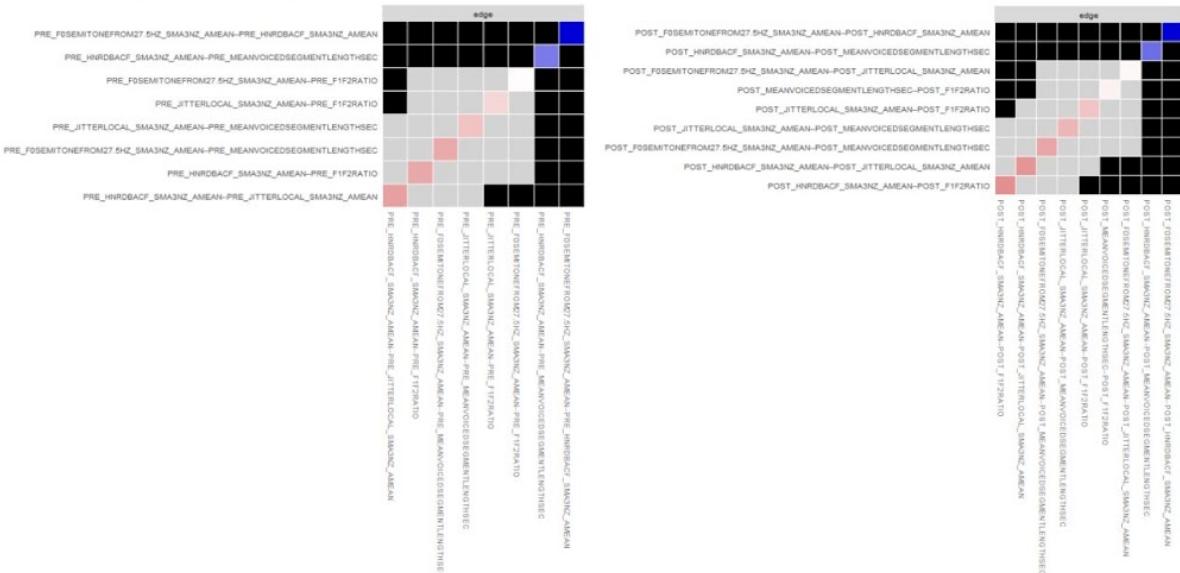
Supplemental Figures



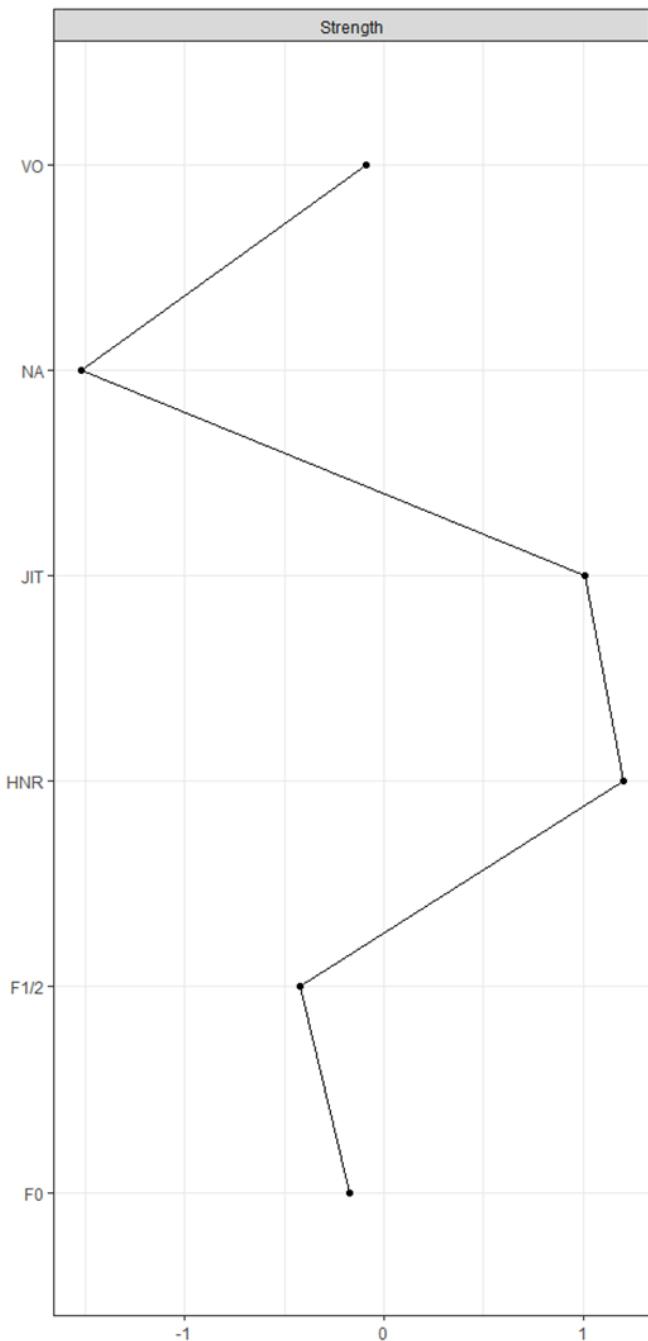
Supplemental Figure 1. Edge accuracy for the Resting state (left) and Stress (right) network models.



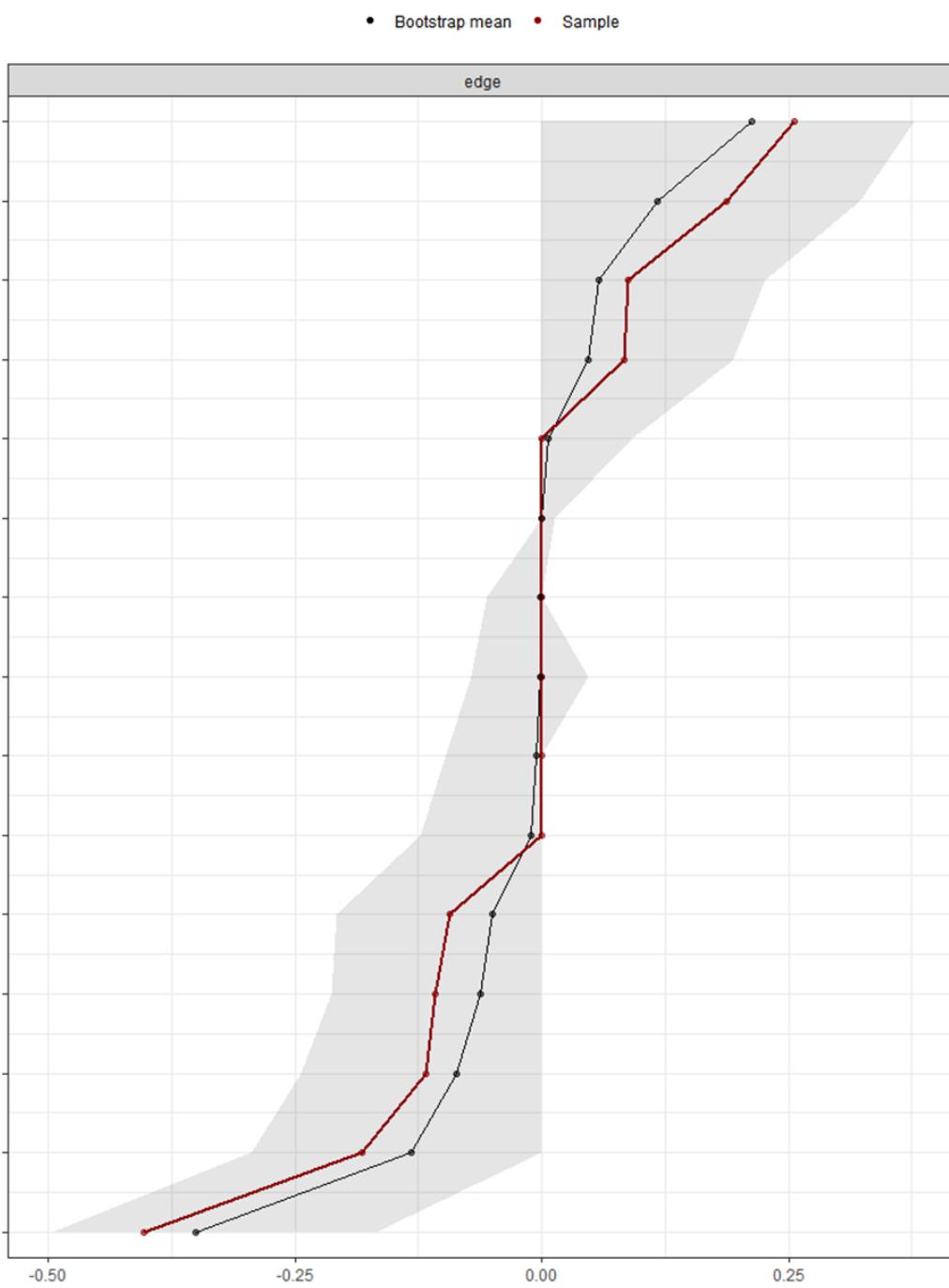
Supplemental Figure 2. Stability of Strength centrality for the Resting state (left) and Stress (right) network models. Note that for both networks, good correlation stability was obtained (.75).



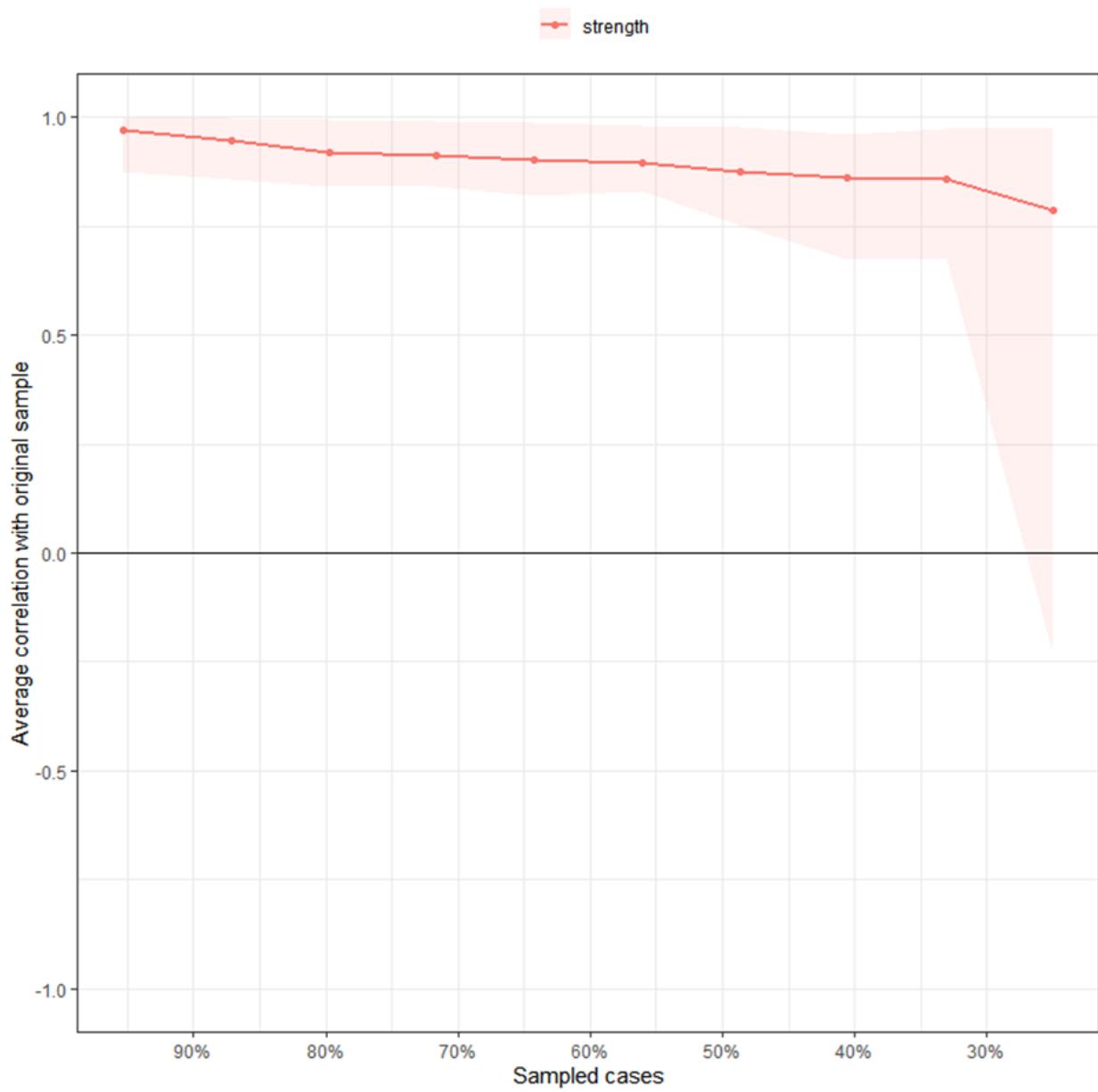
Supplemental Figure 3. Significant edge differences for the Resting state (left) and Stress (right) network models.



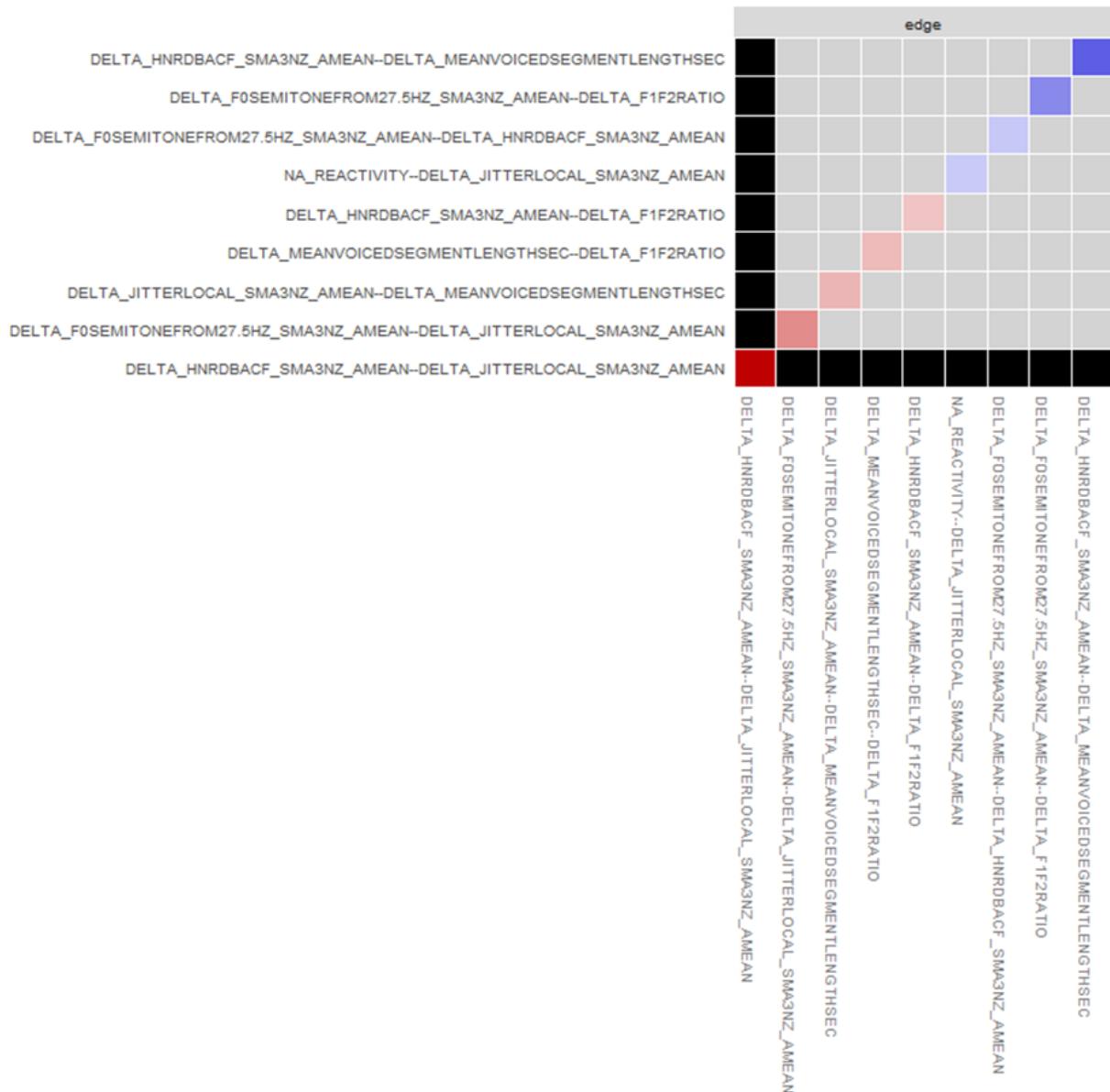
Supplemental Figure 4. Strength centrality for the Stress Reactivity network.



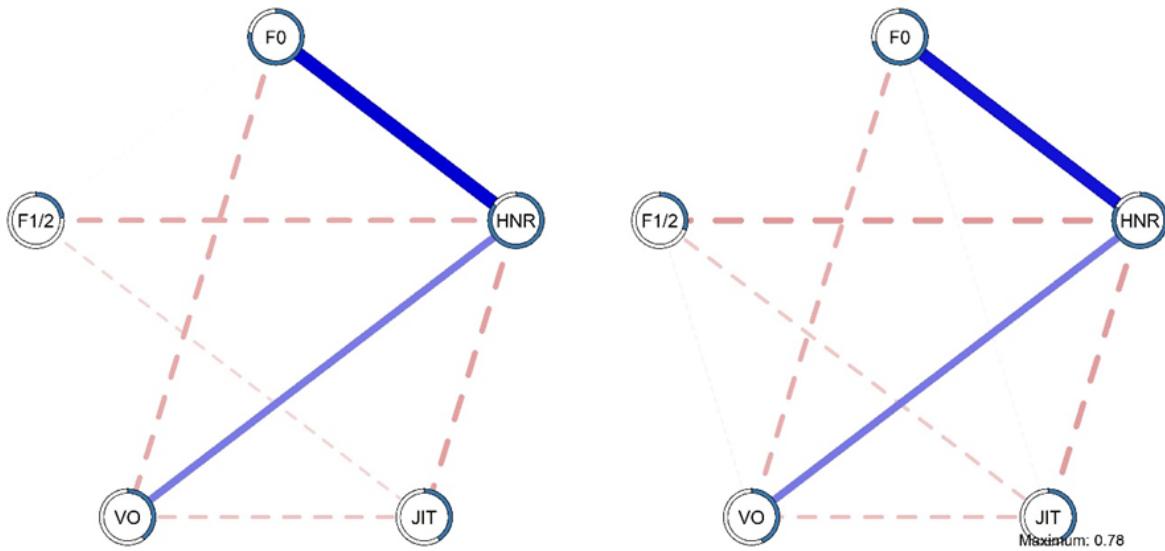
Supplemental Figure 5. Edge accuracy for the Stress Reactivity Network.



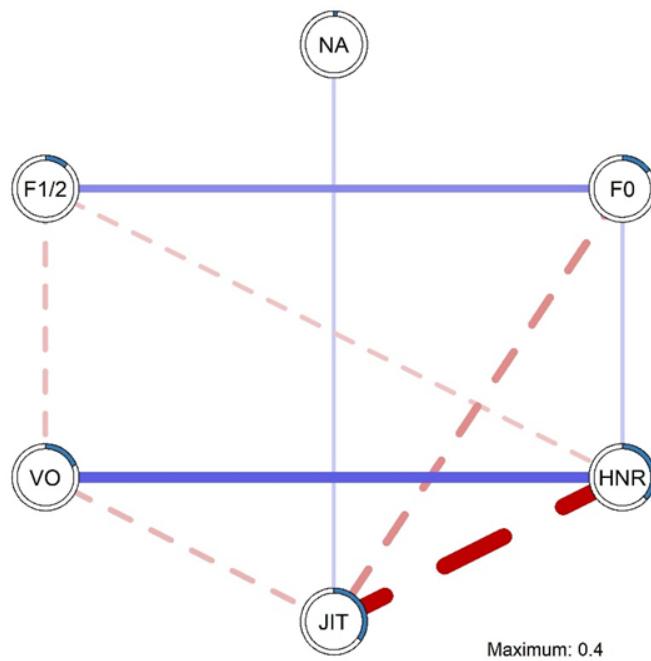
Supplemental Figure 6. Stability of Strength centrality for the Stress Reactivity Network. Note that the correlation stability was adequate (.28).



Supplemental Figure 7. Significant edge differences for the Stress Reactivity Network



Supplemental Figure 8. Circular layout plots for the Resting state (left) and Stress (right) network models.



Supplemental Figure 9. Circular layout plots for the Stress Reactivity Network

Supplemental Tables

Supplemental Table 1. Node predictability for Resting state and Stress network

Node	R ² Resting state network	R ² Stress network
F0	.78	.72
HNR	.85	.83
JIT	.41	.42
VO	.40	.44
F1/2	.24	.31

Supplemental Table 2. Node predictability for Stress Reactivity Network

Node	R ²
NA	.02
F0	.16
HNR	.37
JIT	.35
VO	.19
F1/2	.12

Supplemental Table 3. Descriptive Statistics for Speech Parameters at Resting state

	F0	HNR	JIT	VO	F1/2	RATE
Median	26.33	2.59	.0526	.1721	.3582	2.897
Mean	27.82	3.14	.0558	.1761	.3633	2.921
SE. Mean	.37	.19	.0013	.0026	.0022	.030
Cl. Mean 0.95	.73	.37	.0025	.0051	.0043	.059
Std. Dev.	4.47	2.29	.0452	.0316	.0265	.361

Supplemental Table 4. Descriptive Statistics for Speech Parameters at Post-Stressor state

	F0	HNR	JIT	VO	F1/2	RATE
Median	26.68	2.73	.0521	.178	.361	2.712
Mean	27.88	3.23	.0546	.180	.364	2.755
SE. Mean	.37	.19	.0012	.0015	.002	.027
Cl. Mean 0.95	.74	.37	.0024	.0049	.005	.053
Std. Dev.	4.54	2.28	.0148	.0304	.029	.329

Supplemental Table 5. Descriptive Statistics for Delta scores of Parameters

	NA	F0	HNR	JIT	V0	F1/2	RATE
Median	3.83	-.010	.027	-.0004	.0057	.0005	-.145
Mean	6.45	.055	.090	-.0012	.0040	.0004	-.166
SE. Mean	.88	.068	.049	.0009	.0016	.0009	.015
Cl. Mean 0.95	1.73	.134	.096	.0017	.0031	.0018	.049
Std. Dev.	10.67	.827	.590	.0104	.0188	.0108	3.303

Supplemental Table 6. Edge Weight Matrix Pre-Stressor network

	F0	HNR	JIT	V0	F1/2
F0	0.00	0.78	0.00	-0.26	-0.01
HNR	0.78	0.00	-0.30	0.41	-0.27
JIT	0.00	-0.30	0.00	-0.18	-0.12
V0	-0.26	0.41	-0.18	0.00	0.00
F1/2	-0.01	-0.27	-0.12	0.00	0.00

Supplemental Table 7. Edge Weight Matrix Post-Stressor network

	F0	HNR	JIT	V0	F1/2
F0	0.00	0.73	-0.02	-0.25	0.00
HNR	0.73	0.00	-0.30	0.41	-0.32
JIT	-0.02	-0.30	0.00	-0.21	-0.17
V0	-0.25	0.41	-0.21	0.00	-0.04
F1/2	0.00	-0.32	-0.17	-0.04	0.00

Supplemental Table 8. Edge Weight Matrix Stress Reactivity network

	NA	F0	HNR	JIT	V0	F1/2
NA	0.00	0.00	0.00	0.08	0.00	0.00
F0	0.00	0.00	0.09	-0.18	0.00	0.19
HNR	0.00	0.09	0.00	-0.40	0.26	-0.09
JIT	0.08	-0.18	-0.40	0.00	-0.12	0.00
V0	0.00	0.00	0.26	-0.12	0.00	-0.11
F1/2	0.00	0.19	-0.09	0.00	-0.11	0.00

Feature calculation:

We used OpenSmile 2.3.0 (Eyben et al., 2010) with the GeMAPS configuration (Eyben et al., 2015) to extract the used speech features. Here we will describe the calculation of every feature with reference to the GeMAPS configuration paper (Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202. <https://doi.org/10.1109/TAFFC.2015.2457417>).

F0

F0 variable name in GeMAPS configuration output:
F0SEMITONEFROM27.5HZ_SMA3NZ_AMEAN

In the GeMAPS paper the calculation is described on page 198 as:

“The fundamental frequency (F0) is computed via subharmonic summation (SHS) in the spectral domain as described by [60]. Spectral smoothing, spectral peak enhancement, and auditory weighting are applied as in [60]. 15 harmonics are considered, i.e., the spectrum is octave shift-added 15 times, and a compression factor of 0.85 is used at each shifting ([60]). F0 $\frac{1}{4} 0$ is defined for unvoiced regions. The voicing probability is determined by the ratio of the harmonic summation spectrum peak belonging to an F0 candidate and the average amplitude of all harmonic summation spectrum bins, scaled to a range $\frac{1}{2}0$; 1_. A maximum of 6 F0 candidates in the range of 55–1000 Hz are selected. Online Viterbi post-smoothing is applied to select the most likely F0 path through all possible candidates. A voicing probability threshold of 0:7 is then applied to discern voiced from unvoiced frames. After Viterbi smoothing the F0 range of 55–1000 Hz is enforced by setting all voiced frames outside the range to unvoiced frames (F0 $\frac{1}{4} 0$). The final F0 value is converted from its linear Hz scale to a logarithmic scale – a semitone frequency scale starting at 27.5 Hz (semitone 0). However, as 0 is reserved for unvoiced frames, every value below semitone 1 (29.136 Hz) is clipped to 1.

HNR

HNR variable name in GeMAPS configuration output: HNRDBACF_SMA3NZ_AMEAN

In the GeMAPS paper the calculation is described on page 199 as:

“The HNR gives the energy ratio of the harmonic signal parts to the noise signal parts in dB. It is estimated from the short-time autocorrelation function (ACF) (60 ms window) as the logarithmic ratio of the ACF amplitude at F0 and the total frame energy, expressed in dB, as given by [61]:

$$HNR_{acf,log} = 10 \log_{10} \left(\frac{ACF_{T_0}}{ACF_0 - ACF_{T_0}} \right) dB.$$

where ACFT0 is the amplitude of the autocorrelation peak at the fundamental period (derived from the SHS-based F0 extraction algorithm described above) and ACF0 is the zeroth ACF coefficient (equivalent to the quadratic frame energy). The logarithmic HNR value is floored to -100 dB to avoid highly negative and varying values for low-energy noise.

Jitter

Jitter variable name in GeMAPS configuration output: JITTERLOCAL_SMA3NZ_AMEAN

In the GeMAPS paper the calculation is described on page 198 as:

“Jitter, is computed as the average (over one 60 ms frame) of the absolute local (period to period) jitter $J_{pp,n}$ scaled by the average fundamental period length. For two consecutive pitch periods, with the length of the first period $n_0 - 1$ being $T_0 n_0 - 1 P$ and the length of the second period n_0 being $T_0 n_0 P$, the absolute period to period jitter, also referred to as absolute local jitter, is given as follows [61]:

$$J_{pp}(n) = |T_0(n) - T_0(n - 1)| \text{ for } n > 1$$

This definition yields one value for J_{pp} for every pitch period, starting with the second one. To obtain a single jitter value per frame for N_0 local pitch periods $n_0 \in 1 \dots N_0$ within one analysis frame, the average local jitter J_{pp} is given by:

$$J_{pp} = \frac{1}{N-1} \sum_{n=2}^{N} |T_0(n) - T_0(n - 1)|$$

In order to make the jitter value independent of the underlying pitch period length, it is scaled by the average pitch period length. This yields the average relative jitter, used as the jitter measure in our parameter set:

$$J_{pp,rel} = \frac{\frac{1}{N-1} \sum_{n=2}^N |T_0(n) - T_0(n-1)|}{\frac{1}{N} \sum_{n=1}^N T_0(n)}.$$

Voiced (mean voiced segment length)

Mean voiced segment length variable name in GeMAPS configuration output:
MEANVOICEDSEGMENTLENGTHSEC

In the GeMAPS paper the calculation is described on page 193 as:

“the mean length and the standard deviation of continuously voiced regions ($F0 > 0$),”

F1/2

This feature is feature is computed by calculating the ratio between Formant 1 and Formant 2.

These variables are found in the GeMAPS configuration output under the names:

F1FREQUENCY_SMA3NZ_AMEAN & F2FREQUENCY_SMA3NZ_AMEAN

In the GeMAPS paper the calculation is described on page 199 as:

“Both formant bandwidth and formant centre frequency are computed from the roots of Linear Predictor (LP) [67] coefficient polynomial. The algorithm follows the implementation of [11].”

Relevant references from GeMAPS configuration paper with according numbers:

- [11] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot Int.*, vol. 5, nos. 9/10, pp. 341–345, 2001.

- [60] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, 1988.
- [61] B. Schuller, *Intelligent Audio Analysis (Signals and Communication Technology)*. New York, Ny, USA: Springer, 2013.
- [67] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 5, pp. 561–580, Apr. 1975.

Other supplemental materials

Read-out-loud Text:

"Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij, dus dat duurde wel even. Nu wachten ze tot de trein eraan komt. Het is al vijf over drie, dus het duurt nog vier minuten. Er staan nog veel meer mensen te wachten. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen." From: van de Weijer and Slis (1991)

Van de Weijer, J., Slis, I. (1991). Nasaliteitsmeting met de Nasometer. *Logopedie en Foniatrie*, 63, 97-101.

Translation: "Papa and Marloes are at the station. They are waiting for the train. First, they bought a ticket. There was a very long queue, so it took a while. Now they wait for the train to arrive. It's already five past three, so it's still four minutes. Many more people are waiting. Marloes looks to the left, she sees the train coming in the distance."

This text is 70 words and has been developed as a phonetically balanced text that matches the sound frequencies as they occur in the Dutch language as validated by van den Broecke and described in van de Weijer (1990) and van de Weijer & Slis (1991). During both moments (pre-

and post-stressor) they articulated this text only once. We used the OpenSmile toolbox that computes mean values for each fragment per variable (for more information see: Eyben et al., 2010 and Eyben et al., 2015) resulting in the use of one datapoint per variable per recording.

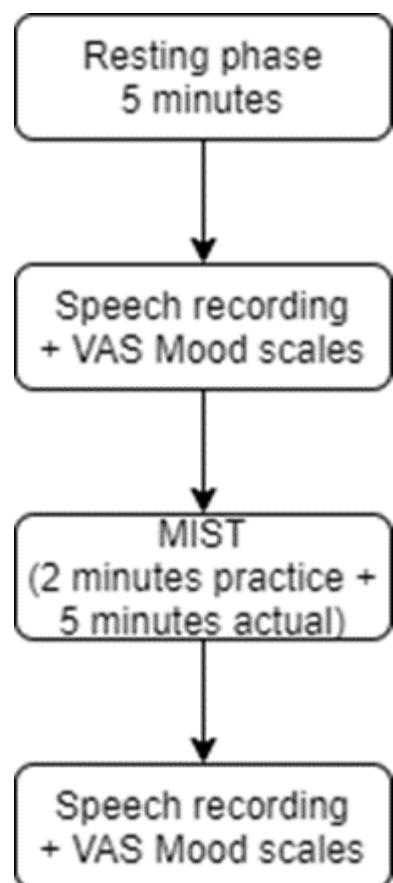
The voiced to voiceless ratio is as follows:

Pre-stress: Mean = 2.103, SD = .533

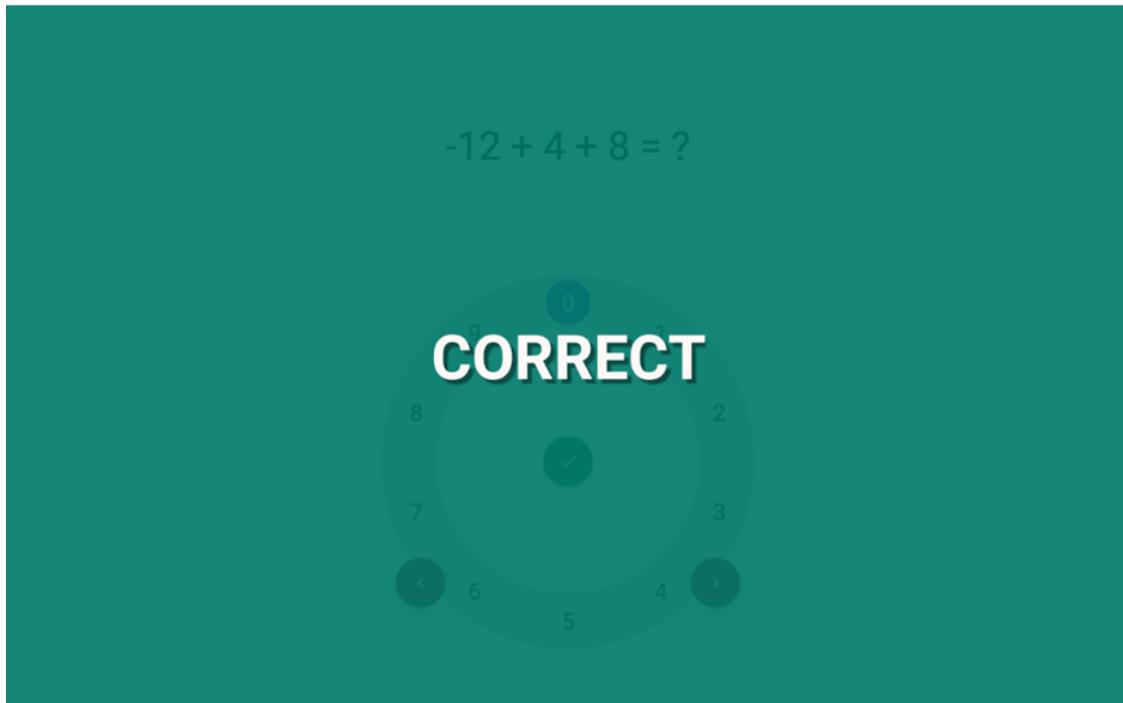
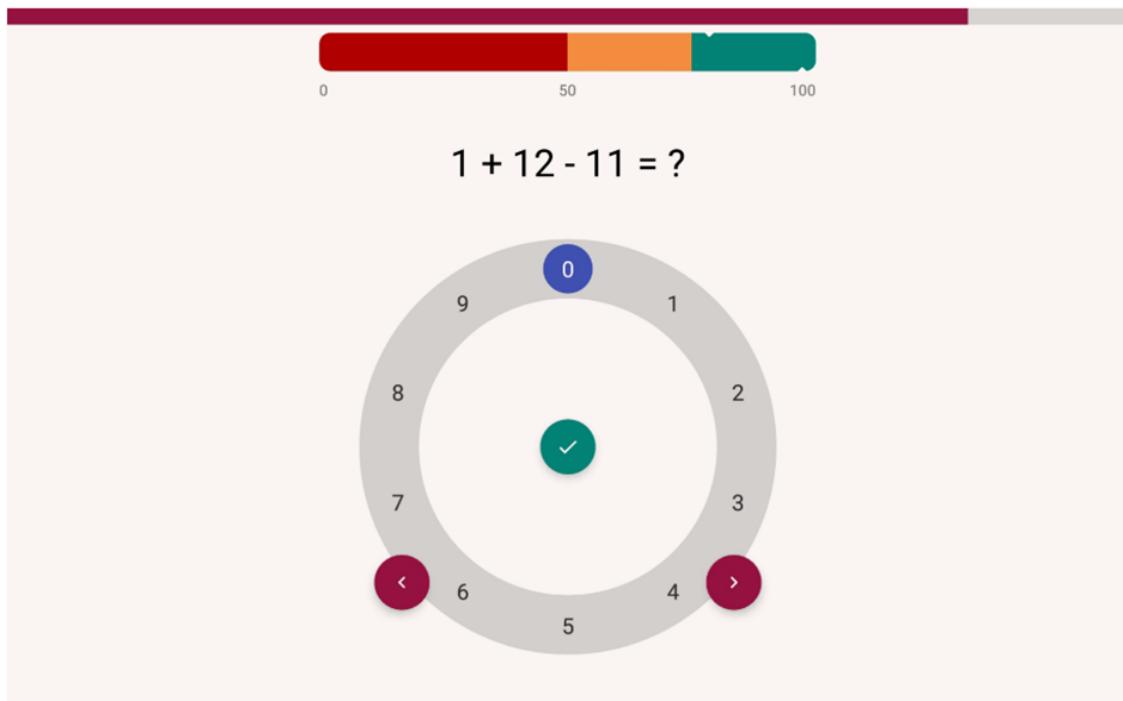
Post-stress: Mean = 2.071, SD = .549

Distributions were similar between the two moments. Data with regards to each individual participants number of voiced and unvoiced frames has been added to OSF.

Study Flowchart

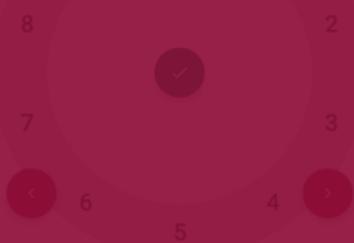


Screenshots



$$-9 + 3 + 9 = ?$$

INCORRECT



$$-22 + 10 + 17 = ?$$

TIMEOUT



Acoustic speech features in social comparison: how stress impacts the way you sound

Mitchel Kappen^{*123}, Jonas van der Donckt^{*4}, Gert Vanhollebeke¹²⁵, Jens Allaert¹²³, Vic Degraeve⁴, Nilesh Madhu⁴, Sofie Van Hoecke⁴, Marie-Anne Vanderhasselt¹²

**Contributed equally*

¹Department of Head and Skin, Ghent University, University Hospital Ghent (UZ Ghent), Department of Psychiatry and Medical Psychology, Ghent, Belgium

²Ghent Experimental Psychiatry (GHEP) Lab, Ghent University, Ghent, Belgium

³Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

⁴IDLab, Ghent University - imec, Ghent, Belgium

⁵Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Published as:

Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van Hoecke, S., & Vanderhasselt, M. A. (2022). Acoustic speech features in social comparison: how stress impacts the way you sound. *Scientific Reports*, 12(1), 22022.

3.1. Abstract

The use of speech as a digital biomarker to detect stress levels is increasingly gaining attention. Yet, heterogeneous effects of stress on specific acoustic speech features have been observed, possibly due to previous studies' use of different stress labels/categories and the lack of solid stress induction paradigms or validation of experienced stress. Here, we deployed a controlled, within-subject psychosocial stress induction experiment in which participants received both neutral (control condition) and negative (negative condition) comparative feedback after solving a challenging cognitive task. This study is the first to use a (non-actor) within-participant design that verifies a successful stress induction using both self-report (i.e., decreased reported valence) and physiological measures (i.e., increased heart rate acceleration using event-related cardiac responses during feedback exposure). Analyses of acoustic speech features showed a significant increase in Fundamental Frequency (F0) and Harmonics-to-Noise Ratio (HNR), and a significant decrease in shimmer during the negative feedback condition. Our results using read-out-loud speech comply with earlier research, yet we are the first to validate these results in a well-controlled but ecologically-valid setting to guarantee the generalization of our findings to real-life settings. Further research should aim to replicate these results in a free speech setting to test the robustness of our findings for real-world settings and should include semantics to also take into account what you say and not only how you say it.

3.2. Introduction

Stress is omnipresent in modern society. Whereas low levels of stress can increase one's performance, the chronic experience of stress is a common risk factor for a variety of different mental and physical health problems (Miller et al., 2002; Slavich, 2016), making it a critical factor in determining human health (Epel et al., 2018). Therefore, frequent, accurate, and affordable stress measurement tools would be of great contribution to society as regular observation of increased acute stress levels would be indicative of a chronically stressed state.

Acute stress, in scientific studies, is often measured by using either self-report assessments or monitoring physiological signals such as electrocardiography or electrodermal activity (Giannakakis et al., 2022). However, the use of speech as a novel biomarker for (acute) stress has rapidly gained attention due to it being non-intrusive (no physical connection necessary to the body), affordable to acquire, and ubiquitous, considering the increasing presence of high-quality microphones in everyday objects (Giddens et al., 2013; Monroe, 2008). As stress influences important factors in speech production such as breathing, cardiac activity, or general pose, it is hypothesized that experienced stress could be detected from acoustic features of one's voice (for an extensive explanation of each step of speech production with regards to these features (See: Van Puyvelde et al., 2018). Moreover, speech has increasingly shown to be a potential sensitive marker for depression, schizophrenia, and autism (Cho et al., 2022; Koops et al., 2021; Voppel et al., 2022). Since stress is considered a key underlying working mechanism of negative mood and a risk factor for the development of mood disorders and the expression of a wide range of psychological diseases, its effects on

speech could further progress the (early) detection of numerous psychological diseases. In addition, stress could affect what words you utter, their complexity, and other prosodic features due to changes in cognitive load (Paulmann et al., 2016; Sandi, 2013). However, to isolate and evaluate the effects of stress on acoustic speech features, it is necessary to exclude both interindividual differences in linguistic capabilities and acoustic effects induced by variations in words and sentences by using read-out-loud speech fragments before moving on to include linguistic features such as syntax and semantics.

As the field of measuring stress in speech is evolving quickly, it is proposed that the vocal response to stress may be as individual and unique as the voice itself, and thus more isolated studies that control for interindividual differences are required (Giddens et al., 2013; Van Puyvelde et al., 2018). Whereas many recent studies use large samples of audio fragments and extract a wide scale of features using easily accessible toolboxes such as PRAAT (Boersma, 2001) and OpenSMILE (Eyben et al., 2010), some limitations can be noted. It is argued that these studies 1) are often between-subject, therefore unable to contain interindividual differences in the stress response, 2) lack a valid verification of the subject's emotional state, or 3) include a wide range of (acoustic) features that lack scientific basis, which increases the risk of overfitted models that would not generalize well to everyday life situations (Giddens et al., 2013; Kappen et al., 2022; Van Puyvelde et al., 2018).

Psychosocial stressors are one of the most potent and ecologically-valid stressors and are induced in situations of social evaluation or exclusion (Dickerson & Kemeny, 2004; Kappen et al., 2022; Kirschbaum & Hellhammer, 1994). Our former study used a similar paradigm, which confirmed the stress induction was successful, but the study was limited to pre-, and

post-stressor measurements, thus lacking a control/neutral condition (Kappen et al., 2022). In our current study, a successful stress induction will be determined based on both self-reports (valence, arousal) throughout the paradigm and physiological activity (cardiac acceleration and deceleration) during the negative versus the control feedback exposure. We expect to find decreases for jitter (vocal frequency variation) and shimmer (vocal intensity variation) as that is the direction of observed effects, however, results are heterogeneous (Giddens et al., 2013; Van Puyvelde et al., 2018). Harmonics-to-noise ratio (HNR; added noise in the voice) has been shown to decrease in the context of a physical stressor (i.e., workout), and mixed results are observed in the context of psychological stress (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Mendoza & Carballo, 1998). We included HNR as it is frequently described in the literature and changes are observed, however, we have no expected direction for this effect. Lastly, we expect participants' speech rates to increase. This feature is not always included in analyses but has shown robust results in free speech settings (Giddens et al., 2010, 2013; Rothkrantz et al., 2004). Therefore, we believe we contribute to the existing literature by 1) a matter of developing a new psychophysiological methodology for stress measurement, and 2) do so by designing a solid experimental paradigm that allows us to induce and validate experienced stress on an individual (within-participant) basis.

Despite mixed results in acoustic changes due to acute stress, some acoustic features get described more often than others in literature. As such, we will focus on these key acoustic speech features from literature in the current study. The most homogeneous results are found for the Fundamental Frequency (F0) of the voice, which refers to the frequency at which the vocal cords vibrate (i.e. pitch), seeing it generally increases with increased stress (Giddens et

al., 2013; Van Puyvelde et al., 2018). We expect to find decreases for jitter (vocal frequency variation) and shimmer (vocal intensity variation) as that is the direction of observed effects, however, results are heterogeneous (Giddens et al., 2013; Van Puyvelde et al., 2018). Harmonics-to-noise ratio (HNR; added noise in the voice) has been shown to decrease in the context of a physical stressor (i.e., workout), and mixed results are observed in the context of psychological stress (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Mendoza & Carballo, 1998). We included HNR as it is frequently described in the literature and changes are observed, however, we have no expected direction for this effect. Lastly, we expect participants' speech rates to increase. This feature is not always included in analyses but has shown robust results in free speech settings (Giddens et al., 2010, 2013; Rothkrantz et al., 2004).

In summary, in the presented study in this paper, we analyze high-quality read-out-loud speech fragments collected from a large (non-actor) sample in a within-subject stress paradigm, containing both a control and a negative feedback condition. In doing so, we can verify the experienced stress by the subjects (based on both self-reports and objective physiological measures). We present trustworthy and ecologically valid information on the distilled effects of (psychological) stress on key acoustic speech features, such as F0 (Fundamental Frequency; pitch), jitter (vocal frequency variation), shimmer (vocal intensity variation), harmonics-to-noise ratio (HNR; added noise in the voice), and speech rate (Giddens et al., 2013; Kappen et al., 2022; Van Puyvelde et al., 2018). These results will be the basis for further deterministic modeling, digital biomarker design, and/or analysis of speech in the context of stress detection and emotion recognition. Moreover, these results will contribute to

the field of vocal markers of neuropsychiatric conditions, considering stress is suggested to be a core psychological mechanism associated with a range of mood disorders (e.g. see: Cho et al., 2022; Koops et al., 2021; Voppel et al., 2022).

3.3. Methods

This study was part of a larger project that investigates the effects of a (psychosocial) stressor on neural correlates. Results of electrophysiological correlates will be published elsewhere. Moreover, collected data that was not part of the current paper's research objectives will only be described in the supplemental materials.

3.3.1. Participants

A convenience sample of 77 subjects (50 women, 27 men, age $M = 23.13$, $SD = 6.19$) was recruited through social media with a post containing the cover story that this study gauges future (academic) success. Upon registration, participants were checked for exclusion criteria (see supplemental material). The study was conducted in accordance with the declaration of Helsinki and received ethical approval from the Ghent University hospital ethical committee (registration number: B670201940636). All participants gave written informed consent before participating and were debriefed afterward on the true purpose of the study. A 30 Euro compensation fee was awarded upon completion through bank transfer.

3.3.2. Apparatus and procedure

3.3.2.1. Read-out-loud text “Marloes”

To collect speech fragments, participants were instructed throughout the experiment to read a standardized text of five sentences out loud. This “Marloes” text is often used in Dutch speech therapy due to the text containing a similar frequency distribution as occurs in the Dutch language (Van de Weijer & Slis, 1991; See supplemental material for full text). Participants were instructed to read the text out loud five times at home prior to the experiment to familiarize themselves with it and to exclude novelty effects (Kappen et al., 2022).

3.3.2.2. On-site experimental session

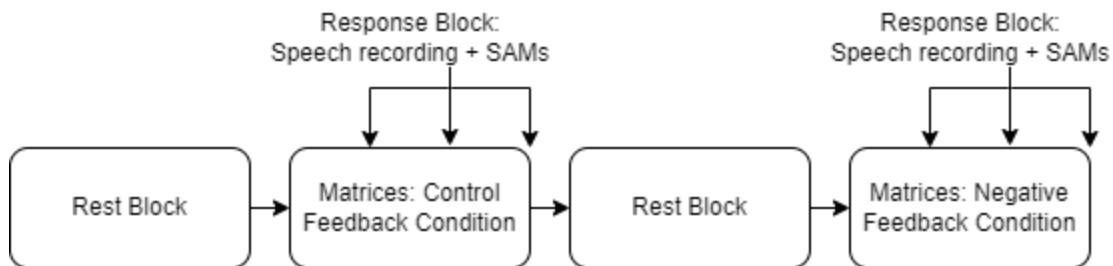
The experiment was conducted in a dedicated room in the Department of Neurology at the Ghent University Hospital. The ECG (ElectroCardioGram) electrodes were applied (1 electrode just below the left collarbone and 1 electrode on the left lower rib), after which the experimental phase commenced. Participants were seated in an upright position in front of a computer screen (Dell E2216H).

The experiment was carried out on a computer (Dell, Windows 10, experiment designed in E-Prime 2.0 (Schneider et al., 2002) and a tablet (Huawei MediaPad M5, custom-designed Android app; see <https://osf.io/78g9s/>). The experimental task was completed on the computer, while self-reports and speech collection were done on the tablet to circumvent any built-in preprocessing of the audio signals in E-Prime 2.0. The experiment started with a 10-minute resting block (to achieve habituation) in which participants closed their eyes to ensure a relaxed state. After this, the *Control* feedback condition commenced. After this

condition, there was another 10-minute resting block, followed by a *Negative feedback condition* (see Figure 1). At three fixed points throughout the two feedback conditions, i.e. at one-third, two-thirds of the way, and at the end of the condition, participants were offered a *Response Block*. The *Response Block* was executed on the tablet and starts with the out-loud reading of the “Marloes” text. After this, participants answered Self-Assessment Manikin scales (SAMs; Valence, Arousal; see supplemental material; Bradley & Lang, 1994) by responding which manikin best represented their feelings. Valence was described as how negative/positive they felt at that instance, whereas arousal was described as how calm/restless they felt (Russell, 1980).

Figure 1

Flowchart of experimental design.



Note. Two feedback conditions (control/negative), both preceded by a 10-minute resting period. Speech and self-reports (SAMs; Self-Assessment Manikins) were recorded at three points in each condition; yielding 3 data points per participant per condition.

3.3.2.3. Trial

During each (experimental) feedback condition (control and negative), participants were offered three subblocks. Each subblock ended when participants either completed 11 trials or 6 minutes had passed since the start of that subblock. During each trial, participants were

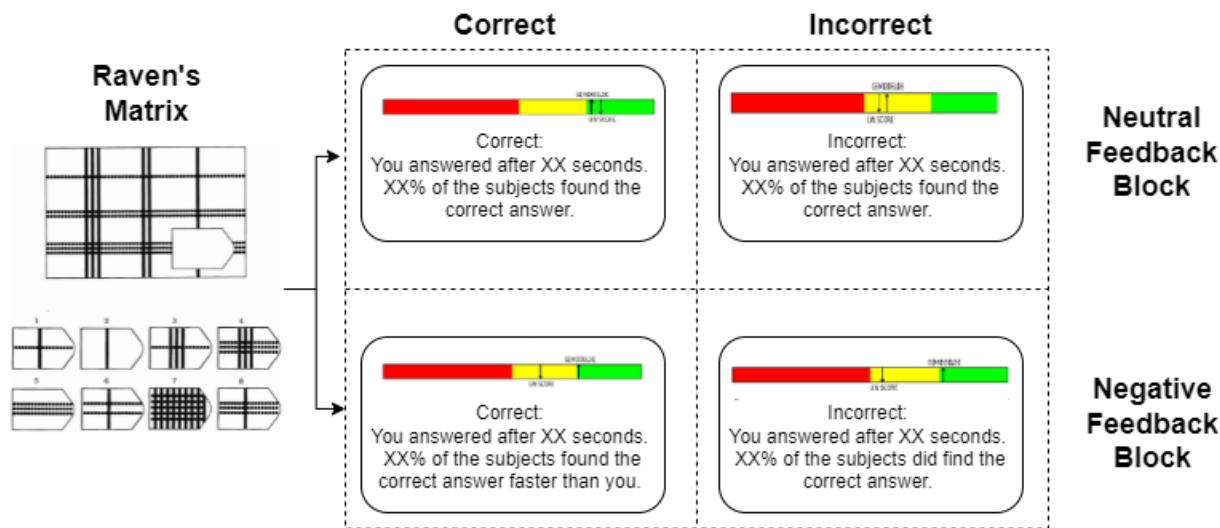
offered a Raven's Matrix (Raven, 2000): a 3x3 raster of illustrations that follows a certain pattern/logic with one spot open which they had to fill in (see Figure 2 for a visual representation of the trial-feedback sequence). Participants could choose from 8 different options to fill in the blank spot and reply with their right hand using the Numpad on a regular US-layout keyboard. Above the raster, a countdown was shown indicating the number of seconds left before the exercise timed out. The allowed time differed per trial and was dependent on the difficulty level; participants had either 20 (for easy levels), 45 (for medium levels), or 100 seconds (for difficult levels) to respond. The difficulty level was validated in a pilot test to have it balanced over subblock as well as over experimental (feedback) condition and randomized per participant. After either a response or a time-out, a feedback screen was shown. The feedback screen consisted of three elements: 1) At the top, a red, yellow, and green bar was displayed with vertical arrows indicating their performance and the average group performance; 2) It was indicated whether their answer was correct or incorrect (or timed-out), this was always in accordance with their actual answer; and 3) At the bottom of the screen, their response time was indicated as well as a textual comparison to the reference group. This paradigm was inspired by the Montreal Imaging Stress Task (MIST; Dedovic et al., 2005), but we used Raven's Matrices rather than mathematical puzzles to reduce the stress that is experienced by people who are not good at math, and because something more similar to an IQ test would fit better in the cover story. In addition, the time pressure did not vary between the control and negative feedback conditions in our paradigm, shifting it from a cognitive stressor to a psychosocial stressor.

3.3.2.4. Stress induction

Both (experimental) feedback conditions were essentially identical, except for the comparison group and the received feedback. During the control feedback condition, to increase the credibility of the cover story, participants are told that they are being compared to a sample of people who are randomly sampled from the population (average individuals). In order to successfully induce stress in a repeated MIST paradigm, there must be a credible social evaluation. As such, a credible cover story as used here contributes to the ecological validity of the stressor (De Calheiros Velozo et al., 2021). When seeing the feedback, they are shown to be performing on par with the reference group to ensure control/neutral feedback. During the negative feedback condition, participants are told that they are being compared to a group of highly educated, well-performing individuals. When observing the colored feedback bar, they are shown to be performing increasingly worse over the course of each subblock, irrespective of their actual performance to enable the negative feedback (see supplemental materials for a visual representation of increasingly negative feedback over subblocks). Moreover, whenever the participant found the correct answer, the feedback would indicate that XX% of the reference group found the correct answer faster than them, still inducing a negative comparison even when they did give a correct answer. Feedback was displayed for 6 seconds, after which the next trial commenced. See Figure 2 for a flow matrix of a trial sequence including examples of the different feedback types for both correct and incorrect responses in both the control and negative feedback conditions. Prior to the experimental task, participants are told a cover story that the study's objective is to predict future success and that this task was commonly used in IQ tests and was valid in this prediction.

Figure 2

A flow diagram of the trial sequence



Note. Flow diagram of a trial sequence including both options (correct/incorrect) for both conditions (control/negative feedback) after a response to a Raven's Matrix. On the colored bar, two arrows indicated their own performance and the comparison group performance (accompanied by text) which shifted after every trial. The displayed colors were indicators of overall performance. Matrices varied per condition but were balanced in difficulty over feedback conditions and subblocks.

3.3.3. Extraction of speech features

All audio fragments were manually checked for quality, and only full and clear (i.e., no excessive clipping or background noise) recordings were included. 31 control and 34 negative feedback recordings were of insufficient quality and were not used in subsequent analyses, resulting in 209 control feedback recordings of 71 out of 77 participants and 206 negative feedback recordings of 69 out of 77 participants. Features were extracted using OpenSmile 2.3.0 (Eyben et al., 2010) with the GeMAPS configuration (Eyben et al., 2015), a minimalistic acoustic feature set frequently used in voice research and affective computing. From this

feature set, Fundamental frequency (F0), Jitter, Shimmer, Harmonics-to-Noise Ratio (HNR), voiced segment length, and mean voiced segments per second (a proxy for speech speed) were selected. It is important to remark that the features were computed locally via a sliding-window and then mean-aggregated over the whole utterance, thus not displaying high temporal changes. For detailed information regarding feature calculation and extraction procedure, we refer the reader to Eyben et al. (2010) and Section 6.1 of Eyben et al. (2015).

3.3.4. Data-Analysis

All data were preprocessed using Python 3.9.6 and statistical analyses were performed using R4.1.1 (for detailed version information of the software and packages used, see supplemental materials). As a part of our manipulation check, we collected ECG data throughout the task and analyzed the event-related cardiac reactivity during feedback exposure (See for a similar approach: Gunther Moor et al., 2010; van der Veen et al., 2019). The recorded IBIs (InterBeat Intervals; time in ms between individual heartbeats) were corrected for artifacts using our custom code (see <https://osf.io/78g9s/>). We assessed the data quality, resulting in the use of 73 out of 77 participants' cardiac responses. Twelve IBIs were selected around the feedback: the IBI/heartbeat closest to feedback onset (from now on called IBI_0), three IBIs preceding the feedback (IBI_{-3} , IBI_{-2} , IBI_{-1}), and eight IBIs during the feedback exposure (IBI_1 to IBI_8). By the 8th IBI collected after feedback onset, 75% of the trials had passed the 6 seconds of feedback exposure (See supplemental materials). In accordance with the literature (Gunther Moor et al., 2010; van der Veen et al., 2014), we referenced IBI difference scores to the second IBI preceding the feedback onset (IBI_{-2}) for each trial. These referenced IBI difference scores are referred to as delta IBIs throughout the manuscript.

To control for the potential effect of sex on the different speech features, sex was considered as a fixed effect for each individual model prior to statistical inference. However, to make sure our models were parsimonious, we bottom-up tested whether adding sex as an independent variable to the model improved each model's fit. For each dependent variable, we compared models that included and excluded sex, and it was only included in the model if it showed to be a significant contributor after comparing models with reducing complexity using χ^2 goodness-of-fit tests within the 'anova()' function. The statistical significance level was set to $p < .05$ and in the results section, we describe for each individual model whether sex was a significant contributor and thus included.

For the manipulation checks (i.e., IBIs and self-reports) and speech features (i.e., F0, jitter, shimmer, harmonics-to-noise ratio, and speech rate), we used the 'lme4' (Bates et al., 2014) and 'car' (Bates et al., 2014; Fox et al., 2012) R packages to fit generalized linear mixed models (GLMMs). The IBI model featured delta IBI (referenced to IBI_{-2} ; relative change of IBI as compared to IBI_{-2} indicating acceleration (i.e., negative delta IBI) and deceleration (i.e., positive delta IBI) of the heart) as a dependent variable with 12 levels (IBI_{-3} to IBI_8), feedback condition (2 levels; control vs negative feedback condition) as a fixed effect, and the subject as a random intercept. The ANOVA comparison for the model including vs excluding sex as a fixed effect showed no significant improvement and sex was thus excluded from the model. The valence and arousal models followed a similar structure. Either valence or arousal as the dependent variable on a 7-point Likert scale, having 2 levels of feedback condition (control vs negative feedback) as a fixed effect, and subject as a random intercept. Again, the ANOVA analysis showed no significant contribution of sex to these models, and sex was thus also excluded from these models as a fixed effect. The models for the speech features were identical to the

valence/arousal models, with feedback condition (2 levels; control vs negative feedback, each containing 3 data points per participant) as a fixed effect, and subject as random intercept whilst controlling for sex by including it as a fixed effect if aforementioned method showed it to have a significant contribution to the model, resulting in the (G)LMM formulas of the following structure (in R notation for ‘lme4’); $DependentVariable \sim Condition + Sex + (1|ID)$ or $DependentVariable \sim Condition + (1|ID)$. Each dependent variable’s specific model is also reported in the results section.

The sum of squares was estimated using the type III approach, and the statistical significance level was set to $p < .05$. Follow-up tests with pairwise comparisons of the EMMs (estimated marginal means) were performed with the ‘emmeans’ package (Lenth, 2018), using false discovery rate (FDR) to correct for multiple testing (Benjamini & Hochberg, 1995).

3.4. Results

3.4.1. Manipulation check

A manipulation check was conducted to verify whether participants experienced increased stress during the negative feedback condition compared to the control feedback condition by setting side-by-side self-reports and physiological activity during both feedback conditions.

3.4.1.1. Self-reports

At three different set moments during each condition, participants answered how they were feeling with regard to *valence* and *arousal* using SAMs (Self-Assessment Manikins). Valence, with the formula $Valence \sim Condition + (1|ID)$, was best described by an LMM (linear

mixed model) with AIC equal to 1055 (Akaike Information Criterion). The LMM showed a significant decrease of valence in the negative feedback condition (Figure 3a) $\chi^2(1) = 83.01$, $b = .594$, $SE = .0652$, $t = 9.111$, $d = .91$, $p < .001$. Arousal, with the formula $Arousal \sim Condition + (1|ID)$, was best described by a GLMM (generalized linear mixed model) with Gamma distribution and identity link, AIC = 1069. The GLMM showed a significant decrease in arousal during the negative feedback condition (Figure 3b) $\chi^2(1) = 4.47$, $b = .135$, $SE = .0639$, $z = 2.116$, $d = .50$, $p = .034$.

Figure 3a

Valence between feedback conditions

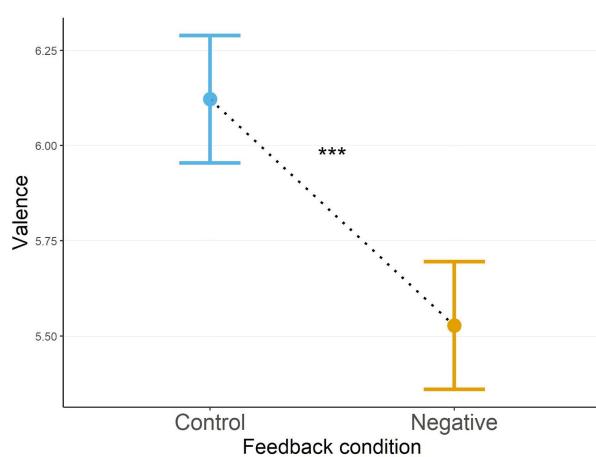
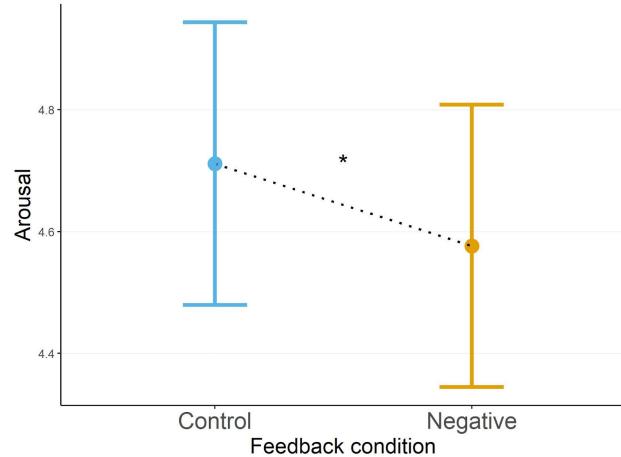


Figure 3b

Arousal between feedback conditions



Note. Estimated marginal means (EMMs) of self-reported valence(a) and arousal(b) during control-, and negative feedback condition after controlling for sex. Error bars depict standard error of the means (SEMs), asterisks indicate significance levels. * indicates $p < .05$. *** indicates $p < .001$.

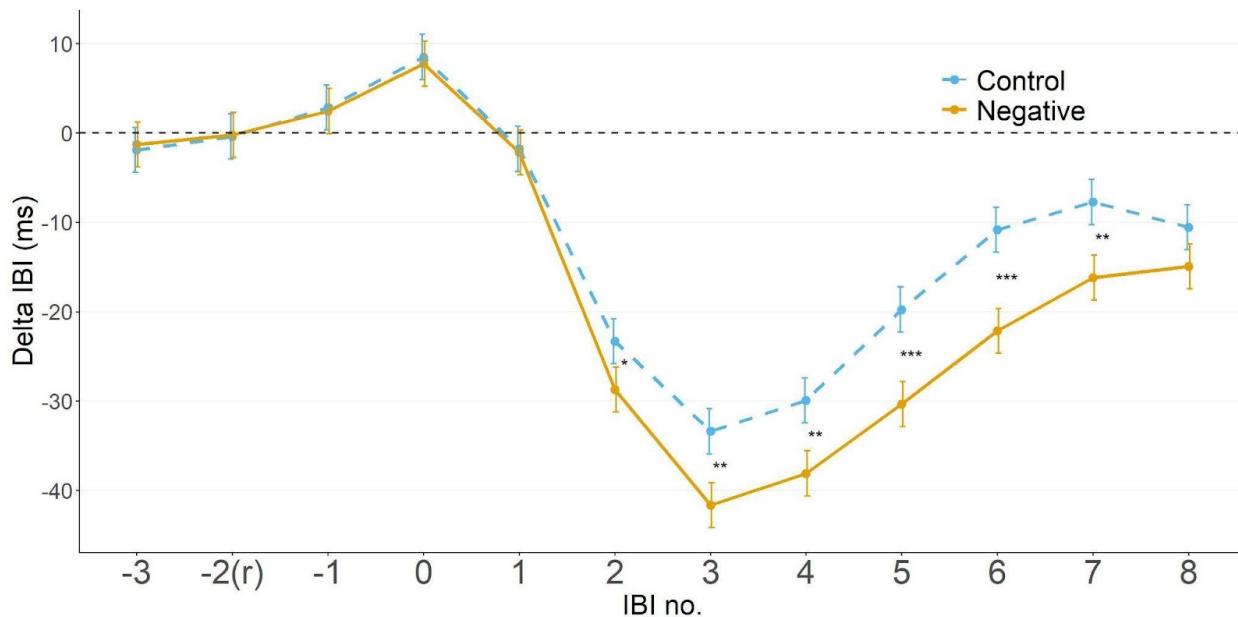
3.4.1.2. Physiological activity

Delta IBI's (cardiac interbeat interval) were calculated during the feedback period (6s) after a trial was completed. Considering the presence of non-positive values (cardiac acceleration: negative delta's), an LMM with the formula $DeltaIBI \sim Condition * IBI_{no} + (1|ID)$

was fit to the data. The LMM showed a Condition \times IBI_{no} interaction effect $\chi^2(11, N = 73) = 33.49$, $p < .001$, showing more acceleration in heart rate during observation of the negative feedback than the control feedback. However, as our main focus is on the IBIs following feedback exposure rather than on all IBIs; follow-up pairwise comparisons were executed between the two conditions at every individual IBI, on which we applied FDR (False Discovery Rate) correction (Benjamini & Hochberg, 1995). We observe significant effects for IBI₂ to IBI₇ (Figure 4, Table 1), showing that heart rate acceleration is larger during exposure to negative feedback as compared to control feedback from IBI₂ to IBI₇ (see Table 1).

Figure 4

Delta Interbeat intervals in response to feedback exposure between feedback conditions



Note. Estimated Marginal Means (EMMs) for delta IBI's (in ms, referenced to IBI₂) of control-, and negative-feedback trials, with IBI₀ being IBI closest to feedback exposure onset. Error bars depict the standard error of the means (SEMs), asterisks indicate significance levels.* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Table 1

Individual contrasts at different IBIs between control-, and negative-feedback condition trials.

	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>
IBI ₀	.760	2.58	.295	.770
IBI ₁	.393	2.58	.153	.879
IBI ₂	5.381	2.58	2.089	.037
IBI ₃	8.282	2.58	3.215	.001
IBI ₄	8.181	2.58	3.176	.002
IBI ₅	10.574	2.58	4.105	<.001
IBI ₆	11.301	2.58	4.388	<.001
IBI ₇	8.437	2.58	3.276	.001
IBI ₈	4.356	2.58	1.691	.091

Note. *b* is the beta coefficient, *SE* is the standard error of the difference, *z* is the z-ratio, *p* is the p-value. P-values are FDR corrected.

3.4.2. Speech feature analysis

For each of the speech features, a series of (G)LMM (generalized linear mixed models) were fitted to increase the likelihood of using a statistical model that best fits the underlying distribution. Model selection was performed using the AIC. To minimize the likelihood of Type 1 errors, FDR correction was applied over all p-values for the different speech features.

3.4.2.1. Harmonics-to-Noise Ratio (HNR)

The distribution for HNR was best represented by an LMM (AIC = 765) and showed a significant main effect for the feedback condition after controlling for sex, with HNR being

significantly higher during the negative-, versus the control-feedback condition, $\chi^2(1) = 8.17$, $b = .127$, $SE = .0444$, $z = 2.858$, $d = .68$, $p = .005$ (Figure 5a).

3.4.2.2. Shimmer

Shimmer was best represented by a GLMM with Gamma distribution and identity-link ($AIC = -927$) and showed a significant main effect for the feedback condition after controlling for sex, with shimmer being significantly lower during the negative-, versus the control-feedback condition, $\chi^2(1) = 8.30$, $b = .019$, $SE = .006$, $z = 2.881$, $d = .68$, $p = .004$ (Figure 5b).

Figure 5a

HNR between feedback conditions

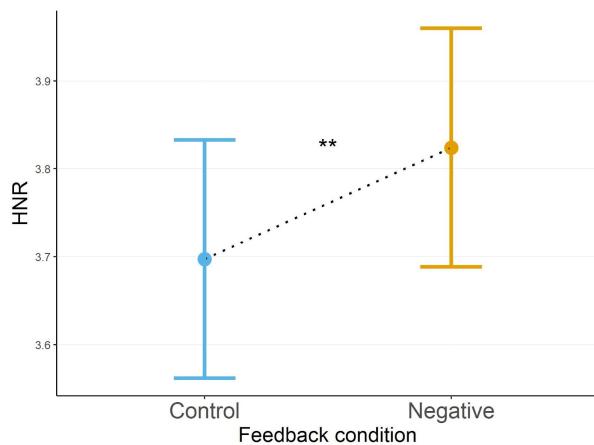
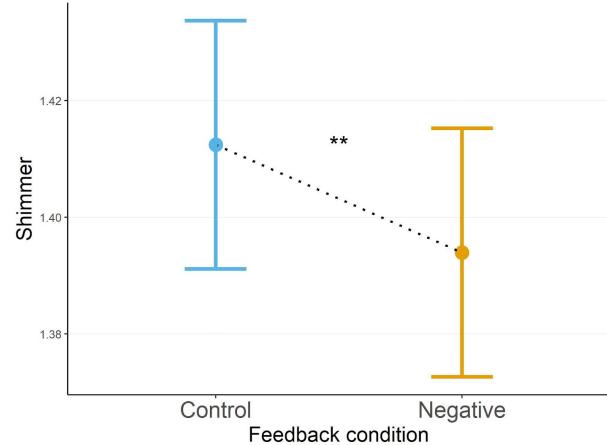


Figure 5b

Shimmer between feedback conditions



Note. Estimated marginal means (EMMs) of HNR(a) and Shimmer(b) during control-, and negative feedback conditions after controlling for sex. Error bars depict standard error of the means (SEMs), asterisks indicate significance levels. ** indicates $p < .01$.

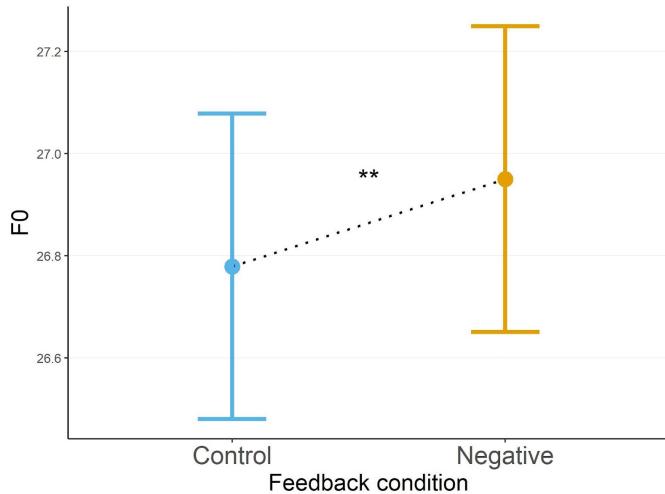
3.4.2.3. Fundamental Frequency (F0)

F0 was best represented by a GLMM with Gamma distribution and identity-link ($AIC = 961$) and showed a significant main effect for the feedback condition after controlling for sex,

with F0 being significantly higher during the negative-, versus the control-feedback condition, $\chi^2(1) = 7.60$, $b = .171$, $SE = .062$, $z = 2.756$, $d = .65$, $p = .006$ (Figure 6).

Figure 6

Fundamental Frequency (F0) between feedback conditions



Note. Estimated marginal means (EMMs) of Fundamental Frequency (F0) during control-, and negative feedback conditions after controlling for sex. Error bars depict standard error of the means (SEMs), asterisks indicate significance levels. ** indicates $p < .01$.

3.4.2.4. Jitter, voiced segments per second, mean voiced segment length

No significant effects were observed for either jitter, voiced segments per second, or mean voiced length. Jitter was fit with an LMM ($AIC = -2725$), $\chi^2(1) = 0.975$, $p = .32$. Voiced segments per second were fit with a GLMM with Gamma fit and identity link ($AIC = 20$), $\chi^2(1) = 2.41$, $p = .12$. Mean voiced segment length was fit with a GLMM with Gamma fit and identity link ($AIC = -2002$), $\chi^2(1) = 3.25$, $p = .07$.

3.5. Discussion

In this study, we aimed to examine the effects of stress, induced by a highly controlled social evaluative threat stressor, on different key speech features (fundamental frequency; F0, harmonics-to-noise ratio; HNR, jitter, shimmer, and speech rate). Participants performed tasks within two conditions, one with control feedback and one with negative comparative feedback, that shared the same overall design. Each feedback condition contained three subblocks of abstract reasoning puzzles (i.e., Raven's matrices) to be solved under time pressure. In the first condition, participants were told that they were being compared to a group of people who were randomly sampled from the population (average individuals) and received feedback after each trial that indicated that they were performing on par with the reference group (control feedback condition). In the second condition, participants were told that they were being compared to individuals that achieved significant academic or professional success (to increase the credibility of a sudden drop in relative performance) and received feedback after each trial, indicating that they were performing increasingly worse compared to the comparison group throughout each subblock (negative feedback condition). During each condition, participants were asked at three moments to read a standardized text out loud. We extracted several key acoustic features from these fragments to gain insight into the effects of acute (psychosocial) stress on speech. These features were selected based on previous research in which they were deemed important in the context of stress (Giddens et al., 2013; Kappen et al., 2022; Van Puyvelde et al., 2018). However, this study is the first to use a within-participant design that verifies a successful stress induction using both self-report and physiological measures in a (non-actor) sample. We verified a successful stress induction based on decreased self-reported

valence scores and increased heart rate acceleration during the negative feedback condition. Increased heart rate acceleration during negative feedback is consistent with the notion of increased sympathetic reactivity due to stress exposure (Taelman et al., 2009; Vrijkotte et al., 2000). However, we did encounter a decrease in self-reported arousal during the negative feedback condition, which is the opposite of what we expected. This result could potentially be an order effect, such as tiredness, due to the negative feedback condition always being subsequent to the control condition. Nonetheless, we conclude a successful stress induction due to the increased heart rate acceleration and decreased self-reported valence during the negative feedback condition.

We observed an increase of F0 (fundamental frequency; pitch) during the negative feedback condition as compared to the control feedback condition. This was expected since increases in F0 in response to an acute (psychosocial) stressor are commonly reported in the literature (Giddens et al., 2013; Van Puyvelde et al., 2018). Furthermore, we observed a significant increase in HNR (harmonics-to-noise ratio; added noise in the voice) in the negative feedback condition as compared to the control feedback condition. In the past, no clear results have been found with regard to this parameter, as it has shown to decrease in the context of physical stress tasks (e.g., workout) and has shown mixed results in the context of cognitive load/psychological stress (Giddens et al., 2013; Kirschbaum et al., 1994; Godin et al., 2012; Godin & Hansen, 2015). We also found a decrease in shimmer (vocal intensity variation) during the negative as compared to control feedback. The effects of stress on shimmer are less pronounced, where some studies indicate no changes and others a decrease in shimmer after different stress induction procedures (Giddens et al., 2013; Mendoza & Carballo, 1998). Nonetheless, we found a clear decrease in shimmer during stress, which make sense due to its

vowel-level relationship with heart rate, a central component in stress reactivity (Giddens et al., 2013; Orlikoff, 1990). Yet, future research should revisit the direct, trial-based relationship between shimmer heart rate.

No effects have been found for jitter (vocal frequency variation). However, as proposed by Van Puyvelde et al. (2018), acoustic speech parameters should not only be considered in their own regard but also as combined patterns of multiple speech parameters that may respond in a simultaneous manner. HNR has been demonstrated to be more sensitive to subtle differences in vocal function than jitter (Awan & Frenkel, 1994; Giddens et al., 2013), and former network analysis has shown a strong negative relationship between changes in jitter and changes in HNR after psychosocial stress induction (Kappen et al., 2022). Moreover, jitter is mainly affected due to a lack of control of the vocal fold vibration (Teixeira et al., 2013). The lack of a significant difference could be explained by the nature of the speech fragments that we analyzed. It could be argued that when people read a text out loud, as opposed to speaking freely, they could be using a ‘reading voice’ that minimizes these types of effects due to read speech being significantly different from spontaneous speech, both acoustically and linguistically (Nakamura et al., 2008). A similar argument could be made to explain a lack of effect found for the speech rate, as when someone reads out loud, one of their focuses is understandability for potential listeners (Nakamura et al., 2008). In addition, since the text and one’s ability to process this both influence a minimal and maximum speech rate, it can be expected that this measure is limited by the speech recording paradigm.

The current results are generated in a well-controlled experimental setting, using a stressor with a control condition that also contains time pressure. Therefore, the presented results are indicative of how speech as a biomarker reacts to actual stress as induced by

negative evaluation, rather than cognitive load or time pressure. However, it should be noted that the current study was limited in its design as the stress condition was always preceded by the control condition rather than being randomized. We chose to use this design because it would enable us to integrate an active control condition which also contained cognitive load and time pressure just as the stress condition, isolating the results to just the experience of negative evaluation. Counterbalancing the order of the control and stress condition would have only been possible by either severely lengthening the design or by testing on multiple days, due to the duration of the recovery phase after a stressor. This limitation could introduce order effects, such as a fatigue or repetition effect, that confound our presented results. Nevertheless, considering our results are in line with previous research, we believe that if this effect influenced our results, they potentially reduced the observed effect sizes. Supplemental analyses indeed show no effect of repetition on the speech features (<https://osf.io/gq7aw>).

The current study evaluated several key acoustic speech features in an isolated situation; by using read-out-loud speech, potential interference from specific word choices was eliminated. However, to work towards a real-world application for speech as a biomarker for stress, features in spontaneous speech fragments should also be tested. Future studies should therefore move towards a speech collection paradigm that enables participants to speak freely. However, certain considerations should be made here, since completely free speech would introduce a number of noise factors that could make the comparison of certain acoustic features between different conditions close to impossible. In order to more closely simulate free speech in a controlled setting, future research should focus on using a speech collection paradigm in which participants are semi-spontaneous in their speech by, for example, controlling the topics they can talk about, whilst limiting any extra cognitive load of active

recall. Shifting towards free speech should give us more insight into the robustness of these acoustic speech features in more spontaneous settings, and will additionally enable us to investigate other linguistic features of speech in the context of stress, such as syntax, prosody, and semantics. Combining semantics with syntax and acoustic features will indicate the potential of real-life applications for stress monitoring using speech signals. Moreover, we invite others to use our data and test other features that they deem promising (<https://osf.io/78g9s/>).

To conclude, we collected repeated read-out-loud speech fragments of participants in a social evaluative threat stress induction paradigm which we validated through self-reports and psychophysiological responses. We were able to give valid and reliable results for the effects of psychosocial stress on F0, HNR, and shimmer, and were not able to find effects on jitter and speech rate. Therefore, we conclude that changes in F0, HNR, and shimmer are shown to be present in speech after stress irrespective of a person's language construction capability. As such, this study shows that speech is a promising biomarker for stress, on top of it being affordable, non-intrusive, and easy to collect and therefore easy to implement in everyday settings. Future studies should focus on replicating our findings to test the robustness of the effect of stress on these acoustic speech features. In addition, different speech production paradigms should be developed and tested in order to move towards more spontaneous speech and test the external validity in more naturalistic settings. Lastly, this would also enable us to increase the range of speech features that can be informative in the context of stress, such as semantics and syntax.

3.6 References

- Awan, S. N., & Frenkel, M. L. (1994). *Improvements in Estimating the Harmonics-to-Noise Ratio of the Voice*. 8(3), 255–262.
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9/10), 341–345.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- Cho, S., Fusaroli, R., Pelella, M. R., Tena, K., Knox, A., Hauptmann, A., Covello, M., Russell, A., Miller, J., Hulink, A., Uzokwe, J., Walker, K., Fiumara, J., Pandey, J., Chatham, C., Cieri, C., Schultz, R., Liberman, M., & Parish-morris, J. (2022). Identifying stable speech-language markers of autism in children: Preliminary evidence from a longitudinal telephony-based study. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 40–46. <https://doi.org/10.18653/v1/2022.clpsych-1.4>
- De Calheiros Velozo, J., Vaessen, T., Pruessner, J., Van Diest, I., Claes, S., & Myin-Germeys, I. (2021). The repeated Montreal Imaging Stress Test (rMIST): Testing habituation,

sensitization, and anticipation effects to repeated stress induction.

Psychoneuroendocrinology, 128, 105217.

<https://doi.org/10.1016/j.psyneuen.2021.105217>

Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., & Pruessner, J. C. (2005). The Montreal Imaging Stress Task: Using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience*, 30(5), 319–325.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), 355–391. <https://doi.org/10.1037/0033-2909.130.3.355>

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, 49(December 2017), 146–169. <https://doi.org/10.1016/j.yfrne.2018.03.001>

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202. <https://doi.org/10.1109/TAFFC.2015.2457417>

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE - The Munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 1459–1462. <https://doi.org/10.1145/1873951.1874246>

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ..., & Heiberger, R. (2012).

Package ‘car.’ *Vienna: R Foundation for Statistical Computing*.

- Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, 13(1), 440–460. <https://doi.org/10.1109/TAFFC.2019.2927337>
- 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.e21-390.e29. <https://doi.org/10.1016/j.jvoice.2012.12.010>
- Giddens, C. L., Barron, K. W., Clark, K. F., & Warde, W. D. (2010). Beta-Adrenergic Blockade and Voice: A Double-Blind, Placebo-Controlled Trial. *Journal of Voice*, 24(4), 477–489. <https://doi.org/10.1016/j.jvoice.2008.12.002>
- Godin, K. W., & Hansen, J. H. L. (2015). Physical task stress and speaker variability in voice quality. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015(1). <https://doi.org/10.1186/s13636-015-0072-7>
- Godin, K. W., Hasan, T., & Hansen, J. H. L. (2012). Glottal waveform analysis of physical task stress speech. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2(January 2012), 1646–1649.
- Gunther Moor, B., Crone, E. A., & van der Molen, M. W. (2010). The Heartbreak of Social Rejection: Heart Rate Deceleration in Response to Unexpected Peer Rejection. *Psychological Science*, 21(9), 1326–1333. <https://doi.org/10.1177/0956797610379236>
- Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54(2), 910–921. <https://doi.org/10.3758/s13428-021-01670-x>
- Kirschbaum, C., & Hellhammer, D. H. (1994). Salivary cortisol in psychoneuroendocrine

research: Recent developments and applications. *Psychoneuroendocrinology*, 19(4), 313–333. <https://doi.org/10.1111/j.0269-8463.2004.00893.x>

Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2021). Speech as a Biomarker for Depression. *CNS & Neurological Disorders - Drug Targets*, 20. <https://doi.org/10.2174/1871527320666211213125847>

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*.

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12(3), 263–273. [https://doi.org/10.1016/S0892-1997\(98\)80017-9](https://doi.org/10.1016/S0892-1997(98)80017-9)

Miller, G. E., Cohen, S., & Ritchey, A. K. (2002). Chronic psychological stress and the regulation of pro-inflammatory cytokines: A glucocorticoid-resistance model. *Health Psychology*, 21(6), 531–541. <https://doi.org/10.1037/0278-6133.21.6.531>

Monroe, S. M. (2008). Modern Approaches to Conceptualizing and Measuring Human Life Stress. *Annual Review of Clinical Psychology*, 4(1), 33–52. <https://doi.org/10.1146/annurev.clinpsy.4.022007.141207>

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171–184. <https://doi.org/10.1016/j.csl.2007.07.003>

Orlikoff, R. F. (1990). Vowel amplitude variation associated with the heart cycle. *Journal of the Acoustical Society of America*, 88(5), 2091–2098. <https://doi.org/10.1121/1.400106>

Paulmann, S., Furnes, D., Bøkenes, A. M., & Cozzolino, P. J. (2016). How psychological stress affects emotional prosody. *PLoS ONE*, 11(11), 1–21. <https://doi.org/10.1371/journal.pone.0165700>

<https://doi.org/10.1371/journal.pone.0165022>

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, 41(1), 1–48. <https://doi.org/10.1006/cogp.1999.0735>

Rothkrantz, L. J. M., Wiggers, P., Van Wees, J. W. A., & Van Vark, R. J. (2004). Voice stress analysis. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3206, 449–456. <https://doi.org/10.4135/9781452229300.n1969>

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178. <https://doi.org/10.1037/h0077714>

Sandi, C. (2013). Stress and cognition. *WIREs Cognitive Science*, 4(3), 245–261. <https://doi.org/10.1002/wcs.1222>

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. (2.0). Pittsburgh: Psychology Software Tools Inc.

Slavich, G. M. (2016). Life Stress and Health: A Review of Conceptual Issues and Recent Findings. *Teaching of Psychology*, 43(4), 346–355. <https://doi.org/10.1177/0098628316662768>

Taelman, J., Vandeput, S., Spaepen, A., & Van Huffel, S. (2009). Influence of Mental Stress on Heart Rate and Heart Rate Variability. In J. Vander Sloten, P. Verdonck, M. Nyssen, & J. Haueisen (Eds.), *4th European Conference of the International Federation for Medical and Biological Engineering* (Vol. 22, pp. 1366–1369). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-89208-3_324

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, 9, 1112–1122. <https://doi.org/10.1016/j.protcy.2013.12.124>

van der Veen, F. M., Burdzina, A., & Langeslag, S. J. E. (2019). Don't you want me, baby? Cardiac and electrocortical concomitants of romantic interest and rejection. *Biological Psychology*, 146, 107707. <https://doi.org/10.1016/j.biopspsycho.2019.05.007>

van der Veen, F. M., van der Molen, M. W., Sahibdin, P. P., & Franken, I. H. A. (2014). The heart-break of social rejection versus the brain wave of social acceptance. *Social Cognitive and Affective Neuroscience*, 9(9), 1346–1351. <https://doi.org/10.1093/scan/nst120>

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>

Voppel, A. E., de Boer, J. N., Brederoo, S. G., Schnack, H. G., & Sommer, I. E. C. (2022). *Semantic and phonetic markers in schizophrenia-spectrum disorders; a combinatory machine learning approach* [Preprint]. Psychiatry and Clinical Psychology. <https://doi.org/10.1101/2022.07.13.22277577>

Vrijkotte, T. G. M., van Doornen, L. J. P., & de Geus, E. J. C. (2000). Effects of Work Stress on Ambulatory Blood Pressure, Heart Rate, and Heart Rate Variability. *Hypertension*, 35(4), 880–886. <https://doi.org/10.1161/01.HYP.35.4.880>

3.7. Supplemental Materials

3.7.1. Acknowledgments

We thank Floor Depetele and Frauke De Craene for their help in collecting the data.

3.7.2. Funding

This research was supported by a Grant for research at Ghent University (BOFSTA2017002501) and a Grant from the King Baudouin Foundation (KBS 2018-J1130650-209563). Jonas Van Der Donckt is funded by a doctoral fellowship of the Research Foundation—Flanders (FWO). Part of this work is done in the scope of the imec.AAA Context-aware health monitoring.imec.AAA Contextaware-aware health monitoring.

3.7.3. Supplemental Materials

All data and corresponding codes are openly available through <https://osf.io/78g9s/>. Code works out of the box with instructions found in the corresponding README.md in OSF directory.

Exclusion criteria:

- Other than native Dutch speakers
- Left-handed
- Born before 1970
- Psychology student
- Personal or family history of epilepsy
- Recent neurosurgical procedures
- Pacemaker or other electronic implants

- Inner ear prosthesis
- Metal objects or magnetic objects in the brain or around the head (only removable earrings & piercings are allowed)
- Pregnancy
- Unstable medical condition
- A current depressive episode
- Other psychiatric disorders
- Skin disorder at the level of the head
- Current addiction
- Current substance abuse
- Current use of psychotropic medication
- Eye disease(s)
- Heart, respiratory, or neurological problems
- Did not drink coffee or smoke 2 hours before the start of the experiment
- Dreadlocks

Response Block:

First: Read out loud text “Marloes”:

“Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij, dus dat duurde wel even. Nu wachten ze tot de trein eraan komt. Het is al vijf over drie, dus het duurt nog vier minuten. Er staan nog veel meer mensen te wachten. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.”

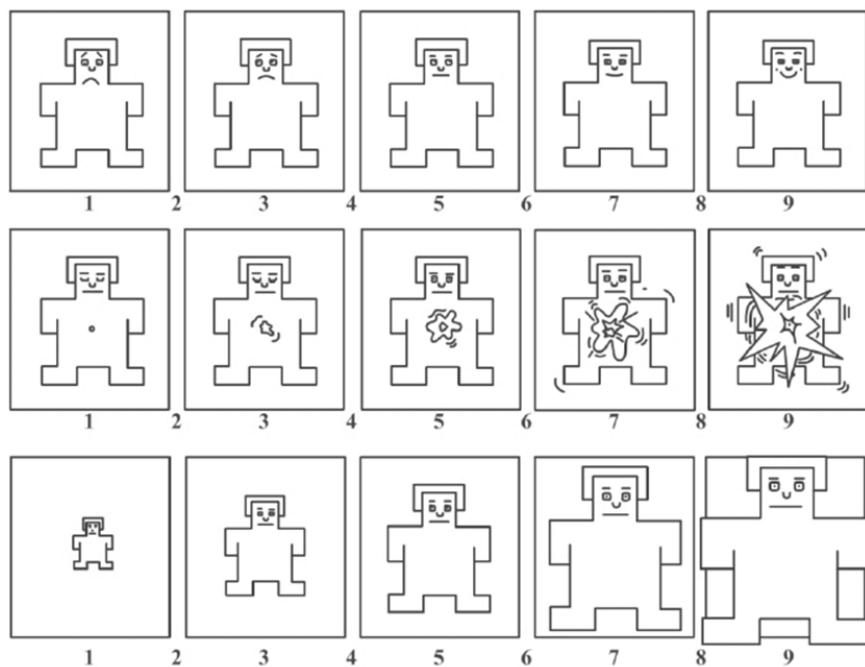
From: van de Weijer and Slis (1991)

Translation:

“Papa and Marloes are at the station. They are waiting for the train. First, they bought a ticket. There was a very long queue, so it took a while. Now they wait for the train to arrive. It's already

five past three, so it's still four minutes. Many more people are waiting. Marloes looks to the left, she sees the train coming in the distance."

Second: Answer Self-Assessment Manikins (SAMs)

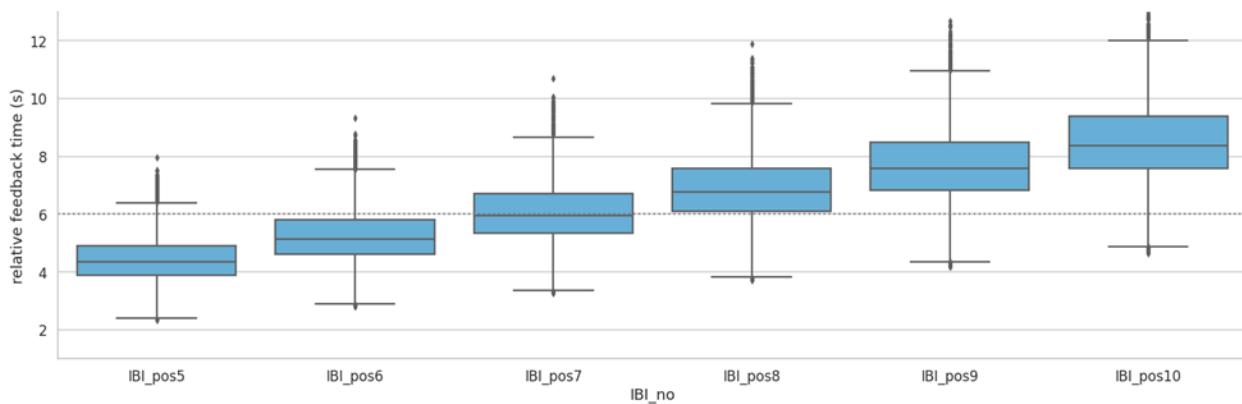


Self-Assessment Manikin (SAM) Scales (Bradley & Lang, 1994)

Third: Read out loud and answer Brief State Rumination Index (BSRI; Marchetti, Mor, Chiorri & Koster, 2018)

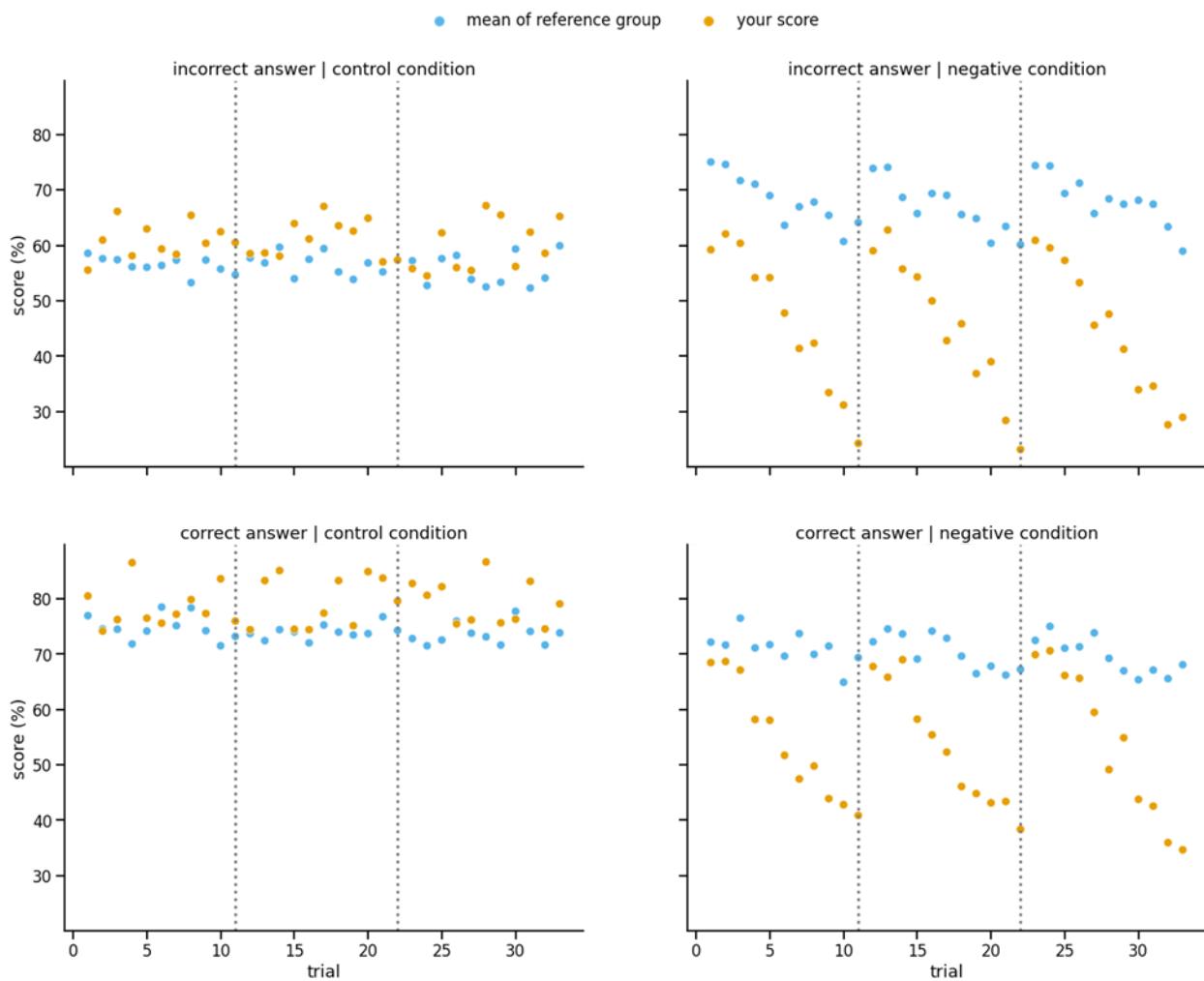
Supplemental Figure 1.

Boxplots indicating time passed since feedback onset for each relative delta IBI. Horizontal bars indicate the lower and upper limit, box borders indicate the 25th and 75th percentile, and lines in the boxplot indicate the median. A horizontal dotted line indicates the 6 seconds mark: total feedback exposure time. This shows that 75% of the trials had passed the 6 seconds of feedback exposure at IBI8.



Supplemental Figure 2.

Performance feedback as lines in the colored feedback bar. Blue dots stand for ‘group performance’, and orange dots stand for ‘your performance’. Y-axis shows time over the condition (left two plots are control condition; right two plots are negative condition). Vertical dashed lines show separations for the subblocks (where a response block was executed).



R Packages and session info:

R version 4.1.1 (2021-08-10)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 10 x64 (build 19044)

Matrix products: default

locale:

```
[1] LC_COLLATE=English_Belgium.1252          LC_CTYPE=English_Belgium.1252
LC_MONETARY=English_Belgium.1252 LC_NUMERIC=C
[5] LC_TIME=English_Belgium.1252
```

attached base packages:

```
[1] stats   graphics grDevices utils   datasets methods  base
```

other attached packages:

```
[1] dplyr_1.0.10      tibble_3.1.8       arrow_5.0.0.2    car_3.0-11      ggpubr_0.4.0
[6] ggplot2_3.4.0     effectsize_0.8.2   effects_4.2-0    effects_4.2-0  carData_3.0-4
lmerTest_3.1-3
[11] reshape_0.8.8     pander_0.6.4      emmeans_1.6.3    lme4_1.1-29
yarrr_0.1.5
[16] circlize_0.4.13  BayesFactor_0.9.12-4.4 Matrix_1.3-4     coda_0.19-4
jpeg_0.1-9
```

loaded via a namespace (and not attached):

```
[1] nlme_3.1-152        pbkrtest_0.5.1      bit64_4.0.5      insight_0.18.8
numDeriv_2016.8-1.1   tools_4.1.1        bit64_4.0.5      insight_0.18.8
[7] backports_1.2.1     utf8_1.2.2         R6_2.5.1        DBI_1.1.1       colorspace_2.0-3
nnet_7.3-16
[13] withr_2.5.0        tidyselect_1.2.0    bit_4.0.4       curl_4.3.2      compiler_4.1.1
textshaping_0.3.6
[19] cli_3.4.1          sandwich_3.0-1     labeling_0.4.2    bayestestR_0.13.0 scales_1.2.1
mvtnorm_1.1-2
[25] pbapply_1.4-3      systemfonts_1.0.4   stringr_1.4.0    digest_0.6.27   foreign_0.8-83
minqa_1.2.4
[31] rio_0.5.27         pkgconfig_2.0.3     readxl_1.3.1    rlang_1.0.6     GlobalOptions_0.1.2
rstudioapi_0.13
```

```

[37] farver_2.1.1      shape_1.4.6       generics_0.1.3    zoo_1.8-9      zip_2.2.0
magrittr_2.0.3
[43] parameters_0.20.1 Rcpp_1.0.7        munsell_0.5.0    fansi_1.0.3    abind_1.4-5
lifecycle_1.0.3
[49] stringi_1.7.4     multcomp_1.4-18   MASS_7.3-60     plyr_1.8.6     grid_4.1.1
parallel_4.1.1
[55] crayon_1.4.1    forcats_0.5.1     lattice_0.20-44  haven_2.4.3    splines_4.1.1
hms_1.1.0
[61] pillar_1.8.1      boot_1.3-28      estimability_1.3  ggsignif_0.6.3  codetools_0.2-18
glue_1.6.2
[67] mitools_2.4       data.table_1.14.0  vctrs_0.5.0     nloptr_1.2.2.2  cellranger_1.1.0
MatrixModels_0.5-0
[73] gtable_0.3.1      purrr_0.3.4      tidyverse_1.1.3   assertthat_0.2.1  datawizard_0.6.5
openxlsx_4.2.4
[79] xtable_1.8-4      broom_1.0.5      survey_4.1-1     rstatix_0.7.0    ragg_1.2.2
survival_3.2-11
[85] TH.data_1.1-0     ellipsis_0.3.2

```

Corresponding feature names as described in GeMAPS (Eyben et al., 2016)

```

Jitter = 'jitterLocal_sma3nz_amean',
Shimmer = 'shimmerLocaldB_sma3nz_amean',
F0 = 'F0semitoneFrom27.5Hz_sma3nz_amean',
HNR = 'HNRdBACF_sma3nz_amean',
Voiced Segments per Second = 'VoicedSegmentsPerSec',
Voiced Segment Length = 'MeanVoicedSegmentLengthSec'

```

Ecologically Valid Speech Collection in Behavioral Research: The Ghent Semi-spontaneous Speech Paradigm (GSSP)

Jonas Van Der Donckt^{*12}, Mitchel Kappen^{*34}, Vic Degraeve¹², Kris Demuynck¹², Marie-Anne Vanderhasselt³⁴⁵, Sofie Van Hoecke¹²

**Contributed equally*

¹ IDLab, Ghent University - imec, Ghent, Belgium

² Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

³ Department of Head and Skin, Ghent University, University Hospital Ghent (UZ Ghent), Department of Psychiatry and Medical Psychology, Ghent, Belgium

⁴ Ghent Experimental Psychiatry (GHEP) Lab, Ghent University, Ghent, Belgium

⁵ Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

Under Review at Behavior Research Methods. Preprint citation:

Van Der Donckt, J., Kappen, M., Degraeve, V., Demuynck, K., Vanderhasselt, M., & Van Hoecke, S. (2023, March 29). Ecologically Valid Speech Collection in Behavioral Research: The Ghent Semi-spontaneous Speech Paradigm (GSSP). <https://doi.org/10.31234/osf.io/e2qzw>

4.1. Abstract

This paper introduces the Ghent Semi-spontaneous Speech Paradigm (GSSP), a new method for acquiring unscripted speech data for affective-behavioral research in both experimental and real-world settings through the description of peer-rated pictures with a consistent affective load. The GSSP was designed to meet five criteria; (1) allowing flexible speech acquisition durations, (2) providing a straightforward and non-interfering task, (3) allow for experimental control, (4) favoring semi-spontaneous speech for its prosodic richness, and (5) require minimal human interference to enable scalability. The validity of the GSSP was evaluated through an online task, in which this paradigm was implemented alongside a fixed-text read-aloud task. The results indicate that participants were able to describe images with an adequate duration, and acoustic analysis demonstrated a trend for most features in line with the targeted speech styles (i.e, unscripted spontaneous speech versus scripted read-aloud speech). A speech style classification model using acoustic features achieved a balanced accuracy of 83% on within-dataset validation, indicating separability between the GSSP and read-aloud speech task. Furthermore, when validating this model on an external dataset that contains interview and read-aloud speech, a balanced accuracy score of 70% is obtained, indicating an acoustic correspondence between the GSSP speech and spontaneous interviewee speech. The GSSP is of special interest for behavioral and speech researchers looking to capture spontaneous speech, both in longitudinal ambulatory behavioral studies and laboratory studies. To facilitate future research on speech styles, acoustics, and affective states, the task implementation code, the collected dataset, and analysis notebooks are available.

4.2. Introduction

Over the last decades, the human voice and speech have been increasingly studied in relation to, amongst others, psychiatric disorders (e.g., depression, schizophrenia), and current psychological (e.g., stress) or physiological (e.g., sleepiness) states (Fagherazzi et al., 2021; Van Puyvelde et al., 2018, Martin et al., 2022). To date, the primary form of speech data used in affective-behavioral research in an experimental setting remains scripted read-aloud speech gathered in highly controlled laboratory environments (Van Puyvelde et al., 2018; Wagner et al., 2015). Scripted lab speech more conveniently allows for systematic experimental control, thus limiting the implicit inclusion of unwanted latent variables. As a result, a smaller sample size is sufficient to capture all degrees of freedom compared to unscripted speech gathered in less controlled environments (Xu, 2010a). However, acoustic properties found in one speech style can be style-specific, which limits the explanatory power of the speech data to other settings (e.g., real-world). Therefore, a promising research direction is to investigate the influence of speech acquisition paradigms on both production and perception (Wagner et al., 2015). On top of this, the scalability of speech acquisition methods should be considered, given that the long-term objective of affective sensing experiments is to facilitate wide-spread, real-world affect monitoring (Slavich et al., 2019). To this end, it is necessary to investigate speech acquisition approaches that can be used in real-life scenarios but still allow for sufficient experimental control.

Prior work has indicated that vocal responses to affective loads may be as individual and unique as the voice itself, requiring more isolated studies that control for inter-individual differences (Giddens et al., 2013; Van Puyvelde et al., 2018). In order to address this issue,

within-subject designs have been proposed, which allow for the collection of both baseline and affective data (Kappen, Hoorelbeke, et al., 2022; Kappen, Van Der Donckt, et al., 2022). However, in these works, the acoustic analysis was conducted on read-aloud speech with a fixed text, which limits the generalizability of conclusions to the more naturalistic and spontaneous speech encountered in real-life settings. Baird and colleagues (2019) tackled this within-participant challenge by using cortisol concentration as a target to examine acoustic features associated with stress. Their spontaneous speech samples were acquired using the Trier Social Stress Tests (TSST; Kirschbaum et al., 1993). In more recent work, Baird and colleagues (2021) assessed the generalizability of spontaneous speech correlates for stress via cortisol, heart rate, and respiration, by using three TSST corpora. The results show an increasing trend towards generalization and explanation power. However, these results are still limited, as the TSST only produces stressed speech under psychosocial load (i.e., during the interview), without consensus on the acquisition of baseline speech.

Furthermore, It has been demonstrated that affective states can influence decisions, working memory, and information retrieval (Mikels & Reuter-Lorenz, 2019; Weerda et al., 2010). Therefore, unscripted speech, which requires larger planning units such as sentences, clauses, and temporal structure, can lead to changes in wording, grammar, and timing of speech under these affective states (Fromkin, 1973; Paulmann et al., 2016; Slavich et al., 2019). These prosodic markers are less pronounced in scripted speech, as fewer planning units are needed (Barik, 1977; Xu, 2010a). However, spontaneous speech rarely allows for controlling the factors that contribute to the phenomena of interest (Xu, 2010a). To address this, more controlled variants of unscripted speech paradigms are employed, such as guided interviews and picture description tasks. For example, language disturbances, at both the acoustic-prosodic and

content level, have been shown to be promising markers for psychiatric diseases such as schizophrenia-spectrum disorders (de Boer et al., 2020). As a result, schizophrenia researchers have employed guided interview protocols as a means of acquiring unscripted speech (Voppel et al., 2021). Recent work in this area has proposed more continuous disorder follow-up, for which such labor-intensive interviews may not be an ideal match (de Boer et al., 2021). The use of picture description tasks, i.e., providing an image stimuli to a participant with the objective of describing the image content out loud, has a successful track record in the field of neurology for aiding in the diagnosis of cognitive disorders such as aphasia and Alzheimer's disease (Goodglass et al., 2001; Mueller et al., 2018). Picture description paradigms are here preferred over spontaneous speech, as the controlled and monological types of content are easier to obtain and analyze in clinical practice (Lind et al., 2009). Furthermore, by letting participants describe stimuli with consistent emotional loads, repeated measures are possible with little change in affect (Helton & Russell, 2011; Kern et al., 2005).

Given the above observations, we established a requirement list for a speech acquisition task that would be useful for both experimental research and real-world applicability. The task should (1) allow for flexible speech acquisition durations, ensuring that it can easily be incorporated into existing paradigms. For example, enabling the inclusion of a task at multiple (time-constrained) moments within an experiment allows for within-participant analysis. Additionally, the task should be (2) straightforward and non-interfering, ensuring that the resulting speech is not affected by the cognitive-emotional load of the acquisition method itself, but only by prior effects induced by the experimental paradigm. The method should be (3) controllable, as experimental control reduces the large number of samples that would be needed elsewhere to marginalize out latent factors (Xu, 2010b). Furthermore, the method

should (4) stimulate participants towards spontaneous speech, as the richness in prosody, semantics, and content has already been proven to be useful to derive markers in affective and cognitive research (Christodoulides, 2016). Unscripted speech should also be more generalizable to everyday speech, enabling the translation of results to real-world settings and applications. Finally, the speech acquisition method should be (5) scalable, by requiring minimal human interference during acquisition to allow for usability in both longitudinal ambulatory studies with repeated measures and studies at scale.

This paper aims to make a significant step towards the application of lab results in a real-world setting by introducing the *Ghent Semi-spontaneous Speech Paradigm (GSSP)*; a controllable and ecologically valid picture description paradigm that complies with the above requirements. By having participants describe an image depicting a neutral social setting that is not complex, and they have not seen before, there will be no cognitive interference of active recall. Whereas speech analysis for (psychosocial) stress and other psychological states is increasingly gaining traction, we propose these stimuli to be congruent with psychosocial (stress) paradigms. That is, offering stimuli that would minimally interfere with elicited psychophysiological states of the experimental paradigm in order to (1) not risk the disruption of observed effects in other constructs (e.g., physiological reactions, rumination, etc.) due to mind wandering and (2) have the collected speech closely resemble the active mental state experienced by participants due to the experimental paradigm.

The selected images are empirically sampled from the PiSCES (Teh et al., 2018) and Radboud (Langner et al., 2010) datasets, based on peer-rated neutral content. In order to minimize additional cognitive task load and biases, we used proper habituation instructions and images with a consistent neutral emotional load. To the best of our knowledge, this is the

first work proposing a picture description task for applied/real-world acoustic analysis of affective-behavioral states.

To summarize, the contributions of this paper are threefold;

- We propose the Ghent Semi-spontaneous Speech Paradigm (GSSP), a novel speech acquisition paradigm using a picture description task for affective-behavioral research. The GSSP enables near-effortless, semi-controlled acquisition of unscripted speech data in both experimental and longitudinal real-life settings.
- To assess the validity of the GSSP regarding speech style, utterance duration, and image subset consistency, a study was performed using a web application. The analysis of the web application data indicated that participants are able to describe the images with an adequate duration, and acoustic analysis hinted that acoustic properties of the GSSP correspond to those of spontaneous speech.
- In order to facilitate the reproducibility of the research outcomes, the materials utilized in the study have been made openly accessible under a research-friendly license. The analysis scripts and web-app code are available on GitHub¹, while the dataset and instruction videos can be accessed through Kaggle datasets².

4.3. Methods

In order to evaluate three key factors pertaining to the GSSP, namely (1) the participant's ability to engage in prolonged discourse, (2) the acoustical similarity between the gathered GSSP speech and spontaneous speech, and (3) the consistency of the initially

¹ https://github.com/predict-idlab/gssp_analysis,

https://github.com/predict-idlab/gssp_web_app

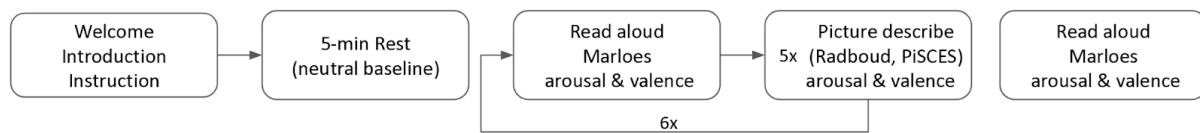
² <https://www.kaggle.com/datasets/jonvdrdo/gssp-web-app-data>

selected image subset, a web application was developed which incorporates the GSSP among a standardized read-aloud task. The following sections describe the web app design and the GSSP procedure, followed by a specification of the participant selection procedure and the speech data processing.

4.3.1. Web app and procedure

Figure 1

Flowchart of the web application experiment.



Note. This results in 7 Marloes, 15 Radboud, and 15 Pisces Utterances per participant.

The web application was developed in Python using the Flask framework (Grinberg, 2018). Screenshots and implementation details are found in Supplemental Material S1 and on GitHub³. As depicted in Figure 1, the experiment was divided into five blocks, with the first block consisting of three consecutive web pages. The first page, labeled “Welcome” (S1.1 Figure 1), provided a general overview of the study’s purpose, i.e., validating the usability of an image set for experimental speech research. The second page, labeled “Introduction” (S1.2 Figure 2), was used to acquire demographics (i.e., age, sex, recording material, highest obtained degree) together with the approval of the informed consent. The introduction page also provided an overview of the general guidelines for the task. In particular, it emphasized the importance of performing the task on a computer in a quiet and distraction-free environment. The complete list of (translated) guidelines can be found in S1.2. The third page, labeled “Task

³ https://github.com/predict-idlab/gssp_web_app

Instruction” (S1.3 Figure 3-5), provided detailed instructions for the components of this study, i.e., a 5-minute resting block (S1.4) to establish a neutral baseline state, followed by the speech acquisition tasks through scripted read speech (i.e., “Marloes”) and the GSSP. The task instruction page also provided three videos, one of which demonstrated the procedure for the reading task, while the other two illustrated the GSSP its picture description process, using an image from both the PiSCES and Radboud dataset, which were not utilized as stimuli in the study. In addition, the instruction page presented the read-out-loud (“Marloes”) text and participants were instructed to read the text out loud. This reading exercise, together with the demonstration videos, aimed to reduce novelty effects for both the GSSP and reading task (Davidson & Smith, 1991; Weierich et al., 2010; Zuckerman, 1990). The study requested participants to provide a description of each image for a minimum of 30 seconds, but no explicit instruction was given to adhere to this duration, nor was the length of the speech recording indicated to the subjects. Finally, as a speech quality control procedure, participants had to record and playback a speech sample, and were only permitted to proceed to the resting block after this microphone assessment was conducted.

The resting block consisted of a blank page featuring the text: “*Close your eyes and try to focus on your breathing. You will hear a sound when the resting block is over*” (translated). This step aimed to bring the participants to a neutral baseline state and is in alignment with (Kappen, Hoorelbeke, et al., 2022; Kappen, Van Der Donckt, et al., 2022).

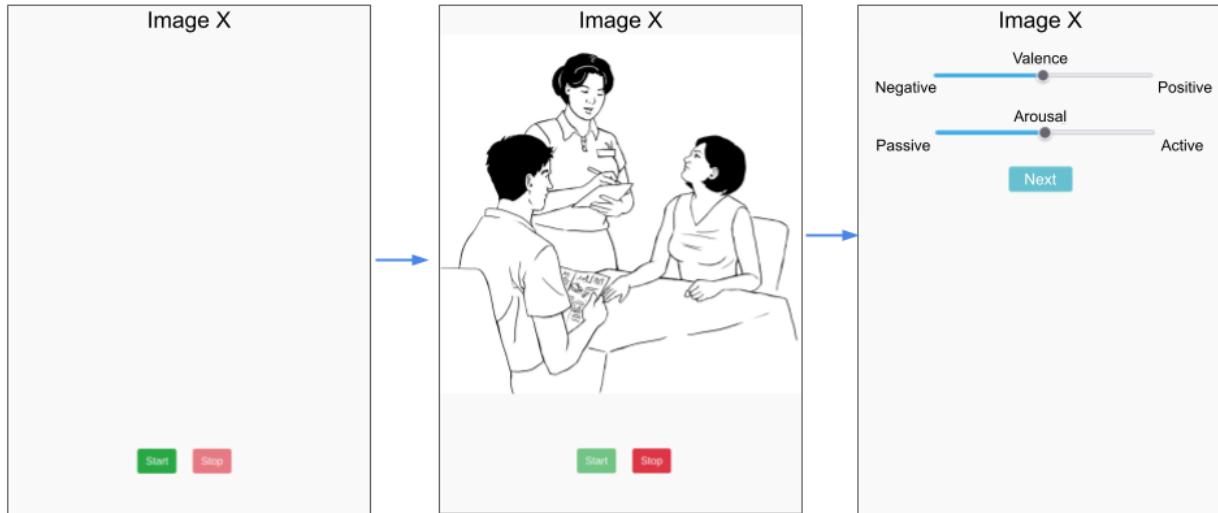
Participants completed six speech cycles, each of which began with one read-out-loud text, followed by GSSP trials, as depicted by Figure 1.

4.3.1.1. Read-out-loud text “Marloes”

To acquire scripted speech fragments, participants were instructed to read aloud a standardized text of five sentences. The text, commonly known as the “Marloes” text, is widely used in Dutch speech therapy due to its phonetic balance (Van de Weijer & Slis, 1991; full text provided in S1.5). As depicted in the speech acquisition flow of Figure 2, the “Marloes” text only became visible (in frame B) after the participant initiated the task by clicking the start button, which should limit the variability in preparation time. Once the segment has been read out loud, participants could proceed to a new page by clicking the stop button. On this page, two sliders were presented, which participants adjusted to indicate their level arousal and valence experienced during the speech task (Figure 2).

Figure 2

Trial flow chart of the web app speech acquisition task



Note. Trial flow of the web app speech acquisition task with the pages translated to English. First, an empty page (a) is displayed with an enabled start button and a disabled stop button. When the participant clicks the start button, (b) the audio recording begins, the stop button will be enabled. The

stimulus in the form of an image (or text for the read-aloud task) will be presented. After the participant completes the stimulus speech acquisition task, he/she or they click on the stop button, triggering the redirection to (c), where the participant will report their experienced arousal and valence values.

4.3.1.2. GSSP picture description speech

The unscripted speech fragments were collected in accordance with the read-aloud task. In order to limit the variability of image description preparation time, all stimuli were presented to the participants at the beginning of the recording upon clicking the start button. This approach ensured a degree of uniformity among participants. The order of the presented images was randomized, alternating between pictures from the PiSCES and Radboud databases. The first image shown was drawn from the PiSCES subset, followed by an image from the Radboud set, and so on. Each cycle consisted of a total of 5 pictures, resulting in a total of 15 images from both the PiSCES and Radboud databases (as shown in Figure 1). To ensure optimal audio quality, speech data was stored within the participant's browser session using the *Recorder.js* JavaScript tool (Matt, 2016). After utterance completion, the audio data was converted into a 16 bit PCM mono WAV file and sent to a secure server, along with the experienced arousal and valence score.

The Radboud Faces Database provides a set of stimuli including both adult and children's faces that have been parametrically varied with respect to displayed expressions, gaze direction, and head orientation (Langner et al., 2010). These stimuli were evaluated based on the facial expression, valence, and attractiveness. To conduct the study, the GSSP utilizes a subset of the neutral expression, front-facing adult images (7 males, 8 females), which were

selected based on their mean valence scores, to minimize the potential for inducing emotional responses in respondents. The used image subset is depicted in Supplemental S1.6. Figure 8.

Similarly, the PiSCES database is a collection of 203 black-and-white line drawings of individuals in social settings (Teh et al., 2018). These stimuli were evaluated based on emotional valence, intensity, and social engagement. To control for emotional responses, a subset of 15 images with neutral valence ratings and high social engagement scores were selected from this database for use in the study. The images are illustrated in Supplemental S1.6. Figure 7.

4.3.1.3. Drinking break

To mitigate vocal fatigue, participants were instructed to take a sip of water after every 9 utterances (Welham & Maclagan, 2003).

4.3.2. Participants

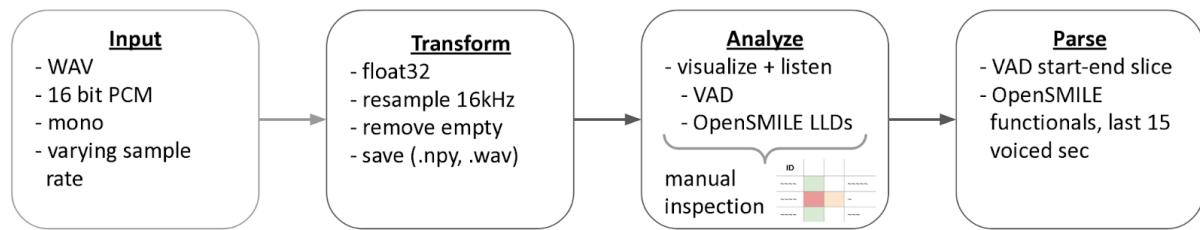
The data was collected in two waves. First, the research groups' networks were leveraged by distributing the study via social network sites. Second, the Prolific platform (Palan & Schitter, 2018) was utilized to gain an adequate number of participants. This resulted in a convenience sample of 89 participants (45 women, 43 men, 1 other) with an average age of 27.54 ($SD = 6.63$). The study only included Dutch-speaking participants residing in Belgium or the Netherlands whose native language is Dutch. On average, participants required one hour to complete the study.

4.3.3. Data processing

The audio data parsing and analysis were carried out in Python 3.8.13 and statistical analyses of the valence-arousal scores were performed using R4.1.1. For detailed version information of the utilized libraries, we refer to the GitHub repository⁴.

Figure 3

Audio data processing flowchart.



4.3.3.1. Audio data processing

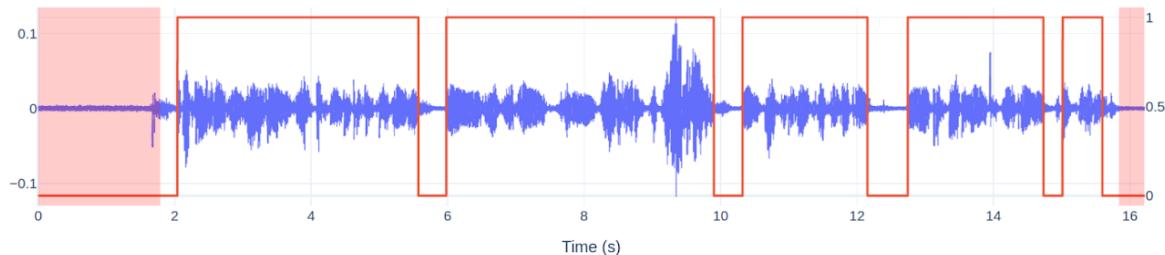
The audio data processing workflow is depicted in Figure 3. The first step is to acquire the input samples (Input step), which are then converted (Transform step) to 16kHz mono audio with 32-bit float precision. Due to technical issues, some recordings were not saved properly, resulting in empty audio-files (24 in total) that are excluded from further analysis during the Transform step. The non-empty transformed outputs are then saved for further processing in the Analyze and Parse steps. Following the Transformation step, a participant-level manual inspection is carried out to assess the audio data quality (Analyze step). The inspection process involves utilizing customized visualizations, as illustrated in Figure 4 and S2. Figure 9, to assist in the analysis process. The outcome of this analysis is a

⁴ https://github.com/predict-idlab/gssp_analysis

manual inspection sheet, which is used to exclude participants with inadequate audio quality. Lastly, a parsing step is performed on the transformed audio for participants whose audio quality was deemed sufficient. This parsing step employs a Voice Activity Detection (VAD) model (*Speechbrain/Vad-Crdnn-Libriparty · Hugging Face*, n.d.) from the SpeechBrain toolkit (Ravanelli et al., 2021) to detect speech segments. The outer bounds of the first and last speech segments are padded with a margin of 0.25 seconds before slicing. The red shaded regions in Figure 4 illustrates the regions that are omitted. As such, each VAD-sliced segment consists of speech data that starts and ends at the same relative time. This approach allows us to make fair comparisons between fixed duration excerpts (relative from VAD-slice beginning or end). Supplemental S2. further details the visualizations that are utilized during the analyze step.

Figure 4

VAD slicing with a 0.25s margin for the first and last voiced segment.



Note. The first voiced regions occur +/- 2 seconds after the participant pressed the “start” button. The slicing ensures that each participant's first/last voiced segment start/end at the same time, allowing to make fair comparisons on fixed-duration excerpts relative to the VAD-slice beginning or end.

4.3.3.2. Acoustic Speech parameter extraction

The final stage of the parsing step entails the extraction of speech parameters. To control for the effects of file duration on acoustic parameters and repetitive start sentences in

the picture description tasks (e.g., “*I see a black and white cartoon*” for the PiSCES database), only the last 15 voiced seconds, as determined by the VAD-slice, were used for both parameter extraction techniques listed below. Therefore, only excerpts with a VAD-slice duration of at least 15 seconds were included, resulting in 2901 samples from 82 participants (554 Marloes, 1184 PiSCES, 1163 Radboud).

The extraction of speech parameters was conducted using the OpenSMILE 3.0.1 Python API (Eyben et al., 2010) and the GeMAPSv01b functional configuration (Eyben et al., 2016). The selection of the GeMAPSv01b configuration was in line with previous research (Baird et al., 2019, 2021; Jati et al., 2018; Kappen, Hoorelbeke, et al., 2022; Kappen, Van Der Donckt, et al., 2022). Moreover, Triantafyllopoulos and colleagues (2019) observed that the eGeMAPS, which is a superset of the GeMAPS, is relatively robust in noisy conditions. A comprehensive explanation of the utilized OpenSMILE feature subset can be found in Supplements S3. During the manual inspection phase of the Analyze step (as illustrated in Figure 2), differences in the values of OpenSMILE Low-Level Descriptors (LLDs) were observed when the original 44.1kHz data was resampled to 16kHz. Further examination of OpenSMILE’s sampling-rate inconsistencies is available in Supplemental S4. This examination led to superposing a small (Gaussian-sampled) noise of -30dB to the resampled audio, which empirically improved the voiced boundary detection.

In addition to the acoustic parameters investigation, visual speech style analysis was performed via deep learning embeddings, generated using the ECAPA-TDNN architecture (Desplanques et al., 2020). These embeddings were projected into a two-dimensional space using t-SNE (Van der Maaten & Hinton, 2008). Further implementation details regarding the

GeMAPSv01b and ECAPA extraction procedures can be found in the [feature extraction](#) and [ECAPA-TDNN](#) notebooks, respectively⁵.

Finally, to evaluate the binary separability of speech styles in a data-driven manner, the OpenSMILE features and ECAPA-TDNN embeddings were also fed to a machine-learning model. Specifically, logistic regression, a linear classification model, was used to assess this separability. The Scikit-learn Python toolkit by Pedregosa and colleagues (2011) was used for this purpose.

4.3.3.3. External dataset “Corpus Gesproken Nederlands”

To validate the generalizability of the data-driven speech style assessment, an external dataset was utilized. Specifically, a subset of the *Corpus Gesproken Nederlands* (CGN), i.e., the Corpus of Spoken Dutch, was leveraged (Oostdijk, 2000). CGN includes recordings of both Flemish and Netherlands Dutch, which are categorized into various components based on speech style and context settings. These components range from spontaneous conversations and news broadcasts, to sports commentaries, sermons, and read-aloud texts. The corpus data is stored as 16-bit PCM 16kHz WAV files, and each recording is orthographically transcribed and diarized.

Two components were chosen from the CGN dataset to serve as our unscripted and scripted speech styles. Component A, “face-to-face conversations”, was deemed unsuitable for the unscripted speech style due to the presence of frequent interruptions and crosstalk in the recordings. Component B, “interviews with Dutch teachers”, was used as unscripted speech style data because the data has a low emotional load and the interviewee’s utterances

⁵ We conducted an acoustic analysis on the duration of the entire utterance and found that the results were consistent with those obtained from the last 15 seconds of voiced data for both the [ECAPA-TDNN](#) projections and [openSMILE](#) features.

have few interruptions and often meet the 15-second duration criterion. Finally, Component O, “read-aloud texts”, served as scripted speech style data in our validation. In accordance with the acoustic parameter extraction performed on the web app data, excerpts of the last 15 seconds (with a margin of 2 seconds) were taken from single speaker segments that met the duration requirement and the OpenSMILE GeMAPSv01b configuration was applied. This resulted in a validation dataset of 3357 segments (1643 scripted read speech (comp. O) and 1714 spontaneous speech (comp. A)).

4.4. Results

This section presents the results of the web app data analysis. In the first subsection, we focus on the affective consistency of the GSSP stimuli and present the arousal and valence scores. Next, the speech style of GSSP is analyzed using renowned acoustic features in relation to the existing literature on speech styles. The jitter and shimmer features trended differently from prior research, prompting a subsection containing a detailed exploration of this inconsistency. The GSSP speech style is further evaluated using data-driven methods, including an ECAPA-TDDN t-SNE projection for analysis and generalizability of the GSSP towards unscripted speech styles beyond the web app dataset.

4.4.1. Arousal & valence scores

As described in the methods section, the PiSCES and Radboud database stimuli were selected by choosing the closest to the middle of the valence scale in its respective validation studies, whilst accounting for potential thematic difficulties that could elicit certain emotional

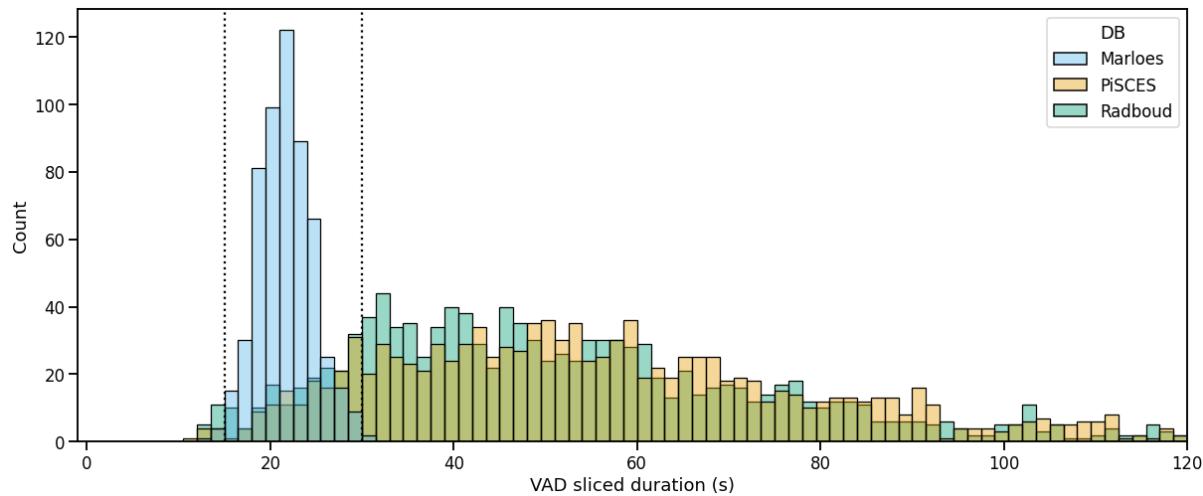
responses in subgroups of people. In doing so, we have compiled a picture subset that could be considered emotionally neutral and therefore appropriate for affective research. Additionally, we have conducted a series of statistical and descriptive approaches to also validate the appropriateness of our picture subset. These tests can be found on the analysis repository⁶, as they are not key findings in this manuscript, yet are of importance to assess the rigidity and validity of the results presented here.

4.4.2. Speech feature analysis

4.4.2.1. Speech duration

Figure 5

Distribution plot of the VAD-sliced utterance durations. The vertical dashed line on the left indicates the voiced duration threshold (15 seconds) and the right line represents the instructed image description duration (30 seconds)



⁶ https://github.com/predict-idlab/gssp_analysis/scripts/1.2_FactorAnalysis.pdf

The web app guidelines, outlined in Supplemental S1.2., instructed the participants to discuss each image for at least 30 seconds. As illustrated by the histogram of Figure 5, 88% of the GSSP utterances met this duration requirement.

4.4.2.2. OpenSMILE acoustics

The OpenSMILE GeMAPSv01b acoustic features were partitioned into three subsets, i.e., a temporal, frequency, and amplitude-related subset. Each subset consists of four distinct features, whose detailed descriptions can be found in Supplemental S3. The visualization of these subsets was conducted using two approaches. The first approach displays the features using a box plot that groups the data on speech acquisition task (Marloes (M), PiSCES (P), Radboud (R)) and speech style (Read, picture description (GSSP)), with each utterance contributing a single data point to the corresponding task (see Figure 6-8). This visualization enables the interpretation of the acoustic features in parameter value space. The second approach employs a violin delta-plot, in which utterances of the same participant and speech task are median-aggregated and then subtracted from other speech task aggregations for the same participant, see Figure 13 of Supplemental S5. This results in each participant contributing one data point for each delta. This violin delta plot reveals the distribution shifts and spreads over the various acquisition tasks. More detailed information regarding the violin delta plot can be found in Supplemental S5.

4.4.2.2.1. Temporal features

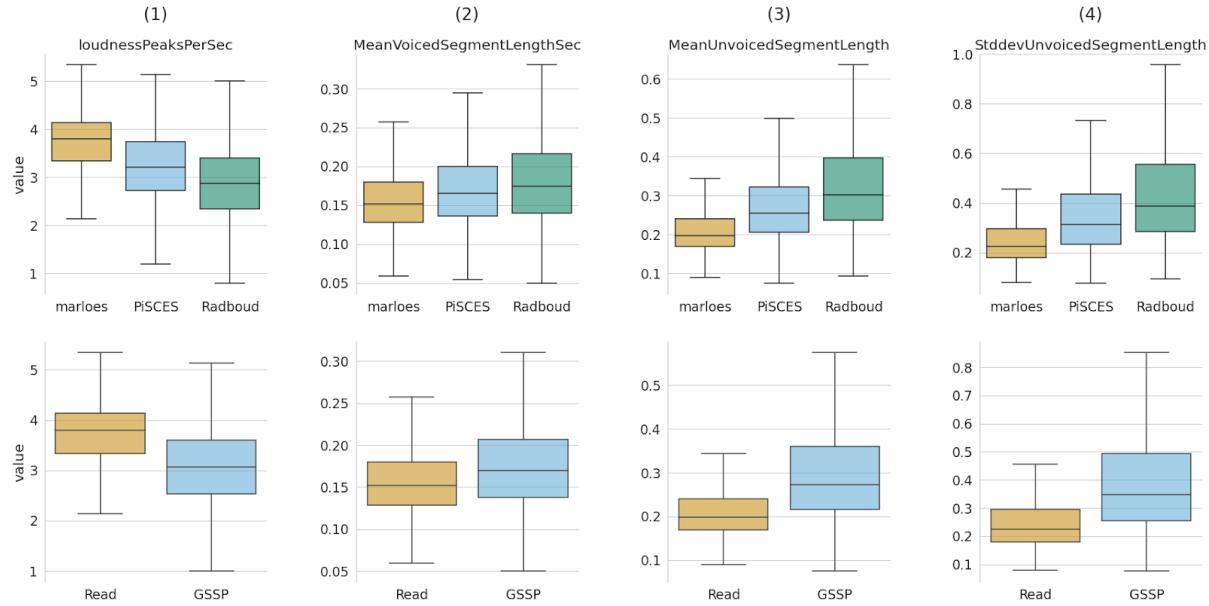
The four temporal features are `loudnessPeaksPerSec`, `MeanVoicedSegmentLengthSec`, `MeanUnvoicedSegmentLength`, and `StddevUnvoicedSegmentLength`, shown in Figure 6.

Column (1) of Figure 6 represents the number of loudness peaks, serving as a proxy for syllable rate (Eyben et al., 2016). The coherent distribution shift of the upper and lower subplot of column (1) indicates that the “Marloes” task has a higher articulation rate than both picture description tasks. This observation is consistent with (Barik, 1977; Levin et al., 1982), which attributes this lower articulation rate to the need for planning time when speaking unprepared. Column (2) illustrates the MeanVoicedSegmentLengthSec, which is the distribution of the mean sound duration, indicating slightly shorter voiced segments for the “Marloes” task than for the picture description tasks. This is in line with the notion of voiced segment duration being inversely proportional to the speaking rate (column (1)). Furthermore, (de Silva et al., 2003) observed a tendency towards longer sound durations for spontaneous speech, which is consistent with our findings. Blaauw (1992) and Laan (1992) found that pauses tend to be more irregular and longer for spontaneous speech, as reflected in the MeanUnvoicedSegmentLength (3) and StddevUnvoicedSegmentLength (4) subplots. Based on these observations, we can conclude that the temporal characteristics of the proposed semi-scripted speech paradigm are highly similar to those of unscripted speech.⁷

⁷ We also observe that the speech rate is lower and the pauses are longer for the Radboud task compared to PiSCES, which might be caused by the homogeneity of the Radboud images, making it substantially harder to describe novel things.

Figure 6

Box plot of temporal features, grouped by acquisition task (row 1) and speech style (row 2).



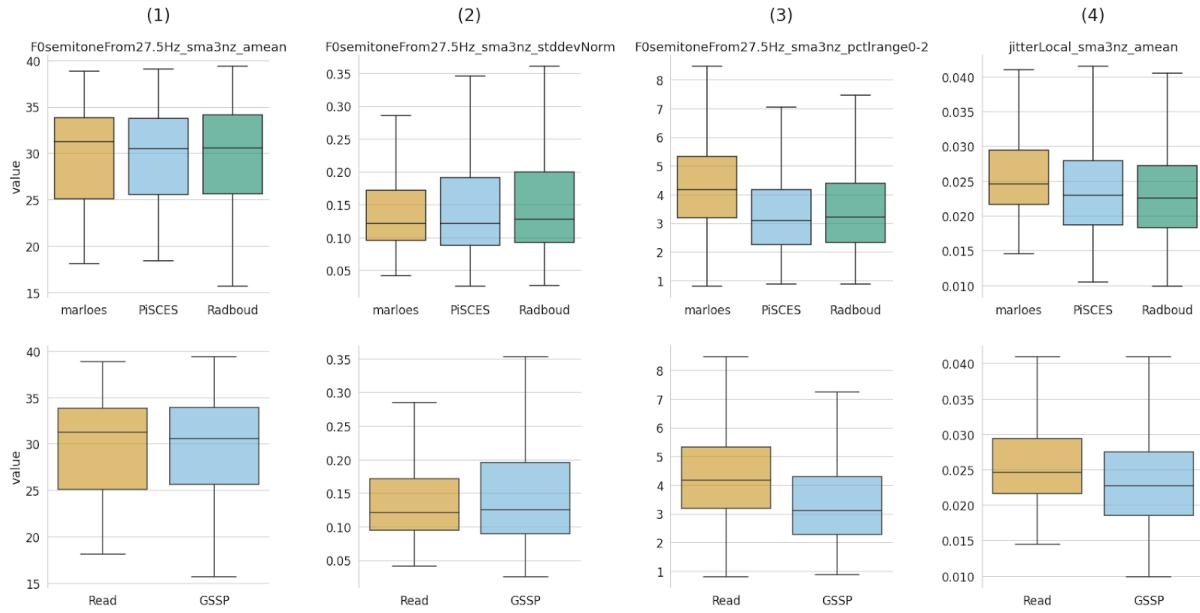
4.4.2.2.2. Frequency-related features

Four frequency-related features were utilized, i.e., F0semitoneFrom27.5Hz_sma3nz_amean, F0semitoneFrom27.5Hz_sma3nz_stddevNorm, F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2, and jitterLocal_sma3nz_amean; the mean frequency perturbation. Columns (1) and (2) of Figure 7 capture the distribution of the fundamental frequency (F0), i.e., its mean and standard deviation respectively. In accordance with de Silva and colleagues (2003), no clear differences are observed between these acoustic parameters and speech styles. Column (3) visualizes the F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2, which covers the F0-range (i.e., 20th to 80th percentile) and has been reported to be larger in read speech (Batliner et al., 1995), consistent

with our findings. (Kraayeveld, 1997; Laan, 1997) observed more jitter in spontaneous speech, but our findings indicate a significant decrease in jitter (4) for spontaneous speech.

Figure 7

Box plot of frequency-related features, grouped by task (row 1) and speech style (row 2).



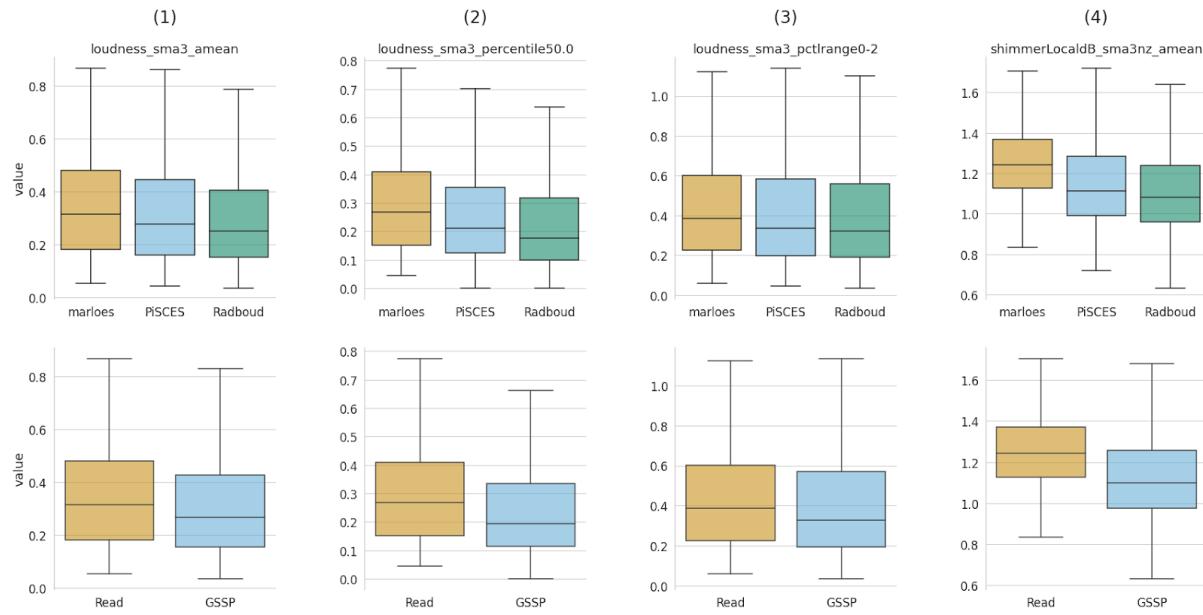
4.4.2.2.3. Amplitude-related features

Also here, four features have been utilized i.e., (1) loudness_sma3_amean; the average loudness, (2) loudness_sma_3_percentile50.0; the median loudness, (3) loudness_sma3_pctlrange0-2; the 20th-to-80th percentile loudness range, and (4) shimmerLocaldB_sma3nZ_amean; the mean amplitude perturbation. To date, few results are available regarding loudness parameters and speech style. (Laan, 1992, 1997, p. 1) even applied amplitude normalization to eliminate loudness differences in their experiments. Columns (1) and (2) of Figure 8 show a slight increase in loudness for the reading task. The loudness range, represented by column (3), is slightly larger for the read-aloud task. We

observe a decrease in shimmer (4) for the picture description task, contradicting the findings of (Kraayeveld, 1997; Laan, 1997).

Figure 8

Box plot of amplitude-related features, grouped by task (row 1) and speech style (row 2).



4.4.2.2.4. Jitter and shimmer inconsistencies

The preceding sections, along with the effect size charts of Supplemental S8, indicated a significant decrease in both jitter and shimmer for the unscripted GSSP task compared to the scripted read-aloud speech. This is in contrast to prior literature that reports the opposite effect, where unscripted speech produces higher jitter and shimmer values than scripted speech. Therefore, we have included this additional section to explore the potential reasons for this inconsistency. Three potential causes for this potential discrepancy are presented below. The first plausible explanation for the acoustic differences could be (1) the nuances in speech styles. The current experiment involved participants being alone in a room and talking to a

computer (recording device); while the previous work that produced contrasting results utilized interview-based spontaneous speech (Kraayeveld, 1997; Laan, 1997). Therefore, a promising research direction is to investigate the acoustic distinctions between these nuanced speech styles (e.g. monologue vs. conversation, the effect of a study taker on monologue unscripted speech, and the effect of the presence of an interviewer in the room). A second potential explanation could be that (2) the OpenSMILE toolkit may not be capable of accurately extracting jitter and shimmer parameters in settings with higher levels of environmental noise. Specifically, sound produced by environmental elements emanating periodic noises such as a (computer) fan could be picked up at the voiced boundaries, i.e., the regions where voicing ends and the environmental elements become more prominent. OpenSMILE could then start to attribute voiced features to these environmental elements. As detailed in Supplemental S4, abnormally high F0 values were encountered near those voiced boundaries, which largely disappeared when resampling the raw data and adding a small amount of dithering (noise). This supplemental also presents the elevated values observed for the shimmer parameter. Given that read-speech contains a greater proportion of voiced segments, as indicated by the higher syllable rate in Figure 6(1), there is an increased frequency of voiced boundaries per time unit. This increase in voiced boundaries potentially contributes to the increase in (abnormally high) augmentation in shimmer and jitter values. A third explanation could be that (3) there is indeed a decreasing trend in shimmer and jitter values when analyzing less scripted speech. As outlined in Supplemental S7, a visualization of the weight coefficients of a logistic regression model revealed that a substantial negative coefficient was identified for the shimmer parameter when the model was fitted on either the web app or CGN dataset. Overall, we can conclude

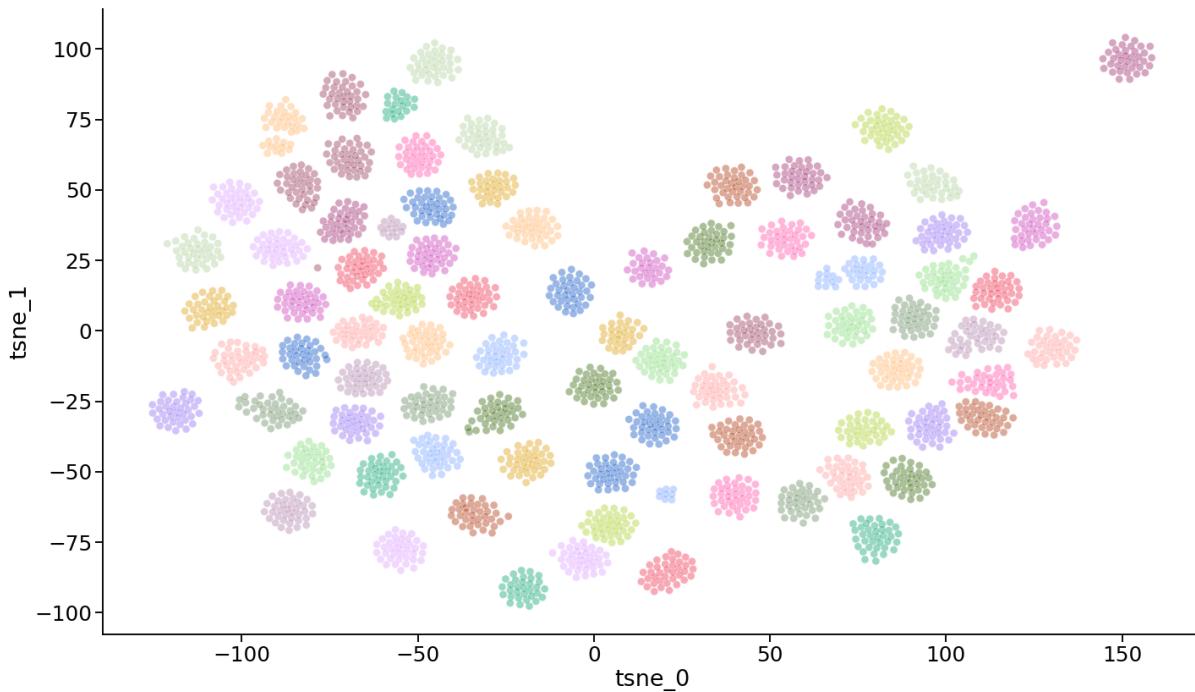
that the trend for the majority of acoustic parameters are in accordance with the findings from the literature.

4.4.3. ECAPA-TDNN projections

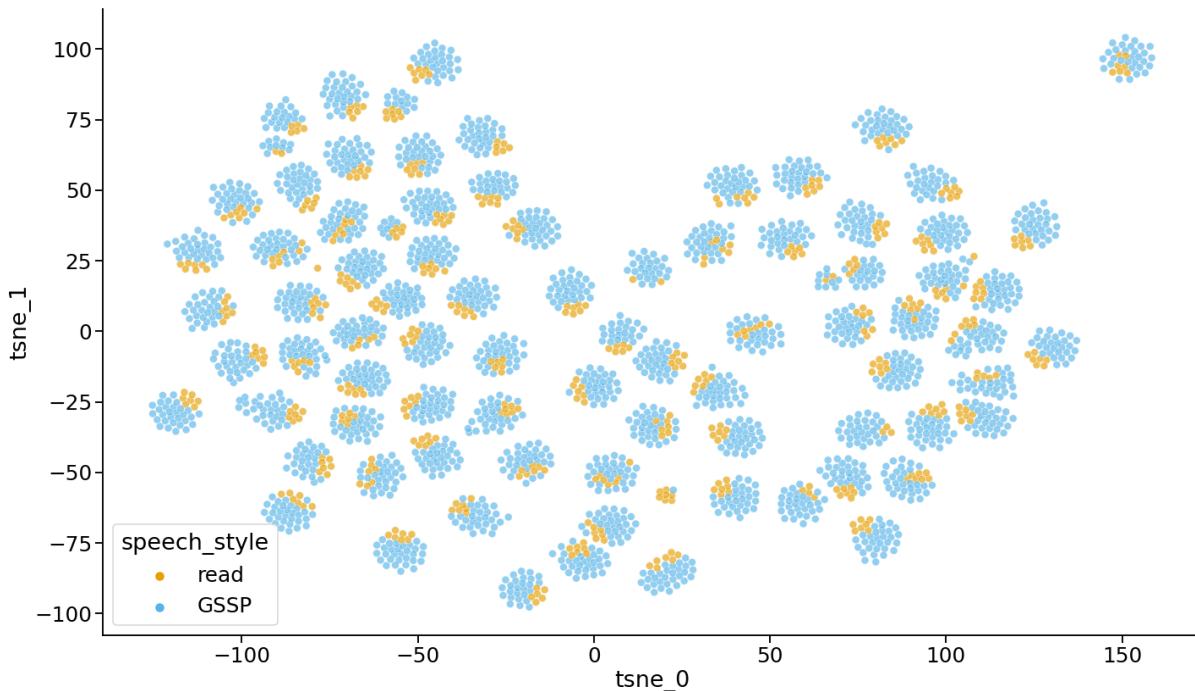
In addition to examining the relationship between acoustic-prosodic features in speech styles and positioning this within the literature, we also wanted to investigate speech styles using more data-oriented techniques. To this end, the ECAPA-TDNN architecture (Desplanques et al., 2020) was used to extract fixed-duration embeddings from the utterances. These embeddings were projected into a lower-dimensional space using t-distributed stochastic neighbor embedding (t-SNE, Van der Maaten & Hinton, 2008), the results of which are depicted in Figure 9. The upper visualization (a) serves as a validation check, as this demonstrates the primary objective of the ECAPA-TDNN architecture, which is speaker identification. Each cluster consists of a single hue-color, indicating that all cluster points originate from the same user, demonstrating the successful separation of speakers. The second visualization (b) employs the same projection parameters as (a) but uses speech style as the hue. We observe that in the majority of individual speaker clusters, the “read” speech style utterances are grouped together. This is noteworthy as the primary goal of ECAPA-TDNN is speaker identification, which implies that it has little advantage in utilizing the silent parts of the utterances and primarily focuses on acoustic properties. This observation leads to the hypothesis that the speech style information resides within the captured acoustic properties of the ECAPA-TDNN architecture.

Figure 9

Two-dimensional t-SNE projection of ECAPA-TDNN utterance embeddings.



(a) Hue determined by speaker ID.



(b) Hue determined by speech style.

Note. In the t-SNE visualization presented, the x and y axes do not represent specific measurable or interpretable variables. Instead, they are abstract dimensions created to optimally place data points such that similar data points are close to each other in this two-dimensional space, and dissimilar ones are farther apart. In this visualization, each marker signifies one speech utterance. As shown in (a), clusters of markers correspond to utterances by a single speaker. By coloring each dot based on its speech (trial) style in (b), we observe a general tendency for individual speech styles to cluster within each speaker's utterances. This suggests a potential separability of speech styles based on speaker identification techniques using acoustic properties. It's important to note that the absolute positions of clusters or points on the x and y axes are not directly interpretable, but rather their relative positions to one another carry significance.

To further validate this claim, a logistic regression model with speech style separability as the objective was fitted on the embeddings. Supplemental Figure 14 illustrates the normality of the embedding features. As such, no further embedding transformations were needed and the features were standardized by removing the mean and scaling to unit variance. The model achieved a balanced accuracy score of 84% \pm 1.5% when using 5-fold cross-validation with the speaker ID as a grouping variable. Model details can be found in the associated notebook⁸.

4.4.4. CGN validation

Speech style separability was also assessed using the GeMAPSv01b features. Supplemental Figure 15 illustrates the distribution of the OpenSMILE features, which demonstrates a non-normal distribution for most features. As a result, a power transformation was applied as a preprocessing step to ensure more Gaussian-like distributions (Yeo & Johnson, 2000). The GeMAPS model achieved a balanced accuracy score of 83% \pm 2.5%, which is comparable to the results obtained from the ECAPA-TDNN model in the above

⁸ [gssp_analysis/notebooks/0.6_ECPA_TDNN.npy.ipynb](#)

section. A 5-fold cross-validation with the speaker ID as the grouping variable was used as the validation setup.

Finally, to ensure maximum generalizability towards the CGN dataset, an educated subset of 24 GeMAPSv01b features was crafted based on their known contribution to speech style representativity. The model achieved a cross-fold score of 81% +/- 2%, using the within web app dataset validation setup as described in the previous paragraph. Subsequently, this model was fitted on the whole web application dataset and validated on the external CGN dataset. This resulted in a balanced accuracy score of 70%, as outlined by Table 1. In addition, a confusion matrix can be found in figure 10, which displays the predicted versus true labels to see how the accuracy was built up. Due to the distribution shift between the training and validation sets, a decrease in accuracy compared to the within-web-app cross-fold accuracy was expected. The obtained performance indicates that the GeMAPSv01b web app data speech style decision boundary also holds predictive power when validated on the “B” and “O” components of the CGN dataset, thus indicating an acoustic correspondence between the picture description GSSP speech (web app) and the interviewee speech (CGN). Additional information regarding the model and feature subset selection can be found in the associated notebook⁹.

⁹ [gssp_analysis/notebooks/1.3_OpenSMILE_ML.ipynb](#)

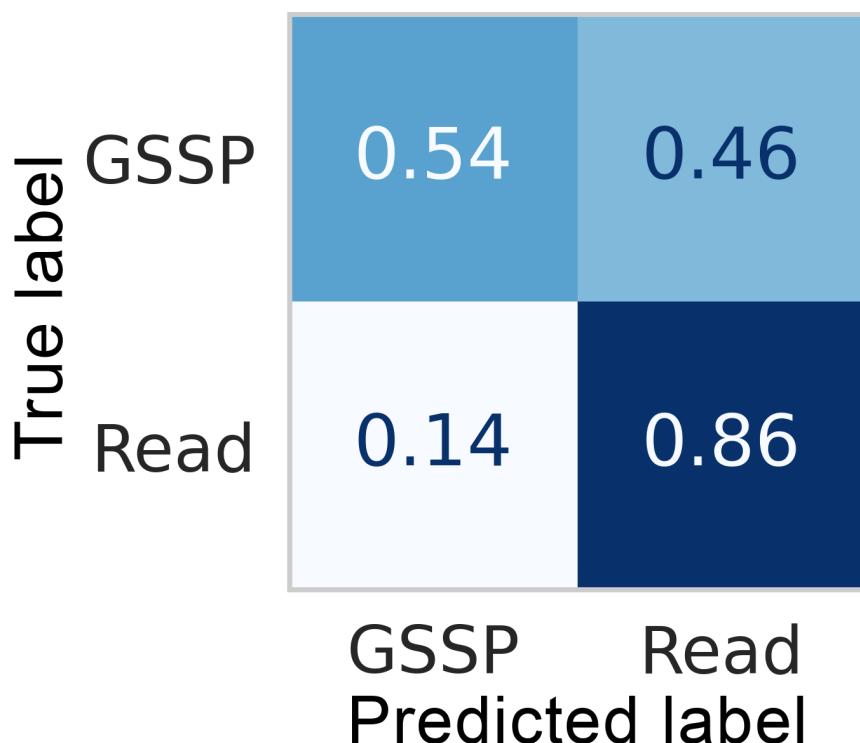
Table 1

CGN validation classification report.

	Precision	Recall	F1-score	Support
Read	0.64	0.87	0.74	1643
Unscripted	0.81	0.54	0.65	1714
accuracy			0.70	3357
macro_avg	0.73	0.70	0.69	3357
weighted_avg	0.73	0.70	0.69	3357

Figure 10

Confusion matrix of predicted labels versus true labels.



4.5. Discussion

This paper presents the Ghent Semi-spontaneous Speech Paradigm (GSSP), a picture description task designed to capture speech data for affective-behavioral research in both experimental and real-world settings. The GSSP was developed based on the requirements identified in the field and literature, which were translated into a list of criteria to which the paradigm should adhere to. Specifically the GSSP was designed to (1) allow for flexible speech acquisition duration, facilitating convenient incorporation into existing paradigms, (2) present a simple and congruent task, ensuring that the obtained speech is not affected by the load of the speech acquisition method itself, (3) be controllable to limit the inclusion of unwanted latent factors, (4) favor unscripted speech for its prosodic richness and generalizability to everyday speech, and (5) require minimal human effort during data acquisition to enable use in remote and real-world settings. The GSSP utilizes image stimuli that are emotionally consistent within their respective image set. This enables stimuli randomization in longitudinal designs, which also mitigates learning effects due to familiarity with the stimuli (as occurs with fixed repeated stimuli). Moreover, both image sets are emotionally neutral, limiting confounding effects when implementing the GSSP in known experimental design. Lastly, we specifically designed one image set (PiSCES) to contain stimuli portraying social settings to supply researchers with emotionally neutral, yet congruent stimuli to be used in experimental designs using psychosocial stressors (commonly used, reliable, and potent stressors), further limiting confounding effects on stress reactions.

The validation of the GSSP was conducted using a web application that collected speech data from participants. In particular, the participants were instructed to repeatedly perform two tasks; a read-aloud text task and the GSSP. A duration analysis indicated that participants were able to describe images with sufficient duration, therefore adhering to the first criteria.

To provide a correct analysis of the study, it is important to ensure that only valid speech samples are utilized. Therefore, an essential contribution of this study is the open-source pipeline utilized to process and evaluate speech data, which has been instrumental in ensuring data quality and determining selection criteria. This methodology is not specific to this research and can be applied in other speech data studies, particularly due to its open-source nature.

To analyze the collected data with regard to speech styles, three analyses were performed. The first analysis was concerned with relating acoustic features and existing literature on scripted vs. unscripted speech styles. Acoustic speech features, extracted using the OpenSMILE GeMAPSv01b functional configuration, exhibited a trend that is consistent with the literature on the targeted speech styles, i.e., scripted read-aloud speech and unscripted spontaneous speech, therefore adhering to the fourth requirement. Nonetheless, the observed trend was not consistent across all the analyzed features. Specifically, the jitter (our fourth frequency-related feature) and the shimmer values (our fourth amplitude-related feature) did not align with existing literature in this field. Jitter and shimmer both were lower for the unscripted GSSP task compared to the scripted read-aloud speech, which contradicts literature that reports the opposite effect. This discrepancy can potentially be attributed to (a combination of) three reasons, which are thoroughly discussed in OpenSMILE acoustics section.

The second analysis is concerned with data-driven techniques. Specifically, the ECAPA-TDNN t-SNE projection, presented in Figure 9, demonstrated that speaker clusters are further sub-grouped according to speech style. A speech style separability experiment on the web app data, utilizing the GeMAPSv01b features, yielded a balanced accuracy of 83%, which is in agreement with the findings of Levin and colleagues (1982), who reported that listeners were able to distinguish between spontaneous and read-aloud speech with an accuracy of 84%, primarily based on temporal characteristics and false starts.

The third analysis assessed the generalization of the web app speech style separability by performing an out-of-dataset validation on the CGN dataset, using scripted read-aloud speech (comp. O) and spontaneous interviewee speech (comp. B). This validation resulted in a lower, but still satisfactory, balanced accuracy score of 70%. These results indicate that there is a clear separation between speech from the read-aloud and GSSP task, and that the acoustic properties of the GSSP task are in accordance with those of spontaneous speech from well-regarded databases.

The significant variation in (the quality of) utilized recording devices, introduced some degree of compromise to the validity of the analysis. Future studies that employ this paradigm are advised to implement stricter guidelines to limit the inclusion of unwanted variables (third criterion). Despite this limitation, the web application demonstrated the ability to deploy the GSSP at scale (fifth criterion) by needing no human interference during acquisition. Furthermore, the unscripted nature (fourth criterion) of this paradigm presents an opportunity to explore semantic-content aspects, as previous research has established the potential of these modalities as markers for various disorders (de Boer et al., 2020; Mueller et al., 2018).

In conclusion, the GSSP demonstrates qualities of intuitiveness, scalability, accessibility, and brevity (i.e, 30-60 seconds), making it a suitable addition to well-established experimental studies for collecting unscripted speech during key moments, such as before and after exposure to stressors or emotional loads. This approach does not compromise other essential outcome variables and can be seamlessly integrated into remote-sensing applications, facilitating research on longitudinal mental well-being using speech and mood correlates (Kappen et al., 2023). We hypothesize that findings obtained from utilizing the GSSP will be easier translatable to real-world settings, such as speech collected in team or board meetings, presentations, or any other social setting. This research aligns with the conclusion from Xu (2010), which states that employed speech acquisition techniques need constant updates to gain increasingly better insights into the full complexity of speech. We are convinced that our presented GSSP, supported by the documented code, data¹⁰, and analysis results, enable behavioral researchers to incorporate an unscripted picture description task in their research studies. Future work should focus on further assessing the nuances in speech styles and investigating environmental effects on (this) paradigm(s), such as the presence of a study taker.

¹⁰ The provided web app dataset can also be used to analyze acoustic effects of repetitive reading, as participants read the same (phonetically balanced) text 9 times.

4.6. References

- Baird, A., Amiriparian, S., Cummins, N., Sturmbauer, S., Janson, J., Messner, E.-M., Baumeister, H., Rohleder, N., & Schuller, B. W. (2019). Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test. *Interspeech* 2019, 534–538. <https://doi.org/10.21437/Interspeech.2019-1352>
- Baird, A., Triantafyllopoulos, A., Zankert, S., Ottl, S., Christ, L., Stappen, L., Konzok, J., Sturmbauer, S., Meßner, E.-M., Kudielka, B. M., Rohleder, N., Baumeister, H., & Schuller, B. W. (2021). An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress. *Frontiers in Computer Science*, 3, 750284. <https://doi.org/10.3389/fcomp.2021.750284>
- Barik, H. C. (1977). Cross-Linguistic Study of Temporal Characteristics of Different Types of Speech Materials. *Language and Speech*, 20(2), 116–126. <https://doi.org/10.1177/002383097702000203>
- Batliner, A., Kompe, R., Kießling, A., Nöth, E., & Niemann, H. (1995). Can You Tell Apart Spontaneous and Read Speech if You Just Look at Prosody? In A. J. R. Ayuso & J. M. L. Soler (Eds.), *Speech Recognition and Coding* (pp. 321–324). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-57745-1_47
- Blaauw, Eleneora. (1992). *Phonetic differences between read and spontaneous speech*. https://www.isca-speech.org/archive_v0/archive_papers/icSLP_1992/i92_0751.pdf
- Christodoulides, G. (2016). *Effects of cognitive load on speech production and perception* [PhD Thesis]. UCL-Université Catholique de Louvain.
- Davidson, R. A., & Smith, B. D. (1991). Caffeine and novelty: Effects on electrodermal activity

and performance. *Physiology & Behavior*, 49(6), 1169–1175.

[https://doi.org/10.1016/0031-9384\(91\)90346-P](https://doi.org/10.1016/0031-9384(91)90346-P)

de Boer, J. N., Brederoo, S. G., Voppel, A. E., & Sommer, I. E. C. (2020). Anomalies in language as a biomarker for schizophrenia: *Current Opinion in Psychiatry*, 33(3), 212–218.

<https://doi.org/10.1097/YCO.0000000000000595>

de Boer, J. N., Voppel, A. E., Brederoo, S. G., Schnack, H. G., Truong, K. P., Wijnen, F. N. K., & Sommer, I. E. C. (2021). Acoustic speech markers for schizophrenia-spectrum disorders: A diagnostic and symptom-recognition tool. *Psychological Medicine*, 1–11.

<https://doi.org/10.1017/S0033291721002804>

de Silva, V., Iivonen, A., Bondarko, L. V., & Pols, L. C. W. (2003). *Common and Language Dependent Phonetic Differences Between Read and Spontaneous Speech in Russian, Finnish and Dutch*. 4.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Interspeech 2020*, 3830–3834. <https://doi.org/10.21437/Interspeech.2020-2650>

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

<https://doi.org/10.1109/TAFFC.2015.2457417>

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the International Conference on Multimedia - MM '10*, 1459. <https://doi.org/10.1145/1873951.1874246>

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers*, 5(1), 78–88.
<https://doi.org/10.1159/000515346>

Fromkin, V. (1973). *Speech errors as linguistic evidence*. Mouton The Hague.

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.e21-390.e29.
<https://doi.org/10.1016/j.jvoice.2012.12.010>

Goodglass, H., Kaplan, E., & Weintraub, S. (2001). *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA.

Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc.

Helton, W. S., & Russell, P. N. (2011). The Effects of Arousing Negative and Neutral Picture Stimuli on Target Detection in a Vigilance Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(2), 132–141.
<https://doi.org/10.1177/0018720811401385>

Jati, A., Williams, P. G., Baucom, B., & Georgiou, P. (2018). Towards Predicting Physiology from Speech During Stressful Conversations: Heart Rate and Respiratory Sinus Arrhythmia. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4944–4948. <https://doi.org/10.1109/ICASSP.2018.8461500>

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54(2), 910–921. <https://doi.org/10.3758/s13428-021-01670-x>

Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van

Hoecke, S., & Vanderhasselt, M. A. (2022). *Acoustic speech features in social comparison: How stress impacts the way you sound* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/kms98>

Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. (2023). Speech as a Promising Biosignal in Precision Psychiatry. *Neuroscience & Biobehavioral Reviews*, 105121.

Kern, R. P., Libkuman, T. M., Otani, H., & Holmes, K. (2005). Emotional Stimuli, Divided Attention, and Memory. *Emotion*, 5(4), 408–417. <https://doi.org/10.1037/1528-3542.5.4.408>

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The ‘Trier Social Stress Test’—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1–2), 76–81.

Kraayeveld, J. (1997). *Idiosyncrasy in prosody: Speaker and speaker group identification in Dutch using melodic and temporal information*. Katholieke Universiteit.

Laan, G. P. M. (1992). *PERCEPTUAL DIFFERENCES BETWEEN SPONTANEOUS AND READ ALOUD SPEECH*. 16.

Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22(1), 43–65. [https://doi.org/10.1016/S0167-6393\(97\)00012-5](https://doi.org/10.1016/S0167-6393(97)00012-5)

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, 24(8), 1377–1388. <https://doi.org/10.1080/02699930903485076>

Levin, H., Schaffer, C. A., & Snow, C. (1982). The Prosodic and Paralinguistic Features of Reading and Telling Stories. *Language and Speech*, 25(1), 43–54.

<https://doi.org/10.1177/002383098202500104>

Lind, M., Kristoffersen, K. E., Moen, I., & Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics*, 23(12), 872–886. <https://doi.org/10.3109/02699200903040051>

Martin, V. P., Rouas, J. L., Boyer, F., & Philip, P. (2021, August). Automatic Speech Recognition systems errors for objective sleepiness detection through voice. In Interspeech 2021 (pp. 2476-2480). ISCA.

Matt, D. (2016). Recorderjs. In GitHub repository. GitHub.
<https://github.com/mattdiamond/Recorderjs#readme>

Mikels, J. A., & Reuter-Lorenz, P. A. (2019). Affective Working Memory: An Integrative Psychological Construct. *Perspectives on Psychological Science*, 14(4), 543–559.
<https://doi.org/10.1177/1745691619837597>

Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9), 917–939. <https://doi.org/10.1080/13803395.2018.1446513>

Oostdijk, N. (2000). *Het corpus gesproken Nederlands*.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
<https://doi.org/10.1016/j.jbef.2017.12.004>

Paulmann, S., Furnes, D., Bøkenes, A. M., & Cozzolino, P. J. (2016). How Psychological Stress Affects Emotional Prosody. *PLOS ONE*, 11(11), e0165022.
<https://doi.org/10.1371/journal.pone.0165022>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., ... Bengio, Y. (2021). *SpeechBrain: A General-Purpose Speech Toolkit* (arXiv:2106.04624). arXiv. <http://arxiv.org/abs/2106.04624>

Slavich, G. M., Taylor, S., & Picard, R. W. (2019). Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22(4), 408–413. <https://doi.org/10.1080/10253890.2019.1584180>

Speechbrain/vad-crdnn-libriparty · *Hugging Face*. (n.d.). Retrieved December 1, 2022, from <https://huggingface.co/speechbrain/vad-crdnn-libriparty>

Teh, E. J., Yap, M. J., & Liow, S. J. R. (2018). PiSCES: Pictures with social context and emotional scenes with norms for emotional valence, intensity, and social engagement. *Behavior Research Methods*, 50(5), 1793–1805. <https://doi.org/10.3758/s13428-017-0947-x>

Triantafyllopoulos, A., Keren, G., Wagner, J., Steiner, I., & Schuller, B. W. (2019). Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement. *Interspeech* 2019, 1691–1695. <https://doi.org/10.21437/Interspeech.2019-1811>

Van de Weijer, J., & Slis, I. (1991). Nasaliteitsmeting met de nasometer. *Logopedie En Foniatrie*,

63(97), 101.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>

Voppel, A., de Boer, J., Brederoo, S., Schnack, H., & Sommer, I. (2021). Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Research*, 304, 114130. <https://doi.org/10.1016/j.psychres.2021.114130>

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12. <https://doi.org/10.1016/j.wocn.2014.11.001>

Weerda, R., Muehlhan, M., Wolf, O. T., & Thiel, C. M. (2010). Effects of acute psychosocial stress on working memory related brain activity in men. *Human Brain Mapping*, 31(9), 1418–1429. <https://doi.org/10.1002/hbm.20945>

Weierich, M. R., Wright, C. I., Negreira, A., Dickerson, B. C., & Barrett, L. F. (2010). Novelty as a dimension in the affective brain. *NeuroImage*, 49(3), 2871–2878. <https://doi.org/10.1016/j.neuroimage.2009.09.047>

Welham, N. V., & MacLagan, M. A. (2003). Vocal Fatigue: Current Knowledge and Future Directions. *Journal of Voice*, 17(1), 21–30. [https://doi.org/10.1016/S0892-1997\(03\)00033-X](https://doi.org/10.1016/S0892-1997(03)00033-X)

Xu, Y. (2010a). In defense of lab speech. *Journal of Phonetics*, 38(3), 329–336. <https://doi.org/10.1016/j.wocn.2010.04.003>

Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve

Normality or Symmetry. *Biometrika*, 87(4), 954–959.

Zuckerman, M. (1990). The Psychophysiology of Sensation Seeking. *Journal of Personality*, 58(1), 313–345. <https://doi.org/10.1111/j.1467-6494.1990.tb00918.x>

4.7. Supplemental Materials

4.7.1. Acknowledgments

The authors would like to thank Jeroen Van Der Donckt and Jennifer Sartor for proofreading the manuscript.

4.7.2. Funding

This research was supported by a grant for research at Ghent University (BOFSTA2017002501) and a grant from the King Baudouin Foundation (KBS 2018-J1130650-209563). Jonas Van Der Donckt is funded by a doctoral fellowship of the Research Foundation – Flanders (FWO 1S56322N). Part of this work is done in the scope of the imec.AAA Context-aware health monitoring. MAV received funding from the FWO and from Ghent University (Grants: G0F4619N and BOF17/STA/030, respectively).

4.7.3. Supplemental Materials

Supplemental Materials are also openly made available online.

All data and code are publicly available at

<https://www.kaggle.com/datasets/jonvdrdo/gssp-web-app-data>

The code is available as well on GitHub at:

https://github.com/predict-idlab/gssp_analysis

https://github.com/predict-idlab/gssp_web_app

Contents:

S1: Web Application details

S2: Speech data parsing

Details regarding manual audio inspection and speech data processing methodology

S3: Description of utilized OpenSMILE feature subset

S4: OpenSMILE sampling rate inconsistency findings

S5: OpenSMILE delta visualizations

S6: ECAPA-TDNN & GeMAPS distribution plots to highlight (non)-normality

S7: Logistic regression weight coefficients

S8: Effect size Shimmer & Jitter

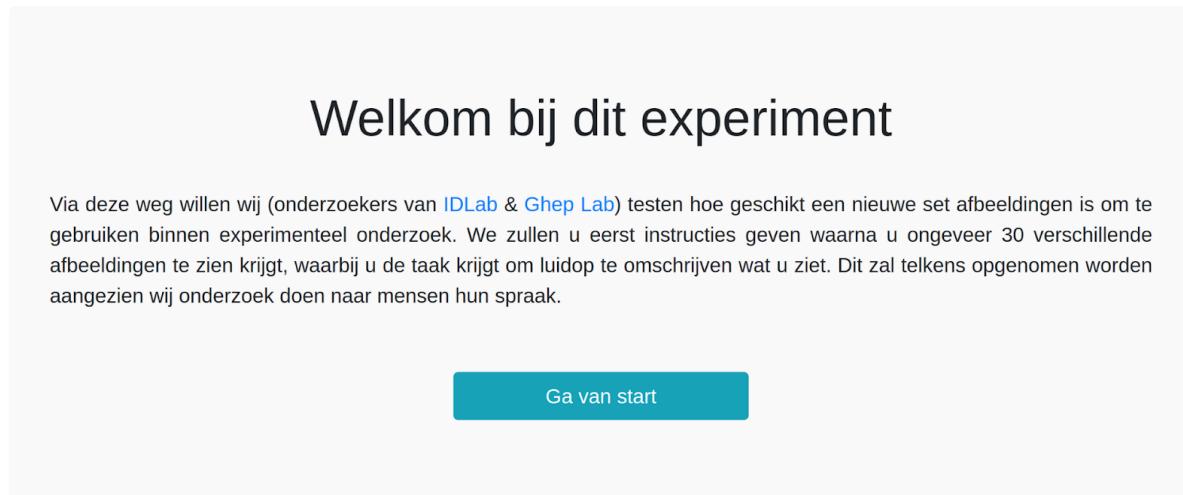
S9: Factor Analysis

4.7.3.S1. Web Application Details

S1.1. Welcome Page

Figure 1

Welcome page Screenshot.



S1.2. Introduction Page

Original instructions:

- Deze taak is enkel uit te voeren op een laptop/computer.
- Gelieve uw oortelefoons/koptelefoon of iets dergelijks te gebruiken met microfoon; dit zorgt voor hoge kwaliteit opnames.
- Indien u zeker bent dat de microfoon van uw computer van voldoende hoge kwaliteit is, mag u deze gebruiken.
- Zorg dat u plaatsneemt in een rustige omgeving waar u 30 minuten omringt wordt door zo min mogelijk afleiding en geluid.
- Zorg dat u een glas water bij uw laptop/computer hebt staan.
- Tijdens de taak zitten een aantal korte drinkpauzes zodat u geen droge keel krijgt door het veelvuldig praten – ook dit zorgt voor hoge kwaliteit opnames.

Translated instructions:

- This task is only to be performed on a desktop.
- We strongly suggest using a headphone, only use your desktops' microphone when you are sure that the recording quality of the device is high.
- Make sure that you are in a quiet and distraction free environment for at least 30 minutes.
- Make sure that you have a glass of water next to your desktop
- During the task, several drinking pauses will occur. This ensures that you will not suffer from a dry throat while speaking.

Figure 2

Introduction page screenshot.

Tot slot vragen we u ook nog om onderstaande gegevens in te vullen:

Geslacht

- Man
- Vrouw
- Anders

Leeftijd

Hoogst behaalde diploma

Device dat je gaat gebruiken voor de audio op te nemen

Informed consent:

Door akkoord te gaan verklaar ik hierbij dat ik als proefpersoon aan een online pilot-studie van de Universiteit Gent deelneem en het eens ben met de volgende punten:

1. Ik heb uitleg gekregen over de aard van de vragen, taken, opdrachten en stimuli die tijdens dit onderzoek zullen worden aangeboden, gekregen en dat mij de mogelijkheid werd geboden om bijkomende informatie te verkrijgen.
2. Ik begrijp dat deelname aan de studie vrijwillig is en dat ik mij op elk ogenblik uit de studie mag terugtrekken zonder een reden voor deze beslissing op te geven en zonder dat dit op enige wijze een invloed zal hebben op mijn verdere relatie met de onderzoekers.
3. Ik geef toestemming om mijn resultaten op vertrouwelijke wijze te bewaren, te verwerken en anoniem te rapporteren.
4. Ik geef toestemming om mijn gepseudonimiseerde dataset (niet meer terug te leiden naar de identiteit van de participant) online beschikbaar te stellen voor onderzoeksdoeleinden
5. Ik geef toestemming dat mijn spraak opnames online beschikbaar gesteld worden in een database voor onderzoeksdoeleinden. Deze opnames zullen niet openbaar gemaakt worden in combinatie met persoonlijke gegevens.
6. Ik begrijp dat auditors, vertegenwoordigers van de opdrachtgever, de Commissie voor Medische Ethiek of bevoegde overheden, mijn gegevens mogelijk willen inspecteren om de verzamelde informatie te controleren. Door dit document te ondertekenen geef ik toestemming voor deze controle. Bovendien ben ik op de hoogte dat bepaalde gegevens doorgegeven worden aan de opdrachtgever. Ik geef hiervoor mijn toestemming, zelfs indien dit betekent dat mijn gegevens doorgegeven worden aan een land buiten de Europese Unie. Te allen tijde zal mijn privacy gerespecteerd worden.
7. Ik begrijp dat persoonlijke gegevens worden verwerkt en bewaard gedurende minstens 20 jaar. Ik stem hiermee in en ben op de hoogte dat ik recht heb op toegang en op verbetering van deze gegevens. Aangezien deze gegevens verwerkt worden in het kader van medisch-wetenschappelijke doeleinden, begrijp ik dat de toegang tot mijn gegevens kan uitgesteld worden tot na beëindiging van het onderzoek. Uw gepseudonimiseerde gegevens kunnen tevens worden gebruikt voor verder onderzoek in het kader van emotieregulatie. Indien ik toegang wil tot mijn gegevens, zal ik mij richten tot de onderzoeker die verantwoordelijk is voor de verwerking.
8. Ik begrijp dat ik het recht heb op de hoogte te zijn van de resultaten van het huidige onderzoek.

Voor meer informatie omtrent het huidige onderzoek of bij vragen over het onderzoek of dit Informed Consent formulier kunt u contact opnemen met Mitchel.Kappen@UGent.be

Ik accepteer de informed consent

S1.3. Instruction Page

Figure 3-5

Instruction page screenshots.

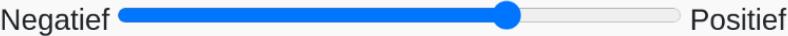
Taak instructie

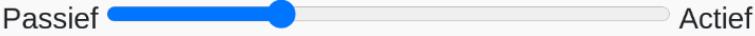
Afbeeldingen luidop omschrijven

Tijdens deze taak zal u zo'n 30 afbeeldingen bespreken. U ziet op de pagina een groene "Start-knop" en een rode "Stop-knop". De afbeelding zal op het scherm verschijnen eens u op start drukt. Dit zal er tevens ook voor zorgen dat de audio opname begint. Nadat u op deze groene knop heeft gedrukt, begint u te omschrijven wat u ziet. Maak u niet te veel druk als u vast loopt, probeer het natuurlijk te doen alsof u de afbeelding omschrijft aan iemand die de afbeelding niet kan zien. Ter indicatie kunt u erop richten om minimaal 30 seconden per afbeelding te omschrijven. Als u klaar bent met uw omschrijving drukt u op **Stop** en gaat u door naar het volgende scherm.

Gemoedstoestand ingeven via sliders

Nadat u de opname gestopt heeft, dient u zo snel mogelijk te antwoorden wat u bij de afbeelding voelt, geef dus aan wat het eerst in u opkomt. Bij elke afbeelding dient u op twee schalen te antwoorden: *opwinding/activiteit* en *valentie/aangenaamheid*, zoals hieronder weergegeven. Deze zullen nu eerst beter uitgelegd worden!

Valentie
Negatief  Positief

Opwinding
Passief  Actief

Volgende

Valentie / aangenaamheid
Valentie / aangenaamheid is een maatstaaf voor waarde. Een afbeelding is POSITIEF als het als goed wordt beschouwd, terwijl de afbeelding NEGATIEF is als het als slecht wordt beschouwd. Geef de valentie van elke afbeelding aan op een *sliding scale* van ZEER NEGATIEF tot ZEER POSITIEF, waarbij het middelpunt NEUTRAAL vertegenwoordigt.
Voorbeelden:

- Als u bijvoorbeeld het gevoel hebt dat "atoombom" een zeer negatieve betekenis heeft, dan antwoordt u ver links op de schaal.
- Als u bijvoorbeeld het gevoel hebt dat "fantastisch" een zeer positieve betekenis heeft, dan antwoordt u ver rechts op de schaal.
- Als u het gevoel heeft dat "spritjes" vrij onaangenaam is voor u, dan antwoordt u een beetje links op de schaal.
- Als u het gevoel heeft dat "ontspannen" een vrij positieve betekenis voor u heeft, dan antwoordt u een beetje rechts op de schaal

Opwinding / activiteit
Activiteit / opwinding is een maatstaaf voor opwinding versus kalme. Een afbeelding is ACTIEF als u zich hierdoor gestimuleerd, opgewonden, zenuwachtig of klaarwakker voelt. Een afbeelding is PASSIEF als u zich er ontspannen, kalm, traag, saai of slaperig van voelt. Geef aan hoe opwindend u elke afbeelding vindt op een *sliding scale* van ZEER PASSIEF tot ZEER ACTIEF, waarbij het middelpunt een matige opwinding vertegenwoordigt.
Voorbeelden:

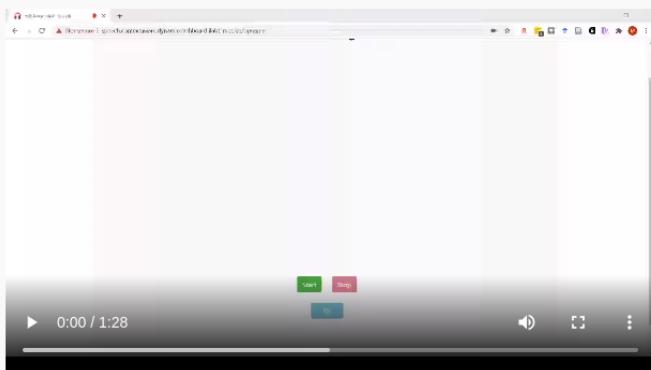
- Als u bijvoorbeeld vindt dat "hangmat" een vrij passieve betekenis heeft, dan antwoordt u vrij ver links op de sliding scale.
- Als u vindt dat "werken" een vrij actieve betekenis heeft, dan antwoordt u vrij ver rechts op de sliding scale.
- Als u vindt dat "mediteren" een zeer kalme betekenis heeft, dan antwoordt u ver links op de sliding scale
- Als u vindt dat "explosief" een zeer opwindende betekenis heeft, dan antwoordt u ver rechts op de schaal.

Er zullen afbeeldingen uit twee verschillende sets worden aangeboden, ofwel gezichten, ofwel tekeningen van bepaalde settings. Hieronder vindt u twee video's (1 voor elk van de twee sets) met hoe een trial eruit ziet: Start -> afbeelding omschrijven -> Stop -> Beoordelen.

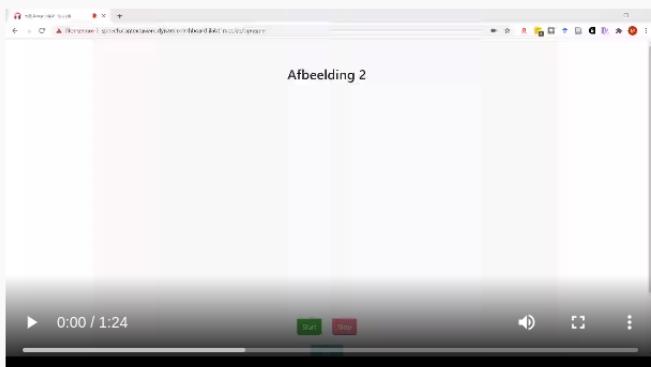
In deze video's ziet u een timer eronder om een indicatie te krijgen van hoe lang deze omschrijving is. Tijdens uw taak hebben we graag dat u minimaal zo'n 30 seconden omschrijft maar zult u geen timer zien. Maak u niet te veel druk mocht u vastlopen.

Kleine opmerking: De blauwe "Volgende" knop bij het ingeven van de gemoedstoestand wordt slechts zichtbaar eens het spraaksegment succesvol naar de server doorgestuurd is!

Gezicht



Tekening



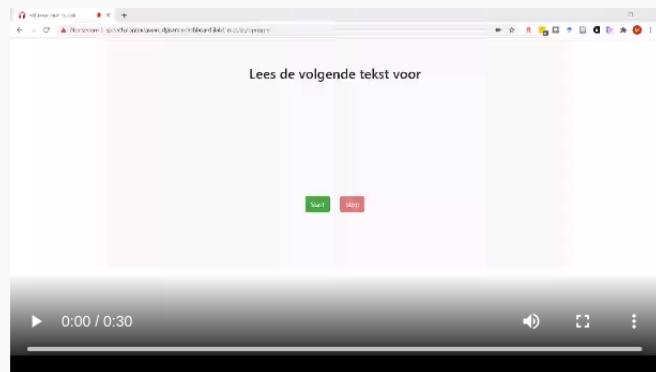
Tekst voorlezen

Tussen de afbeeldingen door krijgt u soms de instructie om een tekstje voor te lezen. Hiervoor wordt hetzelfde principe als bij de afbeeldingen toegepast; groene start-knop -> tekst verschijnt -> voorlezen -> rode stop-knop. De vooraf bepaalde tekst die u zal voorlezen, zal steeds de volgende zijn; Graag vragen we u om deze nu al eens hardop voor te lezen:

Papa en Marloes staan op het station.
Ze wachten op de trein.
Eerst hebben ze een kaartje gekocht.
Er stond een hele lange rij, dus dat duurde wel even.
Nu wachten ze tot de trein eraan komt.
Het is al vijf over drie, dus het duurt nog vier minuten.
Er staan nog veel meer mensen te wachten.
Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

Na het opnemen van deze tekst, zal opnieuw gevraagd worden, om uw gemoedstoestand in te geven via de sliders. Als extra illustratie van dit proces kan u naar onderstaande video kijken.

Tekst voorlezen



Pauzes

Ook krijgt u soms de instructie om een slok water te nemen. Doe dit zeker, aangezien uw keel anders uitdroogt, wat onprettig is en de opname beïnvloedt.

Tot slot willen we u vragen om de audiokwaliteit van uw microfoon te testen. Hiervoor drukt u op start, spreekt u iets in, waarna u de opname kan beluisteren door op pauze te drukken.

Eens dit werkt, en u op de blauwe knop heeft gedrukt, zal u doorverwezen worden naar een **blanco-pagina** die u voor 5 minuten te zien krijgt, het is de bedoeling voor deze periode om uw ogen te sluiten en op de ademhaling te focussen, zodat u helemaal tot rust komt. Eens deze periode voorbij is zal u een toon horen.

Succes!!

Test uw audiokwaliteit:

Start **Stop** **Play**



S1.4. Rest Block

Figure 6

Rest block screenshot.



S1.5. "Marloes" Text

Papa en Marloes staan op het station.

Ze wachten op de trein.

Eerst hebben ze een kaartje gekocht.

Er stond een hele lange rij, dus dat duurde wel even.

Nu wachten ze tot de trein eraan komt.

Het is al vijf over drie, dus het duurt nog vier minuten.

Er staan nog veel meer mensen te wachten.

Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

S1.6. GSSP Web App Image Subsets

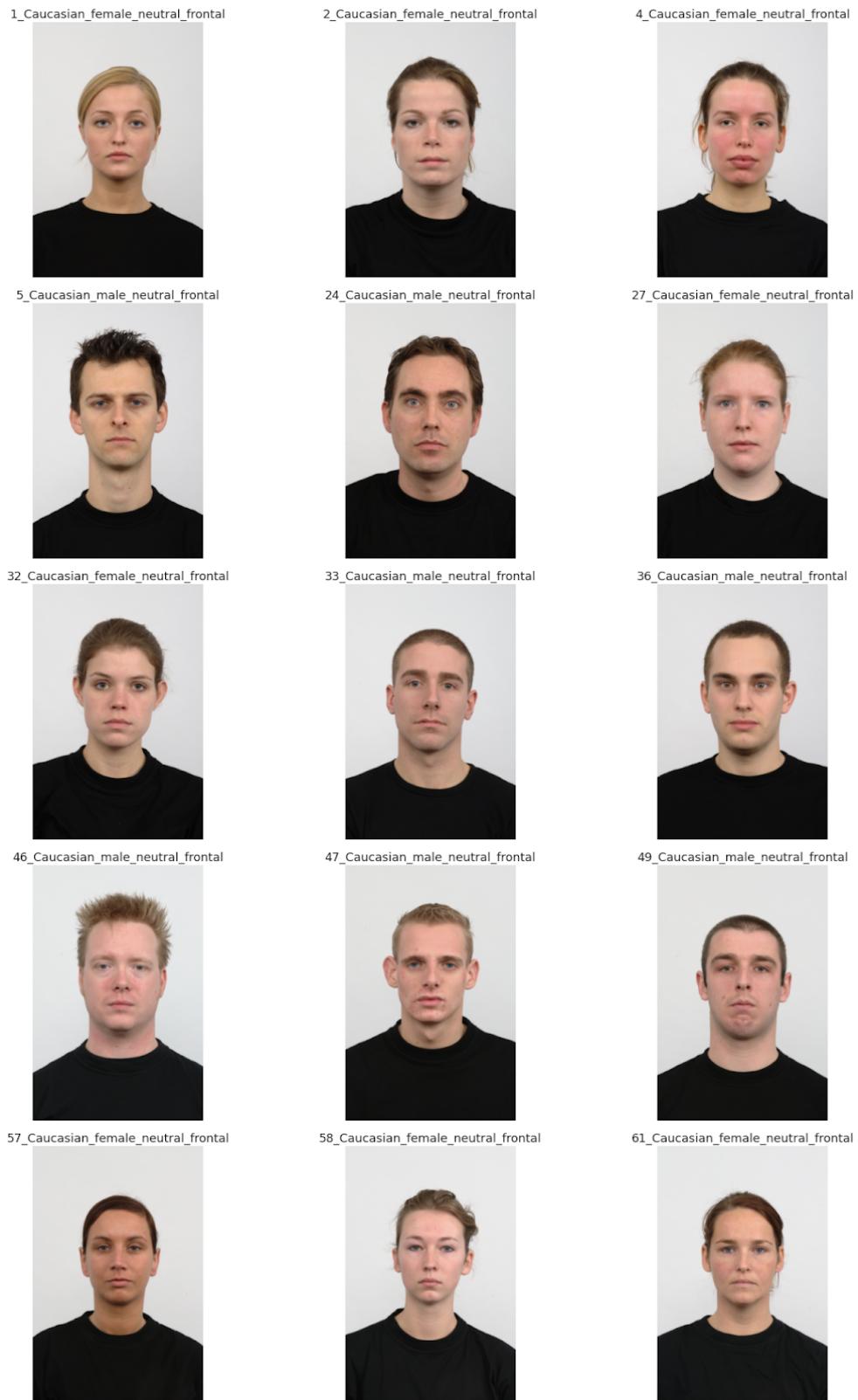
Figure 7

PiCES image subset.



Figure 8

Radboud faces image subset.



4.7.3.S2. Speech Data Parsing

Figure 9

Visualizations employed during the participant audio analysis step.



Note. The upper plot highlights the recorded, non-VAD cropped, utterance duration for each database subset, allowing to detect duration outliers. Below this duration plot, two utterances from the PiSCES subset are analyzed. For each utterance, the raw and transformed audio can be listened to. Below the audio players, a time-series visualization highlights the predictions of a voice activity detection (VAD) model and the extracted openSMILE Low Level Descriptors (LLDs). The VAD predictions are used to detect the first and last speech segments, which on its end determine the regions that will be omitted in the parsing block, i.e., the red shaded areas on the upper subplot. The purpose of the two lower subplots is to assess the ability of OpenSMILE to qualitatively extract speech metrics from the excerpts. The chosen metrics, fundamental frequency (F0) and jitter, are useful indicators of the stability of the feature extraction process. Finally, the table at the bottom of the figure shows the correlation of the extracted speech features with respect to the raw (non-resampled) WAV file. The visualization code can be found here¹¹.

Figure 10

Manual inspection of a participant with a large silent part at the end.



Note. The Voice Activity Detection (VAD) segmentation is able to detect this silence. The red-shaded rectangle indicates that this part will not be included in the parsed segment.

¹¹ https://github.com/predict-idlab/gssp_analysis/notebooks/0.3_Process_audio_Analyze_quality.ipynb

4.7.3.S3. OpenSMILE Feature Subset

Table 1

Description of utilized OpenSMILE GeMAPSv01b Functional features.

GeMAPSv01b name	Description
Temporal	
loudnessPeaksPerSec	The mean rate of loudness peaks, i.e., the number of loudness peaks per second.
MeanVoicedSegmentLengthSec	The mean length of continuously voiced regions ($F_0 > 0$).
MeanUnvoicedSegmentLength	The mean length and the standard deviation of unvoiced
StddevUnvoicedSegmentLength	regions (i.e., $F_0 = 0$; approximating pauses).
Spectral	
F0semitoneFrom27.5Hz_sma3nz_amean	Aggregation of moving windows in which the Logarithmic
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	F_0 is computed on a semitone frequency scale, starting at
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	27.5 Hz (semitone 0). The moving window output is first smoothed by a moving average with a window size of 3 that only includes non-zero values (sma3nz). The aggregations are respectively: mean, standard deviation, and range of 20th to 80th percentile. To convert the semitone frequency ($F_{0_{st}}$) to hertz, the following formula can be applied:
$F_0_{Hz} = 27.5Hz * 2^{F_{0_{st}}/12}$	
jitterLocal_sma3nz_amean	Mean aggregation of moving window in which the deviation of individual consecutive F_0 period lengths is computed.
Amplitude	
loudness_sma3_amean	Aggregation of moving windows of perceived signal
loudness_sma3_percentile50.0	intensity from an auditory spectrum. The aggregation are
loudness_sma3_pctlrange0-2	respectively: mean, median, and range of 20th to 80th percentile.

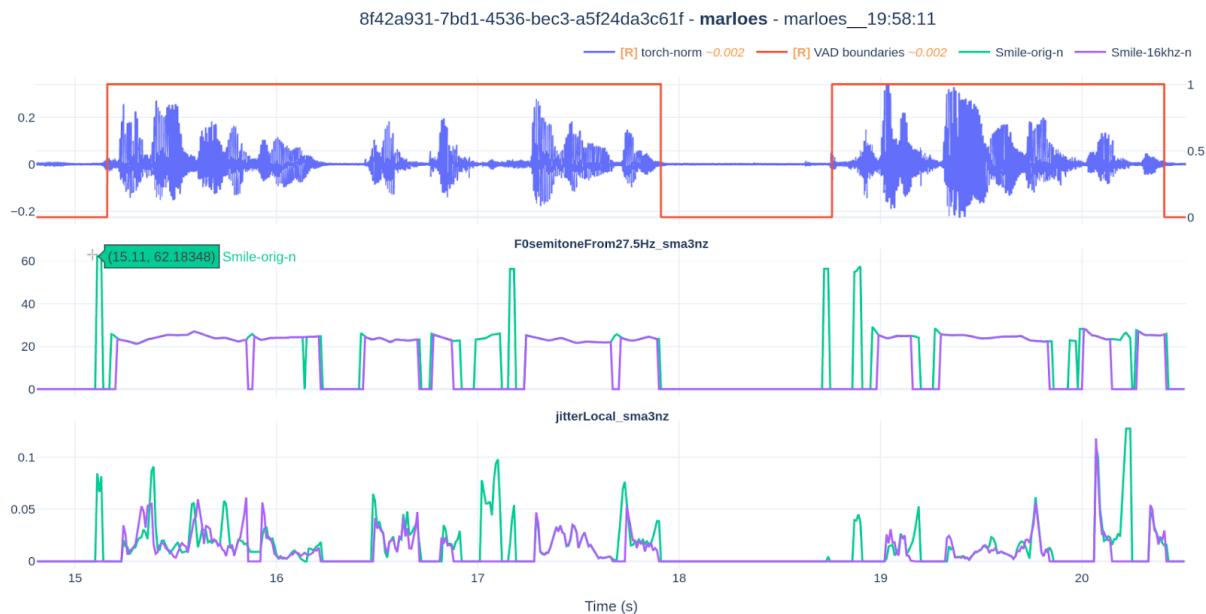
shimmerLocaldB_sma3nz_amean Mean aggregation of moving windows of the differences of the peak amplitudes of consecutive F0 periods.

Note. We further refer to Appendix 6.1 of (Eyben et al., 2016) for implementation details.

4.7.3.S4. OpenSMILE Sampling Rate Inconsistency

Figure 11

Illustration of inconsistency in the GeMAPSv01b Low-Level-Descriptors (LLDs) values when varying the sample frequency. The selection of F0Semitone and jitterLocal as features was based on prior utilization in the current work and their interpretability



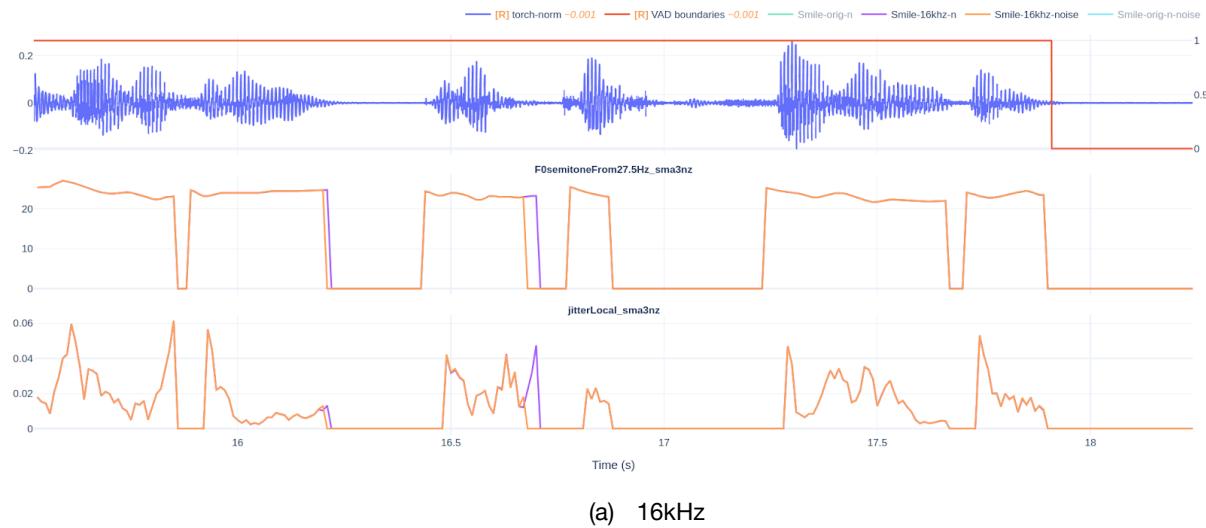
The above figure demonstrates the instability of the GeMAPSv01b Low-Level-Descriptors (LLDs) when the audio sample rate is altered. At approximately second 15 and 17, the *F0semitone* and *jitterLocal* metrics exhibit large values during non-voiced segments in the case of the original 44.1kHz signal (represented by the green trace). In particular, an *F0semitone* value of 62 represents an F0 of 987Hz, which is deemed implausible.

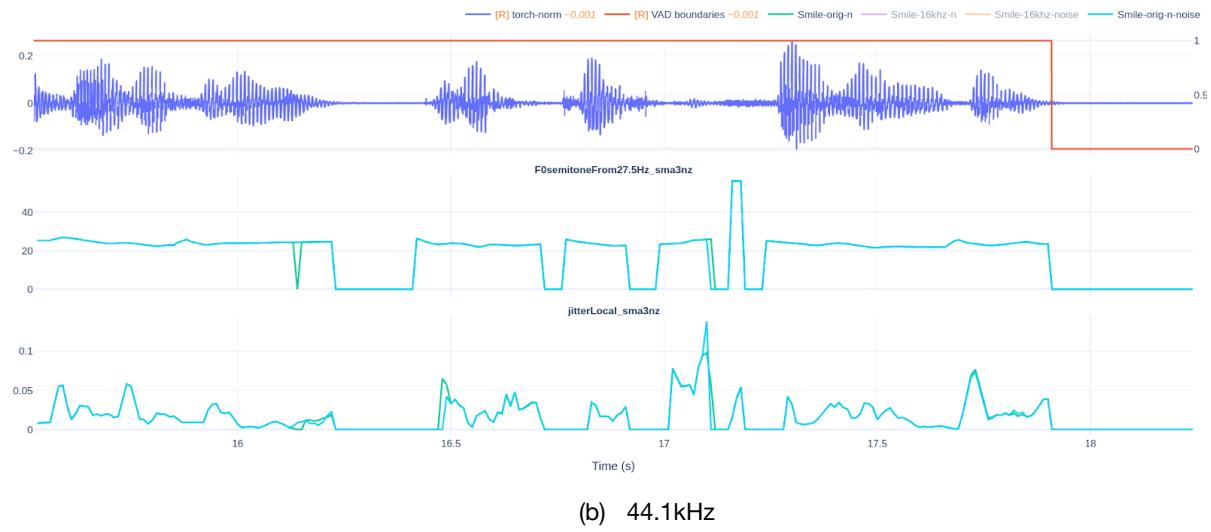
The original 44.1kHz speech signal was resampled to 16kHz using TorchAudio's resample method, which applies sinc-interpolation (Yang et al., 2021).

Our initial hypothesis was that the original 44.1kHz audio contains high-frequency harmonics (e.g., whirring PC-fan) that are more easily picked-up when OpenSMILE is used in certain configurations (in this case a higher sampling rate). To test this hypothesis, we added high-frequency Gaussian noise of -30dB to the audio to determine if it would reduce the ability to detect these harmonics. The results for a single segment are depicted in the figure below. The 16kHz resampled data showed an expected outcome; the signal-to-noise ratio at voiced boundaries for the noisy signal, represented by the orange trace of (a), was slightly trimmed at second 16.25 and 16.75, resulting in a decrease of higher jitter values. Conversely, the addition of noise to the 44.1kHz signal, represented by the blue trace of (b), did not result in an improved detection of unvoiced regions. As such, there was no decrease in F0 or jitter values. Hence, we can conclude that resampling high-frequency seems to contribute more to improved voiced boundary detection than the Gaussian-noise addition.

Figure 12

Impact of Gaussian noise superposition to (resampled) audio.



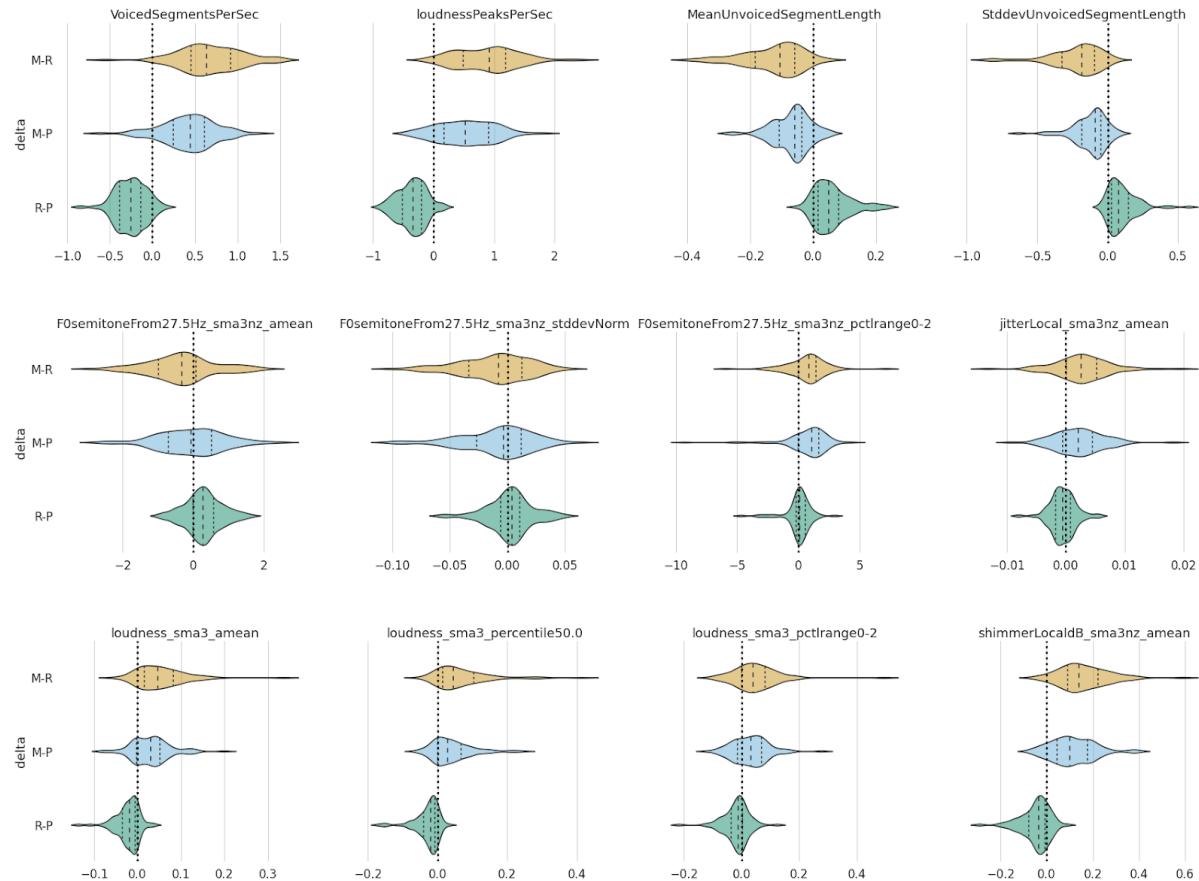


(b) 44.1kHz

4.7.3.S5. OpenSMILE Delta Visualizations

Figure 13

GeMAPSv01b feature subset delta visualizations, divided into temporal (row 1), frequency (row 2), and amplitude (row 3) related features.



The feature subplots in the above graph exhibit a general trend; the deltas between the unscripted-to-scripted speech styles (i.e., M-R, M-P) have a greater variation and a consistent offset. Conversely, the deltas between the unscripted-to-unscripted tasks (i.e., R-P) have a more limited variation and the offset is closer to the 0 delta value, which suggests a greater similarity in the speech features.

4.7.3.S6. ECAPA-TDNN & GeMAPS Distribution Plots

Figure 14

KDE plot, depicting the distribution of the web application ECAPA-TDDN embeddings. A subset of embedding dimensions was chosen, each displaying a normal distribution.

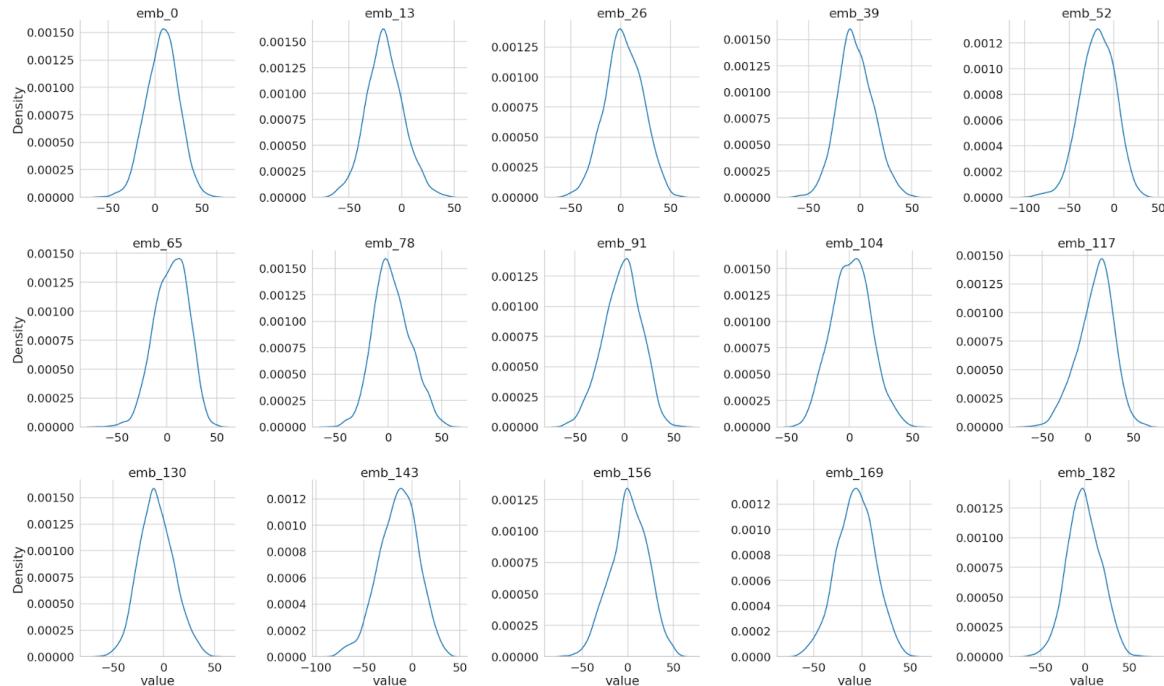
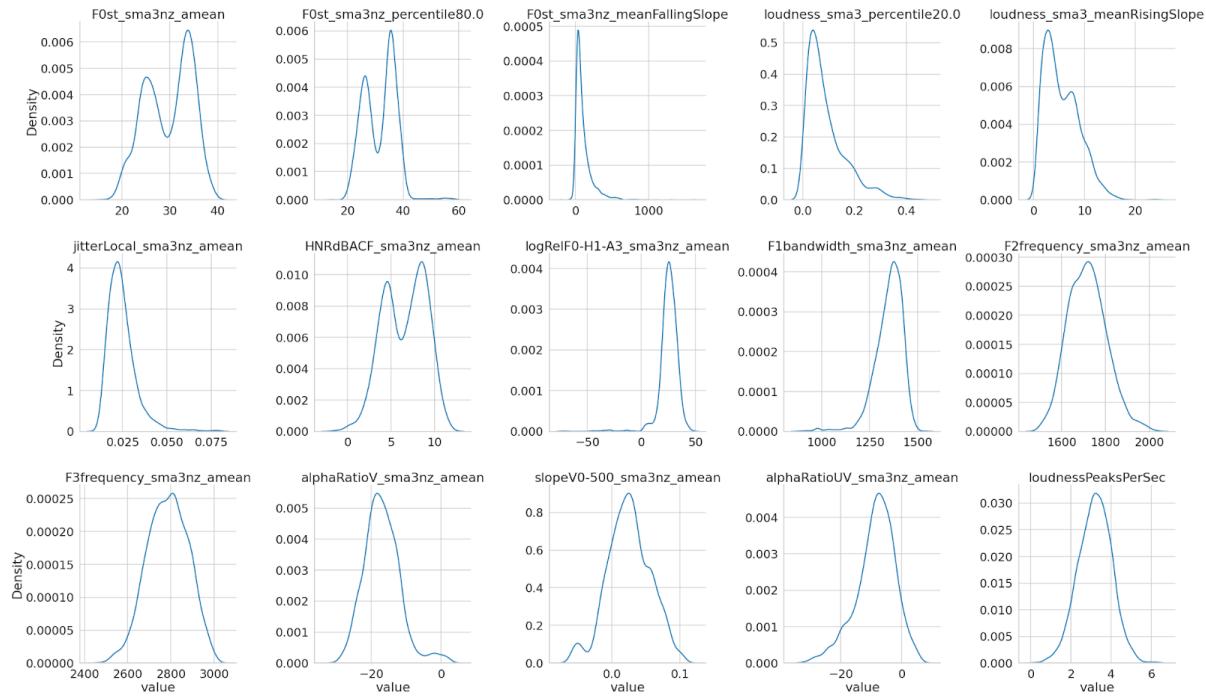


Figure 15

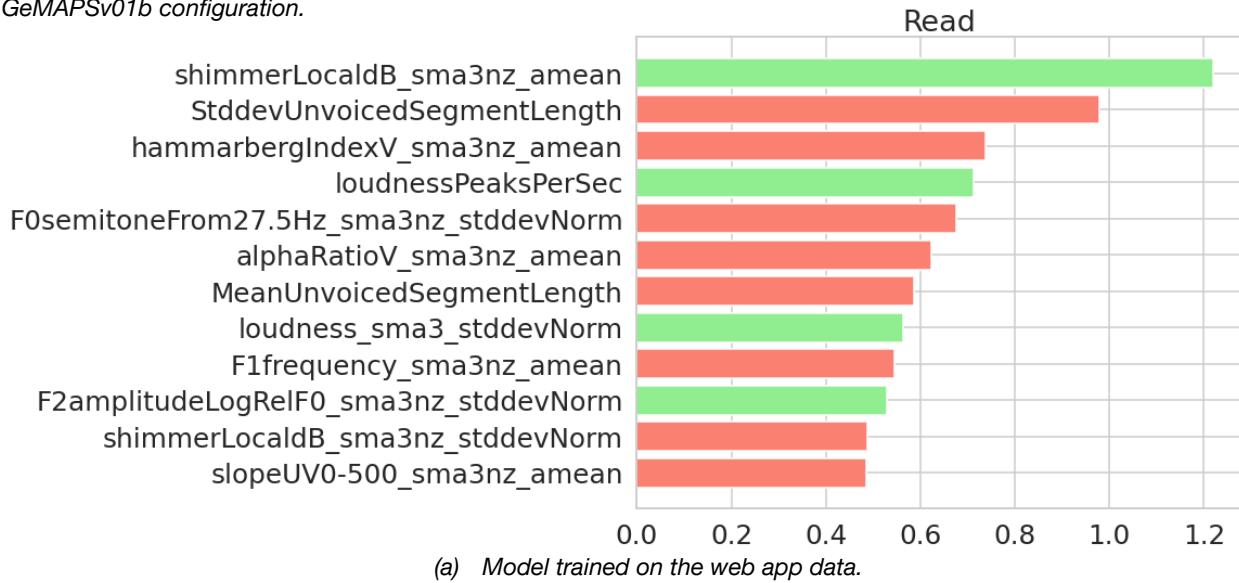
KDE plot of the web application GeMAPSv01b functional features, indicating non-normal distributions.

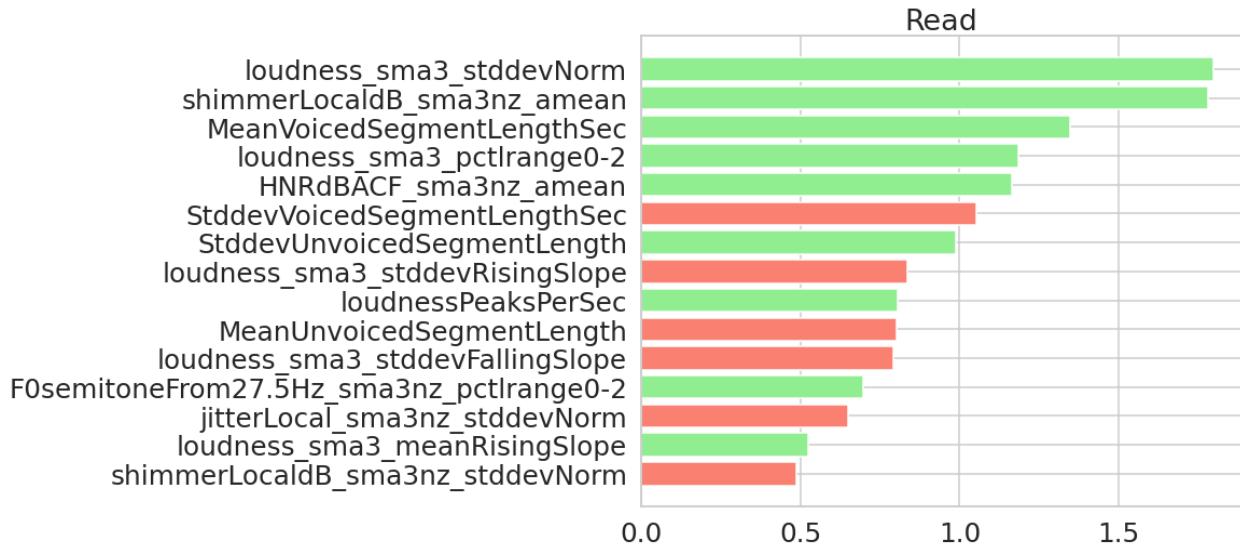


4.7.3.S7. Logistic Regression Weight Coefficients

Figure 16

Visualization of the 15 largest feature coefficients of the logistic regression models that were trained on GeMAPSv01b configuration.





(b) Model trained on CGN data.

Note. The color indicates whether the feature coefficient is positive (green) or negative (red) in relation to the target variable (Read speech).

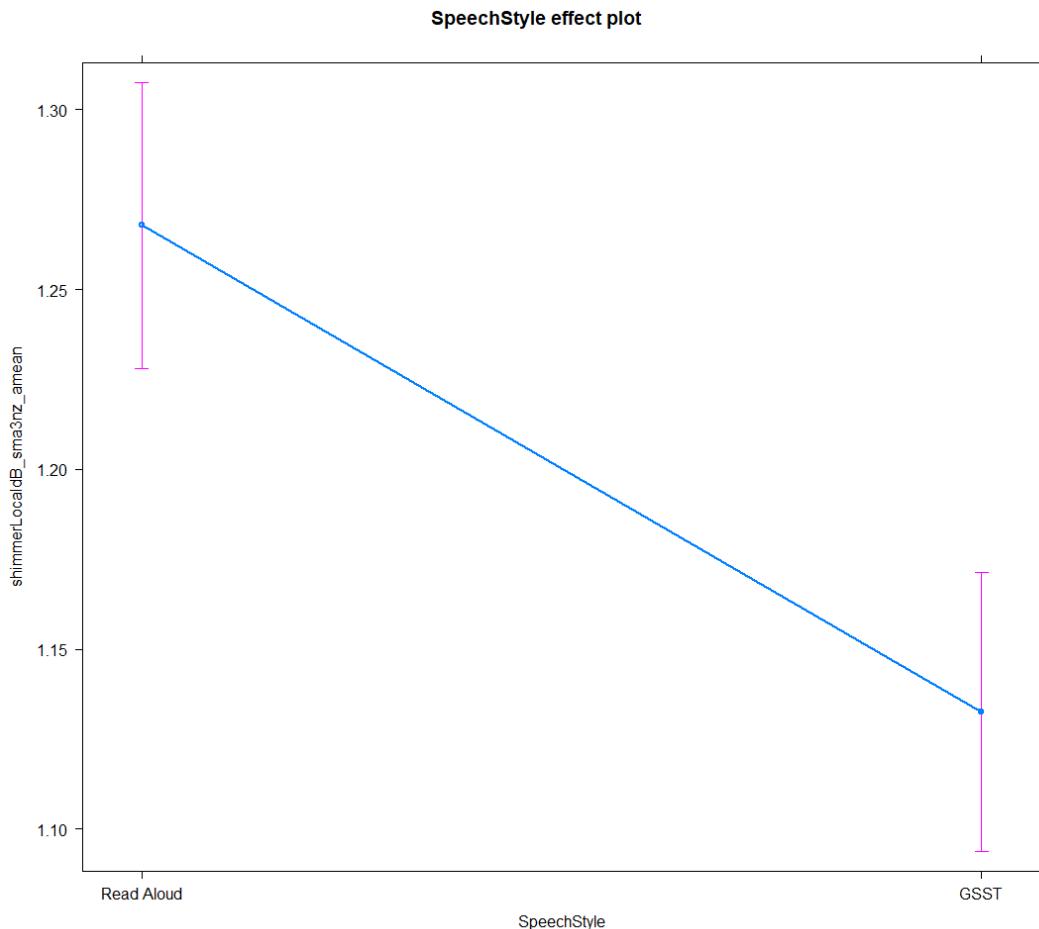
Remark how the weight coefficient for the “shimmerLocaldB_sma3nz_amean” exhibits a substantial positive value in both subplots. This positive sign for the coefficient can be interpreted as indicating that a decrease in shimmer contributes to a higher likeliness of having read-aloud speech, which contradicts existing literature. It is of particular interest that this trend is also observed when fitting a model on the CGN data (b), suggesting that the OpenSMILE GeMAPSv01b shimmer values tend to decrease as the speech becomes less scripted. The “jitterLocal_sma3_stddevNorm” parameter, has a smaller value and is not incorporated in the top 15 values for the CGN model (b).

4.7.3.S8. Effect Size Shimmer & Jitter

S8.1. Shimmer

Figure 17

Effect plot of shimmer (a) and corresponding screenshots of computation process (b-c).



```
> Anova(d0.1, type = 'III')
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: shimmerLocaldB_sma3nz_amean
          chisq  DF Pr(>chisq)
(Intercept) 3658.37  1 < 2.2e-16 ***
Speechstyle  535.32  1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) ANOVA.

```

> emmeans0.1 <- emmeans(d0.1, pairwise ~ speechstyle, adjust = "none", type = "response")
> emm0.1 <- summary(emmeans0.1)$emmeans
> emmeans0.1$contrasts
  contrast      estimate      SE   df t.ratio p.value
Read Aloud - GSST    0.135 0.00585 2829  23.137 <.0001

Degrees-of-freedom method: kenward-roger
> effsummary <- summary(eff_size(emmeans0.1, sigma=sigma(d0.1), edf=df.residual(d0.1)))
Since 'object' is a list, we are using the contrasts already present.
> effsummary
  contrast      effect.size      SE   df lower.CL upper.CL
(Read Aloud - GSST)        1.1 0.0496 2829       1     1.2

sigma used for effect sizes: 0.1232
Degrees-of-freedom method: inherited from kenward-roger when re-gridding
Confidence level used: 0.95

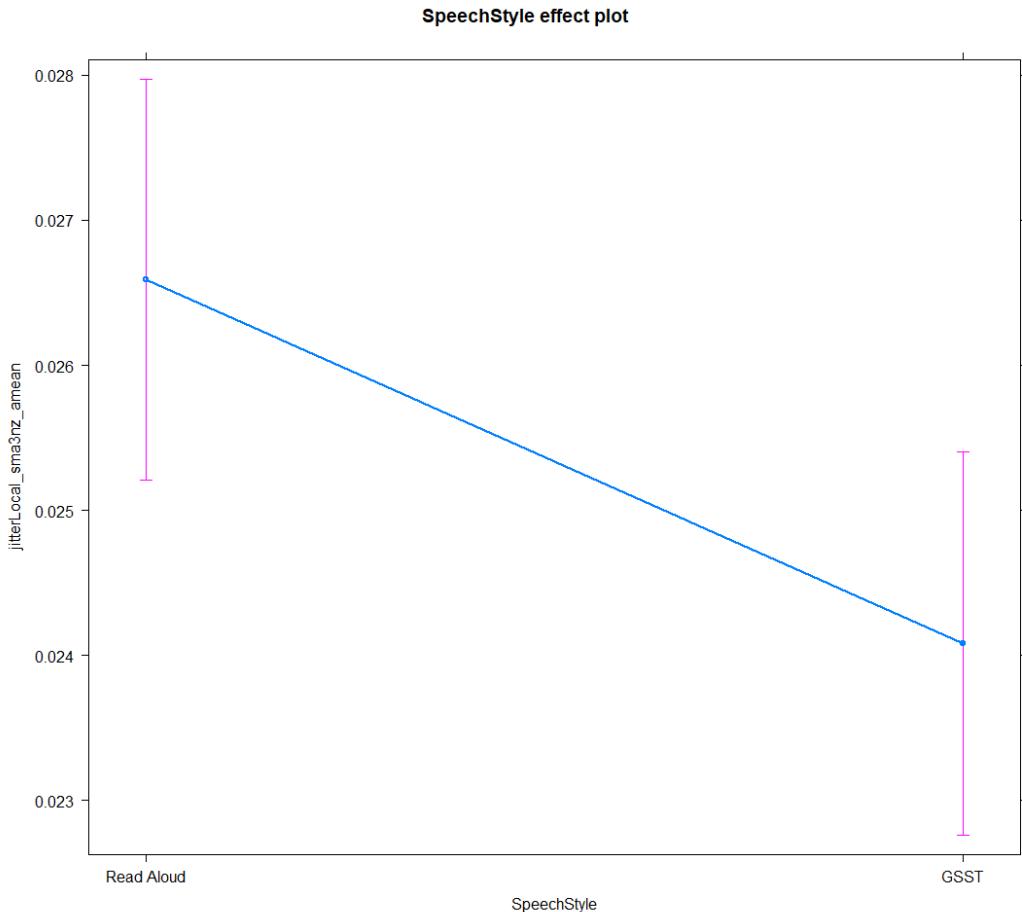
```

(c) Effect size summary.

S8.2. Jitter

Figure 18

Effect plot of jitter (a) and corresponding screenshots of computation process (b-c)



(a) *Effect plot of jitter.*

```
> Anova(d0.1, type = 'III')
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: jitterLocal_sma3nz_amean
          Chisq Df Pr(>chisq)
(Intercept) 1399.394  1 < 2.2e-16 ***
SpeechStyle  88.301  1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) ANOVA.

```
> emmeans0.1 <- emmeans(d0.1, pairwise ~ SpeechStyle, adjust = "none", type = "response")
> emm0.1 <- summary(emmeans0.1)$emmeans
> emmeans0.1$contrasts
  contrast      estimate     SE   df t.ratio p.value
Read Aloud - GSST  0.00251 0.000267 2831   9.397 <.0001

Degrees-of-freedom method: kenward-roger
> effSummary <- summary(eff_size(emmeans0.1, sigma=sigma(d0.1), edf=df.residual(d0.1)))
Since 'object' is a list, we are using the contrasts already present.
> effSummary
  contrast      effect.size     SE   df lower.CL upper.CL
(Read Aloud - GSST)       0.446 0.0478 2831    0.352    0.54

sigma used for effect sizes: 0.005624
Degrees-of-freedom method: inherited from kenward-roger when re-gridding
Confidence level used: 0.95
```

(c) *Effect size summary.*

4.7.3.S9. Factor Analysis

Interactive html RMarkdown file can be found here:

https://github.com/predict-idlab/gssp_analysis/blob/master/scripts/1.2_FactorAnalysis.html

R Markdown

This is an R Markdown document displaying the code and output for the cfa and glmm's ran for valence and arousal for two image sets.

This results in the following (clickable) structure

- 1.0. Pisces Dataset
 - 1.1. Valence
 - 1.1.1. Cronbach's Alpha
 - 1.1.2. CFA
 - 1.1.3. CFA Visualization
 - 1.1.4. Distributions

- 1.2. Arousal
 - 1.2.1. Cronbach's Alpha
 - 1.2.2. CFA
 - 1.2.3. CFA Visualization
 - 1.2.4. Distributions
- [2.0. Radboud faces]
 - [2.1. Valence]
 - 2.1.1. Cronbach's Alpha
 - 2.1.2. CFA
 - 2.1.3. CFA Visualization
 - 2.1.4. Distributions
 - [2.2. Arousal]
 - 2.2.1. Cronbach's Alpha
 - 2.2.2. CFA
 - 2.2.3. CFA Visualization
 - 2.2.4. Distributions

General code

Used to load and prepare dataframes

```
##### Set environment #####
rm(list = ls()) # Clear environment
cat("\014") # Clear console
dev.off() # Clear plot window
options(contrasts=c("contr.sum", "contr.poly")) # Set contrast settings to
# effect coding

# Libraries
library(arrows)
library(lavaan)
library(lavaanPlot)
library(psych)
library(ltm)
library(car)
library(ggplot2)
library(ggstatsplot)
library(Polychrome)

#GLM specific
library(lme4)
library(lmerTest)
library(emmeans)
library(effects)
```

```
##### Loading data #####
imageData
<-as.data.frame(read_parquet("../loc_data/df_session_tot_cleaned.parquet"))

piscesData <- imageData[imageData$DB == 'PiSCES',]
radboudData <- imageData[imageData$DB == 'Radboud',]
marloesData <- imageData[imageData$DB == 'marloes',]
```

1.0. Pisces Dataset

1.1. Valence

```
##### Valence #####
piscesDataClean = piscesData[c("ID", "pic_name", "valence")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean = reshape(piscesDataClean, idvar = "ID", timevar =
"pic_name", direction = "wide")
piscesDataCronbachs = piscesDataClean[, 2:16]
```

1.1.1. Cronbach's Alpha

```
# Calculate Cronbach's alpha using alpha()
alphavar = psych::alpha(piscesDataCronbachs, check.keys = TRUE)
summary(alphavar)

##
## Reliability analysis
## raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
##      0.84      0.84      0.88      0.26 5.4 0.025    57  8     0.27
```

1.1.2. CFA

```
names(piscesDataClean)[2:16] = c("Picture_105", "Picture_82", "Picture_118",
"Picture_65", "Picture_88", "Picture_87", "Picture_59", "Picture_93",
"Picture_56", "Picture_81",
                           "Picture_110", "Picture_96", "Picture_132",
"Picture_80", "Picture_98" )

HS.model <- 'pisces =~ Picture_105 + Picture_82 + Picture_118 + Picture_65 +
Picture_88 + Picture_87 + Picture_59 + Picture_93 + Picture_56 + Picture_81 +
Picture_110 + Picture_96 + Picture_132 + Picture_80 + Picture_98'
```

Fit and visualize

```
## lavaan 0.6-9 ended normally after 56 iterations
##
##   Estimator                               ML
##   Optimization method                    NLMINB
##   Number of model parameters              30
##
##                                         Used    Total
```

```

## Number of observations                                84          89
##
## Model Test User Model:
##
##   Test statistic                               188.181
##   Degrees of freedom                           90
##   P-value (Chi-square)                         0.000
##
## Model Test Baseline Model:
##
##   Test statistic                               466.939
##   Degrees of freedom                           105
##   P-value                                     0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                 0.729
##   Tucker-Lewis Index (TLI)                     0.684
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)                -4979.918
##   Loglikelihood unrestricted model (H1)         -4885.827
##
##   Akaike (AIC)                                 10019.835
##   Bayesian (BIC)                               10092.760
##   Sample-size adjusted Bayesian (BIC)           9998.124
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                    0.114
##   90 Percent confidence interval - lower      0.091
##   90 Percent confidence interval - upper      0.137
##   P-value RMSEA <= 0.05                      0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                    0.099
##
## Parameter Estimates:
##
##   Standard errors                            Standard
##   Information                                Expected
##   Information saturated (h1) model            Structured
##
## Latent Variables:
##                  Estimate Std.Err z-value P(>|z|) Std.lv Std.all
##   pisces =~
##     Picture_105      5.297    1.451    3.651    0.000    5.297    0.407

```

```

##    Picture_82      4.740     1.734     2.733     0.006     4.740     0.311
##    Picture_118     8.769     1.328     6.603     0.000     8.769     0.673
##    Picture_65      8.353     1.519     5.498     0.000     8.353     0.582
##    Picture_88      4.194     1.977     2.122     0.034     4.194     0.244
##    Picture_87     11.781     2.013     5.853     0.000    11.781     0.612
##    Picture_59      5.198     1.336     3.891     0.000     5.198     0.431
##    Picture_93      7.133     1.309     5.451     0.000     7.133     0.578
##    Picture_56      8.063     1.239     6.509     0.000     8.063     0.665
##    Picture_81      9.692     1.413     6.861     0.000     9.692     0.692
##    Picture_110     6.620     1.515     4.369     0.000     6.620     0.478
##    Picture_96      5.934     1.575     3.766     0.000     5.934     0.419
##    Picture_132     6.329     1.508     4.196     0.000     6.329     0.462
##    Picture_80      9.759     1.681     5.807     0.000     9.759     0.608
##    Picture_98      8.113     1.287     6.302     0.000     8.113     0.649
##
## Variances:
##          Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## .Picture_105   141.212  22.442   6.292   0.000 141.212  0.834
## .Picture_82    209.994  32.918   6.379   0.000 209.994  0.903
## .Picture_118    93.033  16.340   5.693   0.000  93.033  0.548
## .Picture_65    136.514  22.772   5.995   0.000 136.514  0.662
## .Picture_88    278.204  43.328   6.421   0.000 278.204  0.941
## .Picture_87    231.777  39.210   5.911   0.000 231.777  0.625
## .Picture_59    118.188  18.868   6.264   0.000 118.188  0.814
## .Picture_93    101.667  16.930   6.005   0.000 101.667  0.666
## .Picture_56     81.859  14.300   5.724   0.000  81.859  0.557
## .Picture_81    101.959  18.198   5.603   0.000 101.959  0.520
## .Picture_110   147.831  23.846   6.199   0.000 147.831  0.771
## .Picture_96    165.490  26.356   6.279   0.000 165.490  0.825
## .Picture_132   148.002  23.780   6.224   0.000 148.002  0.787
## .Picture_80    162.325  27.408   5.923   0.000 162.325  0.630
## .Picture_98     90.502  15.635   5.788   0.000  90.502  0.579
##    pisces        1.000

```

1.1.3. CFA Visualization

Pisces dataset - Valence

1.1.4. Distributions

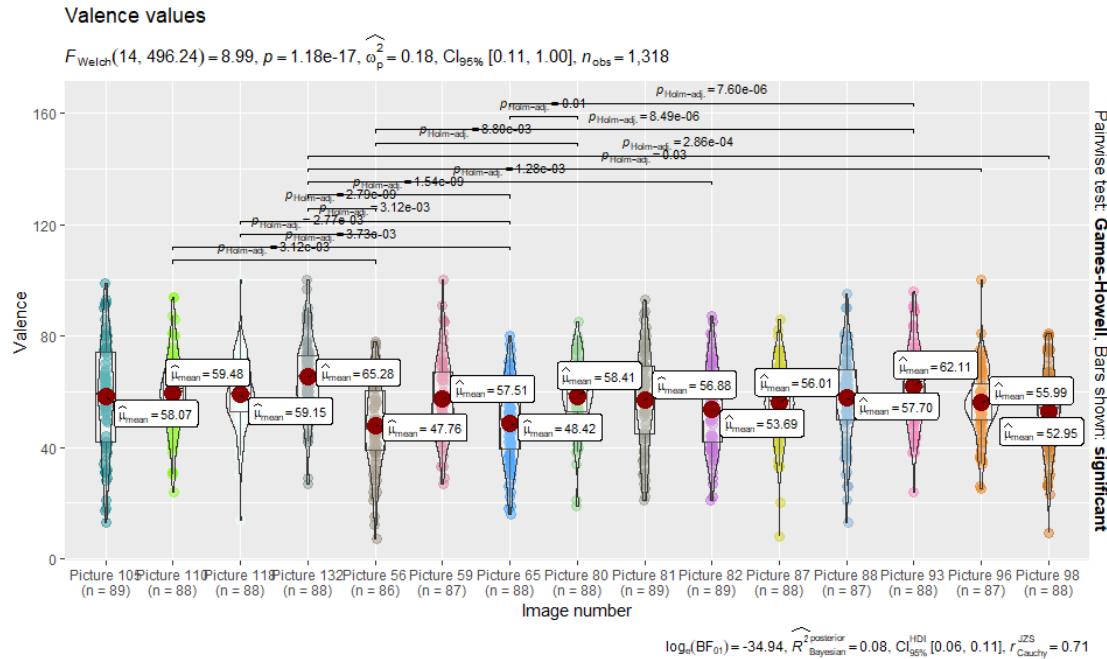
```

# Re-prep data
piscesDataClean = piscesData[c("ID", "pic_name", "valence")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean$ID = as.factor(piscesDataClean$ID)

```

Visualizations

Pisces - Valence



1.2. Arousal

Arousal

```
piscesDataClean = piscesData[c("ID", "pic_name", "arousal")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean = reshape(piscesDataClean, idvar = "ID", timevar =
"pic_name", direction = "wide")
piscesDataCronbachs = piscesDataClean[, 2:16]
```

1.2.1. Cronbach's Alpha

```
# Calculate Cronbach's alpha using alpha()
alphavar = psych::alpha(piscesDataCronbachs, check.keys = TRUE)
summary(alphavar)
```

```
##
```

```
## Reliability analysis
```

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	median_r
0.94	0.94	0.95	0.49	14	0.01	48	14	0.51

1.2.2. CFA

```
names(piscesDataClean)[2:16] = c("Picture_105", "Picture_82", "Picture_118",
"Picture_65", "Picture_88", "Picture_87", "Picture_59", "Picture_93",
"Picture_56", "Picture_81",
"Picture_110", "Picture_96", "Picture_132",
"Picture_80", "Picture_98" )
```

```
HS.model <- 'pisces =~ Picture_105 + Picture_82 + Picture_118 + Picture_65 +
Picture_88 + Picture_87 + Picture_59 + Picture_93 + Picture_56 + Picture_81 +
Picture_110 + Picture_96 + Picture_132 + Picture_80 + Picture_98'
```

Fit and visualize

```
## lavaan 0.6-9 ended normally after 19 iterations
##
##   Estimator                      ML
## Optimization method            NLMINB
## Number of model parameters    30
##
##   Used      Total
## Number of observations       84      89
##
## Model Test User Model:
##
##   Test statistic             193.015
##   Degrees of freedom          90
##   P-value (Chi-square)        0.000
##
## Model Test Baseline Model:
##
##   Test statistic             858.041
##   Degrees of freedom          105
##   P-value                     0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)  0.863
##   Tucker-Lewis Index (TLI)     0.840
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0) -5201.631
##   Loglikelihood unrestricted model (H1) -5105.123
##
##   Akaike (AIC)                 10463.261
##   Bayesian (BIC)                10536.186
##   Sample-size adjusted Bayesian (BIC) 10441.550
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                      0.117
##   90 Percent confidence interval - lower 0.094
##   90 Percent confidence interval - upper 0.139
##   P-value RMSEA <= 0.05           0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                       0.070
##
## Parameter Estimates:
```

```

## Standard errors
## Information
## Information saturated (h1) model
## Standard
## Expected
## Structured

## Latent Variables:
## Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## pisces =~
##   Picture_105    11.707  2.013  5.816  0.000  11.707  0.591
##   Picture_82     16.310  1.828  8.923  0.000  16.310  0.812
##   Picture_118    15.903  2.012  7.904  0.000  15.903  0.747
##   Picture_65     13.560  1.953  6.944  0.000  13.560  0.680
##   Picture_88     14.346  1.846  7.771  0.000  14.346  0.738
##   Picture_87     13.571  1.747  7.770  0.000  13.571  0.738
##   Picture_59     16.185  1.898  8.528  0.000  16.185  0.788
##   Picture_93     14.186  1.891  7.502  0.000  14.186  0.720
##   Picture_56     15.444  1.835  8.415  0.000  15.444  0.781
##   Picture_81     12.237  1.831  6.682  0.000  12.237  0.660
##   Picture_110    7.739   1.935  4.000  0.000   7.739  0.427
##   Picture_96     13.904  1.818  7.648  0.000  13.904  0.730
##   Picture_132    13.627  1.914  7.121  0.000  13.627  0.693
##   Picture_80     13.176  1.872  7.039  0.000  13.176  0.687
##   Picture_98     14.812  1.906  7.772  0.000  14.812  0.738
##
## Variances:
## Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## .Picture_105   255.417 40.745  6.269  0.000 255.417  0.651
## .Picture_82    137.429 24.055  5.713  0.000 137.429  0.341
## .Picture_118   199.787 33.412  5.979  0.000 199.787  0.441
## .Picture_65    214.218 34.882  6.141  0.000 214.218  0.538
## .Picture_88    171.575 28.568  6.006  0.000 171.575  0.455
## .Picture_87    153.637 25.580  6.006  0.000 153.637  0.455
## .Picture_59    160.072 27.447  5.832  0.000 160.072  0.379
## .Picture_93    187.074 30.896  6.055  0.000 187.074  0.482
## .Picture_56    152.821 26.069  5.862  0.000 152.821  0.391
## .Picture_81    194.121 31.433  6.176  0.000 194.121  0.565
## .Picture_110   267.851 41.901  6.392  0.000 267.851  0.817
## .Picture_96    169.447 28.105  6.029  0.000 169.447  0.467
## .Picture_132   201.360 32.925  6.116  0.000 201.360  0.520
## .Picture_80    194.607 31.758  6.128  0.000 194.607  0.529
## .Picture_98    182.901 30.454  6.006  0.000 182.901  0.455
##   pisces        1.000               1.000  1.000

```

1.2.3. CFA Visualization

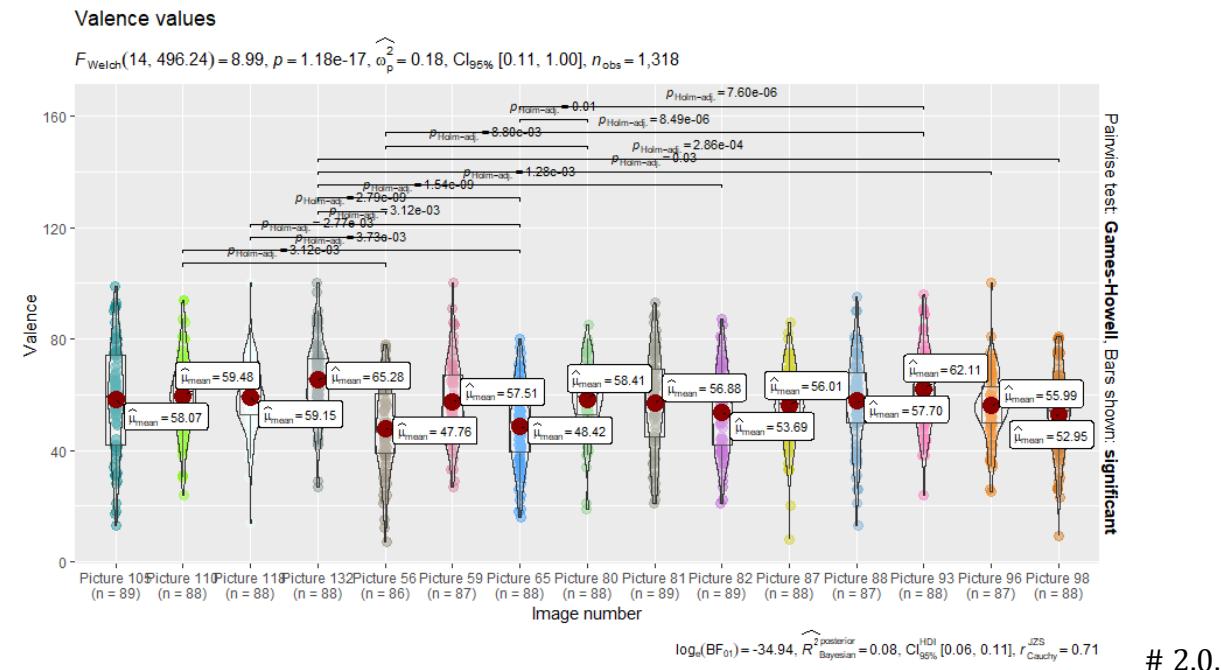
Pisces dataset - Arousal

1.2.4. Distributions

```
# Re-prep data
piscesDataClean = piscesData[c("ID", "pic_name", "valence")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean$ID = as.factor(piscesDataClean$ID)
```

Visualizations

Pisces - Arousal



Radboud faces ## 2.1. Valence

```
##### Valence #####
radboudDataClean = radboudData[c("ID", "pic_name", "valence")]
radboudDataClean$pic_name = as.factor(radboudDataClean$pic_name)
radboudDataClean = reshape(radboudDataClean, idvar = "ID", timevar =
"pic_name", direction = "wide")
radboudDataCronbachs = radboudDataClean[, 2:16]
```

2.1.1. Cronbach's Alpha

```
# Calculate Cronbach's alpha using alpha()
alphavar = psych::alpha(radboudDataCronbachs, check.keys = TRUE)
summary(alphavar)

##
## Reliability analysis
## raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
##      0.89      0.89     0.91      0.36 8.3 0.017    51 8.6     0.35
```

2.1.2. CFA

```
names(radboudDataClean)[2:16] = c('Face_01', 'Face_36', 'Face_32', 'Face_61',
'Face_04', 'Face_24', 'Face_02', 'Face_49', 'Face_58', 'Face_46', 'Face_05',
'Face_33', 'Face_57', 'Face_47', 'Face_27')

HS.model <- 'radboud =~ Face_01 + Face_36 + Face_32 + Face_61 + Face_04 +
Face_24 + Face_02 + Face_49 + Face_58 + Face_46 + Face_05 + Face_33 + Face_57
+ Face_47 + Face_27'
```

Fit and visualize

```
## lavaan 0.6-9 ended normally after 20 iterations
##
##    Estimator                      ML
## Optimization method            NLMINB
## Number of model parameters      30
##
##                                Used    Total
## Number of observations        85      89
##
## Model Test User Model:
##
##    Test statistic              174.182
##    Degrees of freedom           90
##    P-value (Chi-square)        0.000
##
## Model Test Baseline Model:
##
##    Test statistic              571.377
##    Degrees of freedom          105
##    P-value                     0.000
##
## User Model versus Baseline Model:
##
##    Comparative Fit Index (CFI)   0.819
##    Tucker-Lewis Index (TLI)       0.789
##
## Loglikelihood and Information Criteria:
##
##    Loglikelihood user model (H0) -4927.772
##    Loglikelihood unrestricted model (H1) -4840.681
##
##    Akaike (AIC)                  9915.544
##    Bayesian (BIC)                 9988.824
##    Sample-size adjusted Bayesian (BIC) 9894.180
##
## Root Mean Square Error of Approximation:
##
##    RMSEA                         0.105
```

```

## 90 Percent confidence interval - lower      0.081
## 90 Percent confidence interval - upper      0.128
## P-value RMSEA <= 0.05                   0.000
##
## Standardized Root Mean Square Residual:
##
## SRMR                                0.078
##
## Parameter Estimates:
##
## Standard errors                      Standard
## Information                           Expected
## Information saturated (h1) model     Structured
##
## Latent Variables:
##                         Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## radboud =~
##   Face_01        7.066   1.485   4.757  0.000   7.066  0.505
##   Face_36        7.284   1.263   5.767  0.000   7.284  0.594
##   Face_32        8.577   1.308   6.556  0.000   8.577  0.658
##   Face_61        7.407   1.319   5.617  0.000   7.407  0.581
##   Face_04        8.736   1.527   5.723  0.000   8.736  0.590
##   Face_24        7.528   1.344   5.600  0.000   7.528  0.580
##   Face_02       10.139   1.364   7.433  0.000  10.139  0.723
##   Face_49        9.735   1.498   6.499  0.000   9.735  0.653
##   Face_58        8.523   1.404   6.070  0.000   8.523  0.619
##   Face_46        7.598   1.506   5.045  0.000   7.598  0.531
##   Face_05        7.625   1.377   5.537  0.000   7.625  0.575
##   Face_33        9.031   1.364   6.620  0.000   9.031  0.663
##   Face_57        6.207   1.432   4.334  0.000   6.207  0.466
##   Face_47        9.368   1.350   6.941  0.000   9.368  0.687
##   Face_27        7.324   1.228   5.962  0.000   7.324  0.610
##
## Variances:
##                         Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## .Face_01        145.570  23.206   6.273  0.000 145.570  0.745
## .Face_36         97.307  15.881   6.127  0.000  97.307  0.647
## .Face_32         96.528  16.166   5.971  0.000  96.528  0.568
## .Face_61        107.425  17.461   6.152  0.000 107.425  0.662
## .Face_04        142.697  23.260   6.135  0.000 142.697  0.652
## .Face_24        111.813  18.166   6.155  0.000 111.813  0.664
## .Face_02         94.130  16.420   5.733  0.000  94.130  0.478
## .Face_49        127.330  21.279   5.984  0.000 127.330  0.573
## .Face_58         116.898  19.251   6.072  0.000 116.898  0.617
## .Face_46        146.669  23.518   6.237  0.000 146.669  0.718
## .Face_05        117.988  19.138   6.165  0.000 117.988  0.670
## .Face_33        104.206  17.496   5.956  0.000 104.206  0.561
## .Face_57        139.054  22.001   6.320  0.000 139.054  0.783
## .Face_47         98.264  16.722   5.876  0.000  98.264  0.528

```

##	.Face_27	90.419	14.840	6.093	0.000	90.419	0.628
##	rajboud	1.000				1.000	1.000

2.1.3. CFA Visualization

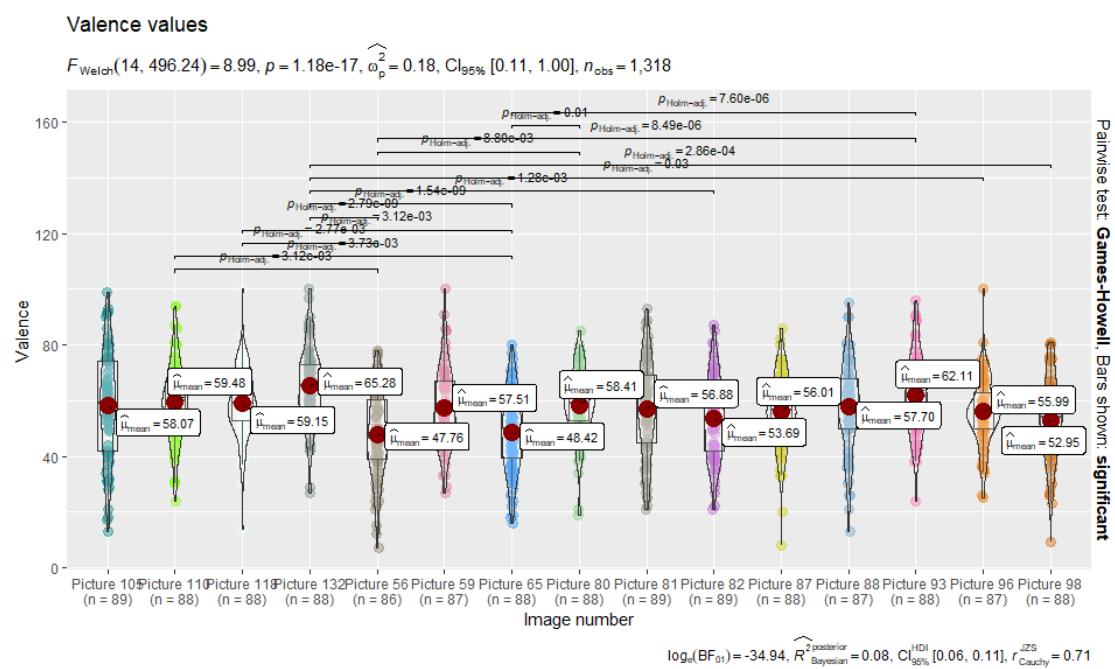
Radboud dataset - Valence

2.1.4. Distributions

```
# Re-prep data
piscesDataClean = piscesData[c("ID", "pic_name", "valence")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean$ID = as.factor(piscesDataClean$ID)
```

Visualizations

Pisces - Valence



2.2.

Arousal

```
##### Valence #####
rajboudDataClean = rjboudData[c("ID", "pic_name", "arousal")]
rajboudDataClean$pic_name = as.factor(rajboudDataClean$pic_name)
rajboudDataClean = reshape(rajboudDataClean, idvar = "ID", timevar =
"pic_name", direction = "wide")
rajboudDataCronbachs = rjboudDataClean[, 2:16]
```

2.2.1. Cronbach's Alpha

```
# Calculate Cronbach's alpha using alpha()
alphavar = psych::alpha(rajboudDataCronbachs, check.keys = TRUE)
summary(alphavar)
```

```

##  

## Reliability analysis  

## raw_alpha std.alpha G6(smc) average_r S/N      ase mean sd median_r  

##       0.95      0.95     0.96      0.57  20 0.0075   36 14      0.56

2.2.2. CFA  

names(radboudDataClean)[2:16] = c('Face_01', 'Face_36', 'Face_32', 'Face_61',
'Face_04', 'Face_24', 'Face_02', 'Face_49', 'Face_58', 'Face_46', 'Face_05',
'Face_33', 'Face_57', 'Face_47', 'Face_27')

HS.model <- 'radboud =~ Face_01 + Face_36 + Face_32 + Face_61 + Face_04 +
Face_24 + Face_02 + Face_49 + Face_58 + Face_46 + Face_05 + Face_33 + Face_57
+ Face_47 + Face_27'

```

Fit and visualize

```

## lavaan 0.6-9 ended normally after 17 iterations
##  

##  

##   Estimator                               ML  

##   Optimization method                    NLMINB  

##   Number of model parameters             30  

##  

##  

##   Used                      Total  

##   Number of observations            85          89  

##  

##  

## Model Test User Model:  

##  

##   Test statistic                  222.273  

##   Degrees of freedom                 90  

##   P-value (Chi-square)              0.000  

##  

##  

## Model Test Baseline Model:  

##  

##   Test statistic                  1087.748  

##   Degrees of freedom                 105  

##   P-value                          0.000  

##  

##  

## User Model versus Baseline Model:  

##  

##   Comparative Fit Index (CFI)        0.865  

##   Tucker-Lewis Index (TLI)           0.843  

##  

##  

## Loglikelihood and Information Criteria:  

##  

##   Loglikelihood user model (H0)      -5070.572  

##   Loglikelihood unrestricted model (H1) -4959.436  

##  

##  

##   Akaike (AIC)                     10201.145  

##   Bayesian (BIC)                   10274.424  

##   Sample-size adjusted Bayesian (BIC) 10179.780

```

```

## 
## Root Mean Square Error of Approximation:
## 
##   RMSEA                               0.131
##   90 Percent confidence interval - lower 0.110
##   90 Percent confidence interval - upper 0.153
##   P-value RMSEA <= 0.05                0.000
## 
## Standardized Root Mean Square Residual:
## 
##   SRMR                                0.062
## 
## Parameter Estimates:
## 
##   Standard errors                      Standard
##   Information                           Expected
##   Information saturated (h1) model      Structured
## 
## Latent Variables:
## 
##   radboud =~
##     Face_01        13.568    1.776    7.640    0.000    13.568    0.723
##     Face_36        13.139    1.701    7.724    0.000    13.139    0.729
##     Face_32        14.518    1.659    8.753    0.000    14.518    0.796
##     Face_61        14.030    1.776    7.901    0.000    14.030    0.741
##     Face_04        13.858    1.790    7.743    0.000    13.858    0.730
##     Face_24        13.351    1.706    7.827    0.000    13.351    0.736
##     Face_02        13.987    1.668    8.387    0.000    13.987    0.773
##     Face_49        12.272    1.577    7.780    0.000    12.272    0.733
##     Face_58        13.383    1.589    8.420    0.000    13.383    0.775
##     Face_46        13.872    1.852    7.490    0.000    13.872    0.713
##     Face_05        13.171    1.561    8.435    0.000    13.171    0.776
##     Face_33        15.258    1.575    9.687    0.000    15.258    0.850
##     Face_57        13.971    1.773    7.882    0.000    13.971    0.740
##     Face_47        14.586    1.535    9.504    0.000    14.586    0.840
##     Face_27        14.357    1.677    8.559    0.000    14.357    0.784
## 
## Variances:
## 
##   .Face_01        Estimate  Std.Err  z-value  P(>|z|)  Std.lv  Std.all
##   .Face_36        167.812   27.096   6.193   0.000   167.812   0.477
##   .Face_32        152.107   24.605   6.182   0.000   152.107   0.468
##   .Face_61        122.170   20.341   6.006   0.000   122.170   0.367
##   .Face_04        161.597   26.245   6.157   0.000   161.597   0.451
##   .Face_24        150.716   24.436   6.168   0.000   150.716   0.458
##   .Face_02        131.867   21.695   6.078   0.000   131.867   0.403
##   .Face_49        129.747   21.014   6.174   0.000   129.747   0.463
##   .Face_58        119.136   19.620   6.072   0.000   119.136   0.399
##   .Face_46        186.223   29.977   6.212   0.000   186.223   0.492

```

##	.Face_05	114.656	18.891	6.069	0.000	114.656	0.398
##	.Face_33	89.525	15.584	5.745	0.000	89.525	0.278
##	.Face_57	161.447	26.209	6.160	0.000	161.447	0.453
##	.Face_47	88.986	15.323	5.807	0.000	88.986	0.295
##	.Face_27	129.495	21.419	6.046	0.000	129.495	0.386
##	radboud	1.000				1.000	1.000

2.2.3. CFA Visualization

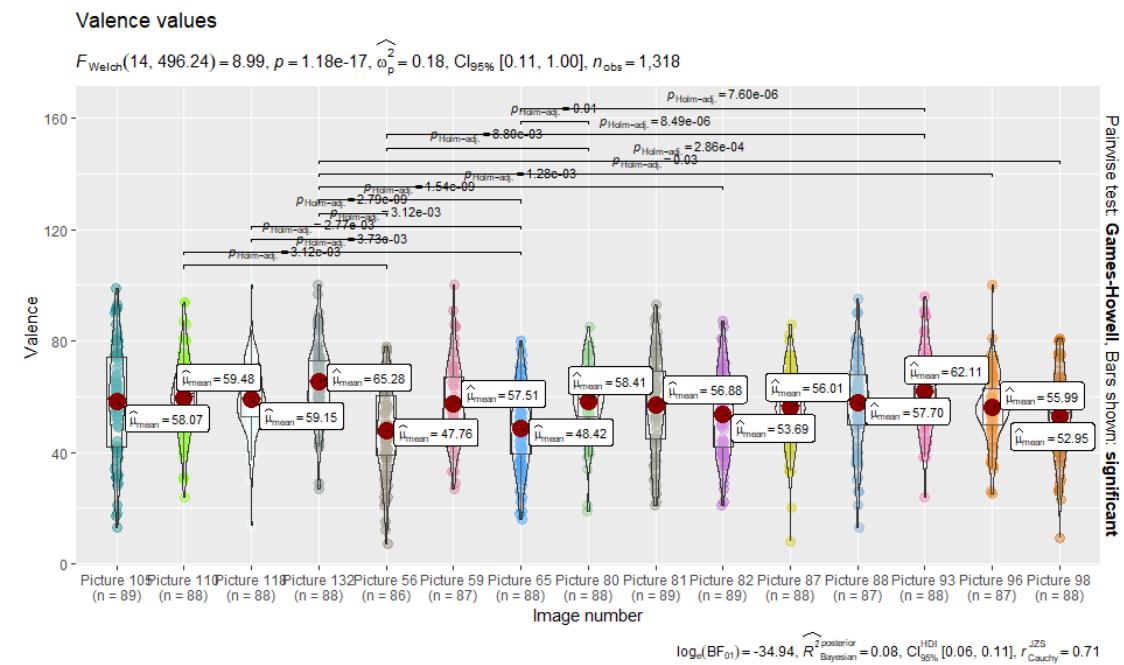
Radboud dataset - Arousal

2.2.4. Distributions

```
# Re-prep data
piscesDataClean = piscesData[c("ID", "pic_name", "valence")]
piscesDataClean$pic_name = as.factor(piscesDataClean$pic_name)
piscesDataClean$ID = as.factor(piscesDataClean$ID)
```

Visualizations

Pisces - Valence



Acoustic and Prosodic Speech Features Reflect Physiological Stress but not Isolated Negative Affect: A Multi-Paradigm Study on Psychosocial Stressors

Mitchel Kappen^{*123}, Gert Vanhollebeke¹²⁵, Jonas Van Der Donckt⁴⁵, Sofie Van Hoecke⁴⁵, Marie-Anne Vanderhasselt¹²

¹Department of Head and Skin, Ghent University, University Hospital Ghent (UZ Ghent), Department of Psychiatry and Medical Psychology, Ghent, Belgium

²Ghent Experimental Psychiatry (GHEP) Lab, Ghent University, Ghent, Belgium

³Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

⁴IDLab, Ghent University - imec, Ghent, Belgium

⁵Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Under review at Scientific Reports. Preprint ref:

Kappen, M., Vanhollebeke, G., Van Der Donckt, J., Van Hoecke, S., & Vanderhasselt, M. (2023, March 29). Acoustic and Prosodic Speech Features Reflect Physiological Stress but not Isolated Negative Affect: A Multi-Paradigm Study on Psychosocial Stressors.

<https://doi.org/10.31234/osf.io/wyd3c>

5.1. Abstract

Heterogeneity in speech under stress has been a recurring issue in stress research, potentially due to varied stress induction paradigms. This study investigated speech features in semi-guided speech following two distinct psychosocial stress paradigms (Cyberball and MIST) and their respective control conditions. Only negative affect increased during Cyberball, while self-reported stress, skin conductance response rate, and negative affect increased during MIST. Fundamental frequency (F0), speech rate, and jitter significantly changed during MIST, but not Cyberball; HNR and shimmer showed no expected changes. The results indicate that observed speech features are robust in semi-guided speech and sensitive to stressors eliciting additional physiological stress responses, not solely decreases in negative affect. These differences between stressors may explain literature heterogeneity. Our findings support the potential of speech as a stress level biomarker, especially when stress elicits physiological reactions, similar to other biomarkers. This highlights its promise as a tool for measuring stress in everyday settings, considering its affordability, non-intrusiveness, and ease of collection. Future research should test these results' robustness and specificity in naturalistic settings, such as freely spoken speech and noisy environments while exploring and validating a broader range of informative speech features in the context of stress.

5.2. Introduction

Stress is a physiological and psychological response to internal or external stimuli that are perceived as threatening to an individual's well-being (Lazarus & Folkman, 1984; McEwen,

2007). Whether it be personal or professional, acute or chronic, stress is a common aspect of modern life that impacts people of all ages and backgrounds. Acute stress is a normal part of the human experience and can be adaptive in the short term by enabling individuals to respond to challenges and adapt to their environment (McEwen, 2007). However, when stress becomes chronic, it can have serious and long-lasting effects on an individual's physical and mental health, such as cardiovascular disease, cognitive impairment, depression, anxiety, and other (mental) health disorders (Slavich, 2016; Yaribeygi et al., 2017). As a result, accurately measuring and regularly monitoring stress levels is crucial for maintaining optimal health and well-being (Crosswell & Lockwood, 2020; Epel et al., 2018).

Given chronic stress's effects on mental and physical health, various methods have been developed for assessing peoples' stress levels including physiological, self-report, and behavioral methods (Allen et al., 2014). While each method has advantages, they also have unresolved limitations, such as cost, validity, intrusiveness, or lack of accuracy in natural settings (Slavich et al., 2019). Consequently, speech has been suggested as a non-intrusive and cost-effective method capable of measuring stress over extended periods. Speech recordings can be obtained from various sources, such as phone calls or meetings, without the need for specialized equipment, making it cost-effective and allowing for data collection in naturalistic settings, which reduces intrusiveness (Giddens et al., 2013; Kappen, Hoorelbeke, et al., 2022; Slavich et al., 2019).

Acoustic changes in speech have been observed in response to (acute) stress (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018). However, whereas most former analyses were done in studies that used either voice actors or read-out-loud speech paradigms, it is important to shift towards evaluating the potential of

freely spoken speech as it would occur in daily life (Van Der Donckt et al., 2023) to not limit the ecological validity of the results. In addition, whereas former studies rarely used validated stress paradigms or failed to validate the stress experience of participants, our recent studies addressed these limitations by utilizing validated stress induction techniques and gathering self-reports. However, they are still limited by 1) the use of one single stress induction paradigm and 2) read-out-loud speech (Kappen, Hoorelbeke, et al., 2022; Kappen, Van Der Donckt, et al., 2022).

In the current study, we focus on a key set of features that, although varying in the frequency of appearance in the literature, have consistently demonstrated their relevance across studies. We refer to these features as acoustic (physical properties of speech) and prosodic (suprasegmental aspects of speech contributing to the overall rhythm, intonation, and stress patterns), with all chosen features belonging to either one or both categories. These include the Fundamental Frequency (F0), a measure of the vocal cord's vibration frequency, that generally increases with stress (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018), jitter (vocal frequency variation) and shimmer (vocal intensity variation) which have been observed to decrease due to stress (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018), and the Harmonics-to-Noise Ratio (HNR; relative amount of noise in comparison to harmonics in the voice) which has been shown to decrease in the context of a physical stressor (any physical event or stimulus that elicits stress) and has mixed results in the context of psychological stress (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Kappen, Van Der Donckt, et al., 2022; Mendoza & Carballo, 1998). Additionally, we will investigate the effect of stress on changes in speech rate

(talking speed) which has been shown to increase during stress in free speech samples (Giddens et al., 2010, 2013; Rothkrantz et al., 2004).

While previous studies have identified links between specific (acoustic) features of speech and stress, more research is needed to fully understand the current heterogeneity, robustness, and sensitivity of these relationships (Giddens et al., 2013; Kappen, Hoorelbeke, et al., 2022; Van Puyvelde et al., 2018). This can only be done if we do not limit our studies to single stress paradigms, especially considering that different stressors used in these paradigms elicit different stress responses. Therefore, exposing participants to different stress paradigms, but with similar experimental setups (i.e., active control task vs stress task), will allow us to better understand the basis of the observed effects on speech under stress. We aim to understand whether the observed changes in speech features that occur are related to one's changes in mood (e.g., increased negative affect) or to physiological reactions (by activation of the hypothalamic-pituitary-adrenocortical (HPA) axis), by using two well-established stress induction paradigms that specifically elicit these changes. We employed the Cyberball (Williams et al., 2000) and the Montreal Imaging Stress Task (MIST) (Dedovic et al., 2005) paradigms to address these limitations and further uncover the sensitivity and robustness of speech features under stress. Both paradigms use a psychosocial stressor, include an active control condition, and unscripted speech will be collected by having participants describe screenshots from the paradigm. See Van Der Donckt and colleagues for a thorough comparison of speech styles and considerations in speech collection paradigms (2023).

The main difference between these two paradigms is that the Cyberball induces stress in the form of feelings of negative mood by means of *ostracism* due to excluding the participant from the task (Williams, 2007; Williams et al., 2000), whereas the MIST induces stress by

adding components of *social evaluative threat* (SET) to a cognitively challenging task (Dedovic et al., 2005). Ostracism has been shown to worsen one's mood but is mostly limited to psychological responses and does not show a neuroendocrine (cortisol) response (Helpman et al., 2017; Williams, 2007; Zwolinski, 2012), whereas SET elicits a strong physiological, neuroendocrine (cortisol) response in addition to a decreased mood (Allen et al., 2014; Bosch et al., 2009; Dickerson, 2008; Dickerson & Kemeny, 2004).

5.2.1. Research Objectives & Hypotheses

We will gauge the stress response based on increased skin conductance response rate (SCRR), as well as self-reports on increased experienced stress and negative affect during the stress block as compared to the control block. Moreover, this is the first study to use a picture-describe paradigm to capture semi-guided speech that closely resembles natural speech in order to yield ecologically valid results. For more details, see Van Der Donckt and colleagues (2023). *Negative Affect.* We expect increases in Negative Affect after the stress blocks compared to the control blocks for both paradigms. *Self-reported Stress.* We expect increased self-reported stress during the stress block for the MIST. However, we do not expect increases in self-reported stress for the Cyberball, as its effect is inconsistent and strongly mediated by traits such as the need to belong, limiting the observance of this effect in a general population (Beekman et al., 2016). *Skin Conductance Response Rate (SCRR).* We expect an increase in SCRR during the stress block for the MIST, but not for the Cyberball. *Speech features.* We expect similar results for speech features as observed in earlier studies that used read-out-loud protocols (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018) for the MIST. That is, increases in Fundamental Frequency (F0)

(Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018) and measures of changes in speech rate (Giddens et al., 2013; Rothkrantz et al., 2004), decreases in Jitter (Giddens et al., 2013; Van Puyvelde et al., 2018) and Shimmer (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018), and changes in Harmonics to Noise (HNR; added noise in the voice), but the direction is unclear due to mixed results in the context of psychological stressors (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Kappen, Van Der Donckt, et al., 2022; Mendoza & Carballo, 1998). In the Cyberball paradigm, the occurrence and direction of significant speech feature changes will reveal the sensitivity and heterogeneity of speech as a biomarker for (psychological) stress. Considering the expected difference in the stress reaction in the Cyberball (negative mood) compared to the MIST (negative mood + physiological reaction), the occurrence of significant changes in speech features would show that speech is responsive to mere changes in mood due to stress (therefore occurring in both paradigms). A lack of changes in speech features would, however, illustrate that speech (features) are merely related to physiological stress responses and therefore follow the patterns (i.e., effects in MIST, but not in Cyberball) observed in other biomarkers such as cortisol (Allen et al., 2014; Dickerson, 2008; Dickerson & Kemeny, 2004). Lastly, it is possible that different speech features have varying sensitivity, where some might just be responsive to combined mood and physiological stress responses, and others might be responsive to mere changes in mood. For example, one of the most homogeneously reported speech features to change under stress is F0, which could indicate that this is a sensitive feature to any change in experienced stress (i.e., mood or physiological) since it occurs in many different studies and stress paradigms. Other features have shown to be more heterogeneous (e.g., HNR, Jitter), which could be explained by the use of different paradigms and stressors.

The combination of two different stress induction paradigms, which elicit different stress responses by calling on different psychological constructs (Cyberball; ostracism, MIST; social evaluative threat), tested in the same group of participants, will give unique insights into the robustness (by using semi-guided speech), sensitivity (by comparing a mood only to a mood plus physiological stress reaction), and up to this point heterogeneity (by comparing two commonly used stress paradigms) in a variety of speech features under stress.

5.3. Methods

5.3.1. Participants

A convenience sample of 66 healthy subjects (13 women, 53 men, age $M = 21.29$, $SD = 2.82$) was recruited through social media. Upon registration, participants were checked for exclusion criteria (see supplemental material). The study was conducted in accordance with the declaration of Helsinki and received ethical approval from the Ghent University hospital ethical committee (registration number: B6702020000676). Another part of the study investigates the effects of (psychosocial) stressors on neural correlates. Results of electrophysiological correlates will be published elsewhere. Other collected data that were not part of the current paper's research objectives will only be described in the supplemental materials.

All participants gave written informed consent before participating and were debriefed afterward on the true purpose of the study. A 40 Euro compensation fee was awarded upon completion of both testing days through bank transfer.

5.3.1. Procedure

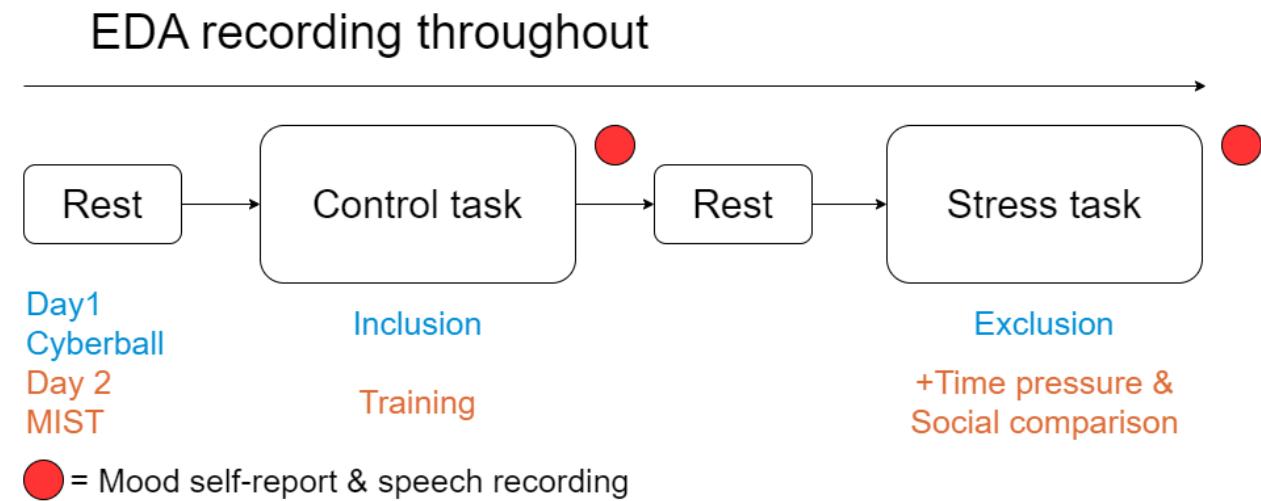
5.3.1.1. On-site experimental session

Participants completed online informed consent and trait questionnaires (beyond this paper's scope) prior to two in-person experimental sessions, which were conducted in a dedicated room in the Department of Adult Psychiatry at Ghent University Hospital. At the start of the first on-site session, participants signed a paper consent form, and experimenters reviewed the cover story (see Cover Story).

The experiment was designed in OpenSesame version 3.2.8 and was carried out on a dedicated computer (Dell, Windows 10). Participants came in on two different days, at least 7 days apart. The experimental sessions were identically structured, but only the task contents differed (Day1; see [Cyberball](#), Day2; see [MIST](#)). Prior to the task, electrodes (ECG, EDA; see [Physiological Data](#)) and an EEG cap were placed (duration 10-30 minutes; beyond this paper's scope). The experiment started with a 10-minute (5 minutes eyes closed, 5 minutes eyes open) resting block to achieve habituation. After this, the Control task started. After this condition, there was another 10-minute resting block, followed by a Stress task and another 10-minute resting block. Subsequent to each task block (i.e., Control task and Stress task), participants were prompted to do a speech trial (see [Speech Data](#)) and respond to self-report psychological state questionnaires (see [Self-Report Data](#)). See figure 1 for a flowchart figure.

Figure 1

Flowchart of experimental design



Note. The two days follow a similar structure except for the presented paradigm and respective control and stress tasks. EDA (electrodermal activity) is collected throughout the paradigm. Speech recordings (picture description) and self-report questionnaires are collected directly after task execution. The colors used in the figure are congruent with the colors presented in the results section.

5.3.1.2. Cyberball - Day 1

The Cyberball paradigm involved a ball-tossing game in which participants played with two computer-generated confederates (one man, one woman, placement counterbalanced across participants), represented by pictures (from Allaert and colleagues (Allaert et al., 2022)). However, participants were told that the other players were humans participating at other universities. The confederates' behaviors were predefined and the game was visualized with a picture of the participant at the bottom center, while the confederate pictures were placed at the top left and right (Williams et al., 2000). Participants could throw the ball to either confederate by pressing an arrow key (right hand), and the ball's movement took 1500 ms. The confederates held possession between two and three seconds (randomly generated) to increase the credibility that they were human. During the **control task** (inclusion phase),

participants received the ball 33% of the time over 150 throws, while in the **stress task** (exclusion phase), participants were excluded in a probabilistic manner from receiving the ball after an initial normal phase of 30 throws. The chance of retrieving the ball increased with each subsequent throw not directed at the participant, with chances ranging from 0 to 100% (adapted from Williams and colleagues (Williams et al., 2000)).

5.3.1.3. MIST - Day 2

During the Montreal Imaging Stress Task (MIST), participants solved mathematical equations of increasing difficulty (Dedovic et al., 2005). Equations were displayed in black on a white background, and the correct answer was always a number between zero and nine, with participants answering each question using the corresponding number on the keyboard's numpad (right hand). The difficulty scales and equation-generating code was identical to the original study, as supplied by prof. Pruessner (Dedovic et al., 2005).

The **control task** included seven difficulty scales, with participants solving up to ten equations per scale. After each equation, feedback was given in the form of "*Correct!*", "*Incorrect!*", or "*Timeout!*", shown in black. The **stress task** employed the same difficulty scales but introduced changes to the task and feedback. Participants were informed that their performance would be compared to that of a group and that they should perform at least on par with the average. Equations were presented with a shrinking bar indicating the remaining time to solve the equation, and the allowed time was set to be 90% of their average response time during the control task. After every three successive correct or incorrect answers, the allowed time was adjusted by 10% to increase/decrease difficulty. Participants also saw a

performance bar with two arrows indicating their personal and group average scores. Their personal arrow moved in steps of 5% of the bar's length after each equation (incorrect/timeout; left, correct; right), whereas the group average arrow was stationary at 83%. If a participant's performance fell below the average group performance after completing five difficulty scales, the experimenter would inform them that their data might not be usable and urge them to improve.

5.3.2. Data Collection

In this study, several types of data were collected for analysis. While our main hypotheses focused on specific data modalities, we also collected additional self-report and cardiac data. To ensure transparency, we have provided an overview of these data modalities and a complete study flowchart in the supplemental materials.

5.3.2.1. Speech Data

On both days, after completion of either task (i.e., control/stress task), participants were prompted to describe a picture out loud, see Figure 1. The image was a screenshot of the task they had just completed to avoid introducing noise to our self-report measures by having their minds wander (for a similar approach and considerations, see Van Der Donckt and colleagues (2023). The participants were instructed to describe the images based on what they saw, as well as how it made them feel. See supplemental materials for screenshots.

5.3.2.2. Self-report Data

On both days, after completion of either task (i.e., control/stress task), participants were asked to rate their current levels of stress and negative affect (“Right now, how much do you feel...”) using 6 negative affect (NA) prompts and 1 stress prompt question, each with a 0-100 sliding scale (0 = Not at all - 100 = Very much), see Figure 1. The six negative affect (NA) prompts are: upset, distressed, scared, angry, anxious, and sad, whereas stress was a single item asking “Right now, how much do you feel stressed?”. Positive activating and soothing affect were also collected but not part of the primary hypotheses, thus only described in the supplemental materials. These scales were adopted from Petrocchi and colleagues (Petrocchi et al., 2017). Given the high internal consistency between the prompts in the NA category, these responses were aggregated to compute a single mean score to be used in the analysis.

5.3.2.3. Physiological Data

Both electrocardiography (ECG; see supplemental material for more info) and electrodermal activity (EDA) were collected throughout the paradigms using the VU-AMS ambulatory monitor¹². ECG data was collected using three electrodes, one placed between the right lower two ribs (ground), one placed at the left lateral side of the chest at the height level of the xiphoid process (V+), and one slightly below the right collar bone four to five cm right from the sternum (V-). EDA data was collected by placing two velcro electrodes (with applied isotonic electrode gel; Biopac) on the middle phalanges of the left index- and middle finger. For

¹² VU University Amsterdam, www.vu-ams.nl, Amsterdam, the Netherlands

more technical information surrounding physiological data collection, please refer to supplemental material ([0.2_EDA.ipynb](#) & [scl_processing.py](#)).

5.3.3. Data Analysis

5.3.3.1. Physiological Data

EDA data was preprocessed using specific Python code, which can be found in the supplemental material ([0.2_EDA.ipynb](#) & [scl_processing.py](#)). To prepare the data for analysis, a 2Hz low-pass filter was applied to the raw signal, which was then decomposed into a tonic and phasic component. From the phasic component, the Skin Conductance Response Rate (SCRR) was extracted by identifying peaks with the SciPy toolkit. The thresholds for rise and fall time, as well as peak parameters, were determined based on established guidelines from the literature (Posada-Quintero & Chon, 2020).

5.3.3.2. Extraction of Speech Features

To ensure data quality, we manually checked all recordings whether they were complete, clear, with limited background noise, and no excessive clipping. In addition, recordings were dropped if there was no complementary self-report scales (due to technical issues). Eight recordings were removed, resulting in 120 control recordings (64 Cyberball, 56 MIST) of 66 out of 66 participants, and 119 stress recordings (64 Cyberball, 55 MIST) of 66 out of 66 participants. Prior to feature extraction, we downsampled the speech samples to 16Khz and applied dithering. These steps were performed in order to make the extracted OpenSMILE metrics less sensitive to environmental harmonics at the voiced boundaries (Van Der Donckt et

al., 2023). To extract features from the recordings, we used OpenSmile 2.3.0 (Eyben et al., 2010) with the GeMAPSv01b configuration (Eyben et al., 2015), a widely-used acoustic feature set in voice research and affective computing. From this feature set (feature names as described in GeMAPS added between brackets), we selected Fundamental Frequency (F0; F0semitoneFrom27.5Hz_sma3nz_amean), Jitter (jitterLocal_sma3nz_amean), Shimmer (shimmerLocaldB_sma3nz_amean), Harmonics-to-Noise Ratio (HNR; HNRdBACF_sma3nz_amean), and Voiced Segment Length (MeanVoicedSegmentLengthSec) and Mean Voiced Segments per Second (a proxy for speech speed; VoicedSegmentsPerSec) to capture changes in speech rate. All features were computed using Python 3.9.6 for a sliding window and then mean-aggregated over the whole recording, thus not displaying high temporal changes. For detailed information regarding feature calculation and extraction procedure, we refer the reader to Eyben et al. (Eyben et al., 2010) and Section 6.1 of Eyben et al. (Eyben et al., 2015).

5.3.3.3. Statistical Analysis

Statistical analyses were performed using R4.1.1 (for detailed version information of the software and packages used, please refer to the supplemental materials).

We used the ‘lme4’ (Bates et al., 2014) package to fit linear mixed models (LMMs) to each of the dependent variables. The sum of squares for each model was estimated using a partial sum of squares (Anova type III approach), and the statistical significance level was set to $p < .05$ (these results are only reported in the supplemental materials). Tests for pairwise comparisons of the EMMs (estimated marginal means) were performed with the ‘emmeans’

package (Lenth, 2018). A false discovery rate (FDR) was used to correct for multiple comparisons correction for each data modality (e.g., all speech comparisons pooled together and penalized accordingly) to minimize the risk of Type 1 errors (Benjamini & Hochberg, 1995) using the ‘`p.adjust()`’ function from the ‘`stats`’ package. In the results section, only corrected p-values will be reported. Moreover, effect sizes (Cohen’s D) and their 95% confidence intervals (CI) are estimated with the ‘`eff_size()`’ function from the ‘`emmeans`’ package (Lenth, 2018). Results are only reported using the effect sizes of within-paradigm comparisons to make a comparison between different dependent variables and data modalities. Note that between-paradigm comparisons (i.e., Cyberbal vs MIST) should be avoided for each respective task (i.e., control and stress task) as the performed tasks and the described pictures are inherently different.

To control for the potential effect of gender on the different dependent variables, gender was considered as a fixed effect for each individual model prior to statistical inference. However, to make sure our models were parsimonious, we bottom-up tested whether adding *gender* as an independent variable to the model improved each model’s fit (Bates et al., 2018). For each dependent variable, we compared models that included and excluded *gender*, and it was only included in the model if it showed to be a significant contributor after comparing models with reducing complexity using χ^2 goodness-of-fit tests within the ‘`anova()`’ function (Fox et al., 2012). The statistical significance level was set to $p < .05$ and based on this, gender was included in the models for F0, Shimmer, HNR, Voiced segment length, and self-reported stress. As such, each model followed the following structure; *DependentVariable* ~ *Phase* * *Task* + *Gender* + (1|*ID*) or *DependentVariable* ~ *Phase* * *Task* + (1|*ID*). With *Phase* having 2 levels

(control vs stress), *Task* having 2 levels (Cyberball vs MIST), and participant as a random intercept.

5.4. Results

Throughout the different paradigms, we focus in our analyses on three main modalities of which two are already more validated in literature (i.e., self-reports and physiological measures) and one is our novel addition to the state-of-the-art (i.e., speech). The models reported will contain the dependent variable, taskPhase (control vs stress task), taskType (MIST vs Cyberball), Gender (man vs woman) if showing to be a significant contributor, and '(1|ID)'; a random intercept for each participant. Per category (i.e., self-reports, physiological, and speech), for each feature, we will only describe the pairwise comparisons of stress vs control block for each individual paradigm as these are directly related to our research questions. All effect sizes and corresponding 95% confidence intervals for the control-stress comparisons per feature, per paradigm, are also displayed in Figure 2. Full model information and corresponding statistics are described in the analyses section of the supplemental materials.

5.4.1. Physiological

5.4.1.1. Skin Conductance Response Rate (SCRR)

A significant increase in SCRR was observed during the stress task in the MIST, $b = 1.30$, $SE = .34$, $t = 3.77$, $p < .001$, $d = .72$, 95% CI [.33, 1.10], but not in the Cyberball, $b = .53$, $SE = .32$, $t = 1.66$, $p = .098$, $d = .29$, 95% CI [-.06, .64].

5.4.2. Self-reports

5.4.2.1. Negative Affect

A significant increase in *Negative Affect* was observed during the stress task in the MIST, $b = 4.29$, $SE = 1.69$, $t = 2.54$, $p = .016$, $d = .48$, 95% CI [.11, .86], as well as in the Cyberball, $b = 5.37$, $SE = 1.57$, $t = 3.42$, $p = .002$, $d = .60$, 95% CI [.25, .96].

5.4.2.2. Stress

A significant increase in self-reported *Stress* was observed during the stress task in the MIST, $b = 17.26$, $SE = 3.84$, $t = 4.49$, $p < .001$, $d = .85$, 95% CI [.47, 1.24], but not in the Cyberball, $b = 2.60$, $SE = 3.58$, $t = .73$, $p = .469$, $d = .13$, 95% CI [-.22, .48].

5.4.3. Speech

5.4.3.1. Fundamental Frequency (F0)

A significant increase in *F0* was observed during the stress task in the MIST, $b = .42$, $SE = .15$, $t = 2.76$, $p = .026$, $d = .52$, 95% CI [.15, .90], but not in the Cyberball, $b = .12$, $SE = .14$, $t = .85$, $p = .531$, $d = .15$, 95% CI [-.20, .50].

5.4.3.2. Voiced segments per second (MVSPS)

A significant increase in *voiced segments per second* was observed during the stress task in the MIST, $b = .23$, $SE = .04$, $t = 5.44$, $p < .001$, $d = 1.03$, 95% CI [.65, 1.42], but not in the Cyberball, $b = .02$, $SE = .04$, $t = .52$, $p = .729$, $d = .09$, 95% CI [-.26, .44].

5.4.3.3. Voiced segment length (MVSL)

A significant increase in *voiced segment length* was observed during the stress task in the MIST, $b = .01$, $SE = .005$, $t = 2.62$, $p = .029$, $d = .50$, 95% CI [.12, .88], but not in the Cyberball, $b = -.001$, $SE = .005$, $t = -.27$, $p = .858$, $d = -.05$, 95% CI [-.40, .30].

5.4.3.4. Harmonics-to-noise ratio (HNR)

No significant change in *HNR* was observed during the stress task in the MIST, $b = .21$, $SE = .09$, $t = 2.31$, $p = .053$, $d = .44$, 95% CI [.06, .82], nor in the Cyberball, $b = .10$, $SE = .08$, $t = 1.19$, $p = .354$, $d = .21$, 95% CI [-.14, .56].

5.4.3.5. Shimmer

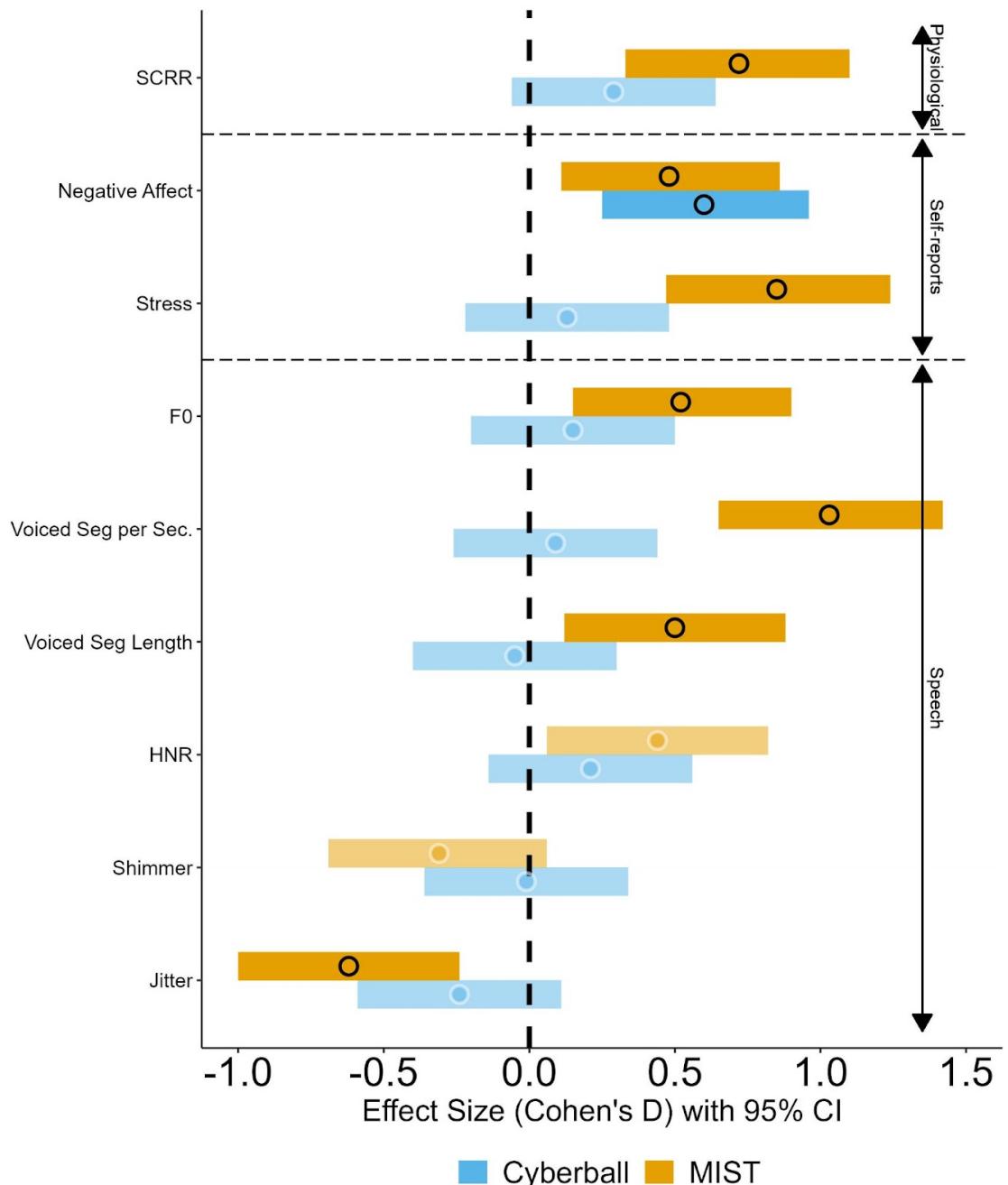
No change in *Shimmer* was observed during the stress task in the MIST, $b = -.02$, $SE = .01$, $t = -1.65$, $p = .201$, $d = -.31$, 95% CI [-.69, .06], nor in the Cyberball, $b = 0$, $SE = .01$, $t = -.07$, $p = .942$, $d = -.01$, 95% CI [-.36, .34].

5.4.3.6. Jitter

A significant decrease in *Jitter* was observed during the stress task in the MIST, $b = -.003$, $SE = .0009$, $t = -3.28$, $p = .007$, $d = -.62$, 95% CI [-1.00, -.24], but not in the Cyberball, $b = -.001$, $SE = .0009$, $t = -1.38$, $p = .292$, $d = -.24$, 95% CI [-.59, .11].

Figure 2

Forest plot of effect sizes and confidence intervals for all control-stress comparisons per paradigm



Note. Effect sizes (dots; Cohen's D) and 95% confidence intervals (bars) for each control vs stress task comparison per stress induction paradigm (Cyberball - MIST). Dependent variables are grouped on their categories (i.e., physiological, self-report, speech) and are FDR corrected within their respective categories. Dots are circled black and ranges are saturated (i.e. non-transparent) if a significant effect is observed after correction.

5.5. Discussion

In this study, we aimed to gain insights into the robustness (by using semi-guided instead of read-out-loud speech), sensitivity (by comparing a mood only to a mood plus physiological stress reaction), and up to this point observed heterogeneity (by comparing two different commonly used stress paradigms) of the effects of stress on acoustic and prosodic speech features. On two different days, participants were exposed to two different stress induction paradigms (i.e., Cyberball and Montreal Imaging Stress Task; MIST) with an expected different stress reaction (i.e., Cyberball; changes in mood, MIST; changes in mood and physiological response). Both paradigms included an active control condition in order to isolate the effects of added stress on their speech. Speech samples were collected directly after each paradigm phase (i.e., control or stress phase) using a picture-describe paradigm (prompting participants to describe a screenshot from the paradigm) to capture semi-guided speech that closely resembles natural speech in order to yield ecologically valid results. For more details, see Van Der Donckt and colleagues (2023).

First, we used validated measures to gauge how the stress responses elicited by the different stress paradigms (Cyberball and MIST), differed. As such, we observed that when considering physiological responses, there was only an increase in skin conductance response rate (SCRR) during the stress phase of the MIST, but not for the Cyberball, which corresponds to our prior hypotheses. In addition to the physiological responses, we also assessed participants' moods using self-reported measures. In line with our expectations, we observed an increase in self-reported negative affect during the stress task of both paradigms. Moreover, participants only reported increased self-reported stress during the MIST. These results are in

line with the literature, as previously mentioned. The Cyberball task affects one's mood due to feelings of ostracization, but it only elicits psychological responses and does not elicit a physiological, neuroendocrine stress response (Helpman et al., 2017; Williams, 2007; Zwolinski, 2012). However, the MIST, using social evaluative threat by means of (negative) social comparison, elicits both physiological and neuroendocrine stress responses, alongside a decrease in mood (Allen et al., 2014; Bosch et al., 2009; Dedovic et al., 2005; Dickerson, 2008; Dickerson & Kemeny, 2004).

Several key acoustic features, described in the literature to be responsive to stress, were extracted from the speech samples (F0; fundamental frequency, HNR; harmonics-to-noise ratio, jitter, shimmer, speech rate, and voiced segment length). Prior results have been heterogeneous, which is possibly due to the use of many different paradigms and stressors which introduce noise rather than robustness in this new modality's early, exploratory stages. We tackle this by doing explicit, side-by-side analysis of two often used stress paradigms in the same sample. In the current study, during the Cyberball task, none of the tested acoustic speech features changed significantly during the stress, compared to the control phase. On the other hand, however, all features except HNR and shimmer changed in the expected direction during the stress phase of the MIST. We observe increases in F0, consistent with literature Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018), speech rate and voiced segment length, consistent with literature (Giddens et al., 2010, 2013; Rothkrantz et al., 2004), and a decrease in Jitter, consistent with literature (Giddens et al., 2013; Van Puyvelde et al., 2018). The observed increase in HNR, related to stress in the MIST, did not survive multiple comparison corrections, which indicates that the observed effect was rather small. This is consistent with the literature, as previous

studies have reported mixed results or conflicting findings in HNR changes (Giddens et al., 2013; Godin et al., 2012; Godin & Hansen, 2015; Mendoza & Carballo, 1998). However, the observed effect in the current study follows the same direction as our former study, which did show a significant increase (Kappen, Van Der Donckt, et al., 2022). This consistency might be indicative of the true direction of the effect, despite the small effect size in the present study. Additionally, no decrease was observed for shimmer during the stress task, whereas mixed results have been observed in the literature (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018). The absence of a significant change in shimmer during the stress phase of the MIST can be related to two things. First, it could be related to the speech collection paradigm used in the current study. We used a semi-guided speech paradigm in which participants were shown a screenshot from the task they just completed and were prompted to describe it. For the MIST, that means participants were shown a mathematical puzzle, similar to the earlier task, which many participants would try and solve out loud. This speech follows a less natural flow than naturally spoken speech, and as such could affect the amount of changes in shimmer. Second, which is arguably related to the first, the absolute observed values for shimmer were rather low as compared to former studies (Kappen, Van Der Donckt, et al., 2022) indicating potential floor effects. However, it should be noted that the observed absolute shimmer values are consistently lower for the recordings in the Cyberball paradigm (see supplemental materials). Nonetheless, due to the presented images being different between the two speech collection paradigms, no formal comparisons between the two should be done, and our interpretations are limited to within-paradigm changes. The methodological choice, to have participants describe screenshots from the task rather than other, off-topic images, should also be noted as the study's biggest limitation. The study's

objective was, first and foremost, to elicit stress using two different psychosocial stress induction paradigms. By having participants describe unrelated images directly after completing the task, it can be argued that they could be distracted from the stressor and thus decrease its potency, confounding the final results. As such, follow-up studies should collect speech samples using a standardized semi-guided speech paradigm consisting of validated images that are congruent to psychosocial paradigms, to describe and keep consistent throughout longitudinal designs as described in Van der Donckt and colleagues (2023). It should also be noted that the sample consisted of predominantly young adults, therefore possibly limiting the generalization of our results to the general population.

The current study used two different psychosocial stress paradigms that are different in their stress responses (i.e., Cyberball; negative mood, MIST; negative mood + physiological reaction). As such, we were able to relate promising acoustic and prosodic speech features to these distinct stress responses. We demonstrate that most features that are described in the literature in relation to stress only changed in the MIST (social evaluative threat paradigm), and not in the Cyberball (ostracism paradigm). These results follow our observed changes in self-report and physiological measures and as such, we conclude that speech as a biomarker is indeed a promising method for detecting changes in stress levels. Speech is comparable to other validated methods (i.e., skin conductance response rate & self-reported stress) as illustrated by the observed effect sizes, in that it does not respond to mere changes in negative affect, but only when physiological changes occur.

This study is the first to demonstrate how speech features change due to different stress paradigms and corresponding stressors, using a within-participant design. These results might explain the current heterogeneity in the literature with regard to these speech features.

We conclude that semi-freely spoken speech (features) are promising for stress detection, and are not affected by stressors that only evoke changes in negative mood. This further outlines its potential in real-world applications, where it appears increasingly promising in passive, remote, non-intrusive tracking of major stressors in daily life that can have severe health implications (Kappen et al., 2023).

5.6. Conclusions

To conclude, we collected repeated semi-guided speech fragments from participants in two different psychosocial stress paradigms, both including active control conditions. We observed distinct stress reactions in the two paradigms through self-reports and psychophysiological responses. A change in self-reported negative affect during the Cyberball, and an additional physiological and self-reported stress reaction during the MIST were found. Similar effects (i.e., effect during MIST, but not during Cyberball) were found for most speech features of interest; F0, voiced segments per second, mean voiced segment length, and jitter, but not for HNR and shimmer. Therefore, we conclude that these effects are robust in (semi-)freely spoken speech (as compared to earlier studies using read-out-loud speech), and are sensitive to stressors that activate the HPA axis, but not to changes in negative affect alone. The difference in observed effects between the two stressors possibly explains the current heterogeneity in the literature. These results further solidify the potential use of speech as a biomarker for stress level assessment in everyday settings, given its affordability, non-intrusiveness, and ease of collection. Future studies should focus on further testing the robustness of these results in increasingly naturalistic settings, such as completely freely

spoken speech and noisy environments while exploring a broader range of speech features that can be informative in the context of stress.

5.7. References

- Allaert, J., De Raedt, R., Sanchez-Lopez, A., Baeken, C., & Vanderhasselt, M.-A. (2022). Mind the social feedback: Effects of tDCS applied to the left DLPFC on psychophysiological responses during the anticipation and reception of social evaluations. *Social Cognitive and Affective Neuroscience*, 17(1), 131–141. <https://doi.org/10.1093/scan/nsaa066>
- Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., & Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test. *Neuroscience & Biobehavioral Reviews*, 38, 94–124.
<https://doi.org/10.1016/j.neubiorev.2013.11.005>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious Mixed Models* (arXiv:1506.04967). arXiv. <http://arxiv.org/abs/1506.04967>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Beekman, J. B., Stock, M. L., & Marcus, T. (2016). Need to Belong, Not Rejection Sensitivity, Moderates Cortisol Response, Self-Reported Stress, and Negative Affect Following Social Exclusion. *The Journal of Social Psychology*, 156(2), 131–138.
<https://doi.org/10.1080/00224545.2015.1071767>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bosch, J. A., de Geus, E. J. C., Carroll, D., Goedhart, A. D., Anane, L. A., van Zanten, J. J. V., Helmerhorst, E. J., & Edwards, K. M. (2009). A general enhancement of autonomic and

- cortisol responses during social evaluative threat. *Psychosomatic Medicine*, 71(8), 877–885. <https://doi.org/10.1097/PSY.0b013e3181baef05>
- Crosswell, A. D., & Lockwood, K. G. (2020). Best practices for stress measurement: How to measure psychological stress in health research. *Health Psychology Open*, 7(2), 2055102920933072. <https://doi.org/10.1177/2055102920933072>
- Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., & Pruessner, J. C. (2005). The Montreal Imaging Stress Task: Using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience*, 30(5), 319–325.
- Dickerson, S. S. (2008). Emotional and Physiological Responses to Social-Evaluative Threat. *Social and Personality Psychology Compass*, 2(3), 1362–1378.
<https://doi.org/10.1111/j.1751-9004.2008.00095.x>
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), 355–391. <https://doi.org/10.1037/0033-2909.130.3.355>
- Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, 49(December 2017), 146–169.
<https://doi.org/10.1016/j.yfrne.2018.03.001>
- Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

<https://doi.org/10.1109/TAFFC.2015.2457417>

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE - The Munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 1459–1462. <https://doi.org/10.1145/1873951.1874246>

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ..., & Heiberger, R. (2012). Package ‘car.’ *Vienna: R Foundation for Statistical Computing*.

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.e21-390.e29.

<https://doi.org/10.1016/j.jvoice.2012.12.010>

Giddens, C. L., Barron, K. W., Clark, K. F., & Warde, W. D. (2010). Beta-Adrenergic Blockade and Voice: A Double-Blind, Placebo-Controlled Trial. *Journal of Voice*, 24(4), 477–489.

<https://doi.org/10.1016/j.jvoice.2008.12.002>

Godin, K. W., & Hansen, J. H. L. (2015). Physical task stress and speaker variability in voice quality. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015(1).

<https://doi.org/10.1186/s13636-015-0072-7>

Godin, K. W., Hasan, T., & Hansen, J. H. L. (2012). Glottal waveform analysis of physical task stress speech. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2(January 2012), 1646–1649.

Helpman, L., Penso, J., Zagoory-Sharon, O., Feldman, R., & Gilboa-Schechtman, E. (2017). Endocrine and emotional response to exclusion among women and men; cortisol, salivary alpha amylase, and mood. *Anxiety, Stress, & Coping*, 30(3), 253–263.

<https://doi.org/10.1080/10615806.2016.1269323>

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech

as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54(2), 910–921. <https://doi.org/10.3758/s13428-021-01670-x>

Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van Hoecke, S., & Vanderhasselt, M.-A. (2022). Acoustic speech features in social comparison: How stress impacts the way you sound. *Scientific Reports*, 12(1), Article 1. <https://doi.org/10.1038/s41598-022-26375-9>

Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. (2023). Speech as a promising biosignal in precision psychiatry. *Neuroscience & Biobehavioral Reviews*, 148, 105121. <https://doi.org/10.1016/j.neubiorev.2023.105121>

Lazarus, R. S., & Folkman, S. (1984). Stress, appraisal, and coping. In *Spirit and Capital in an Age of Inequality*. Springer publishing company.

<https://doi.org/10.4324/9781315413532>

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*.

McEwen, B. S. (2007). Physiology and Neurobiology of Stress and Adaptation: Central Role of the Brain. *Physiological Reviews*, 87(3), 873–904.

<https://doi.org/10.1152/physrev.00041.2006>

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12(3), 263–273.

[https://doi.org/10.1016/S0892-1997\(98\)80017-9](https://doi.org/10.1016/S0892-1997(98)80017-9)

Petrocchi, N., Piccirillo, G., Fiorucci, C., Moscucci, F., Di Iorio, C., Mastropietri, F., Parrotta, I., Pascucci, M., Magrì, D., & Ottaviani, C. (2017). Transcranial direct current stimulation enhances soothing positive affect and vagal tone. *Neuropsychologia*, 96, 256–261. <https://doi.org/10.1016/j.neuropsychologia.2017.01.028>

- Posada-Quintero, H. F., & Chon, K. H. (2020). Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review. *Sensors*, 20(2), Article 2. <https://doi.org/10.3390/s20020479>
- Rothkrantz, L. J. M., Wiggers, P., Van Wees, J. W. A., & Van Vark, R. J. (2004). Voice stress analysis. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3206, 449–456. <https://doi.org/10.4135/9781452229300.n1969>
- Slavich, G. M. (2016). Life Stress and Health: A Review of Conceptual Issues and Recent Findings. *Teaching of Psychology*, 43(4), 346–355. <https://doi.org/10.1177/0098628316662768>
- Slavich, G. M., Taylor, S., Picard, R. W., Slavich, G. M., Taylor, S., & Stress, R. W. P. (2019). *Stress measurement using speech: Recent advancements , validation issues , and ethical and privacy considerations*. 3890. <https://doi.org/10.1080/10253890.2019.1584180>
- Van Der Donckt, J., Kappen, M., Degraeve, V., Demuynck, K., Vanderhasselt, M.-A., & Hoecke, S. V. (2023). *Ecologically Valid Speech Collection in Behavioral Research: The Ghent Semi-spontaneous Speech Paradigm (GSSP)*. PsyArXiv. <https://doi.org/10.31234/osf.io/e2qxw>
- Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>
- Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, 58(1), 425–452. <https://doi.org/10.1146/annurev.psych.58.110405.085641>
- Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored

over the Internet. *Journal of Personality and Social Psychology*, 79(5), 748–762.

<https://doi.org/10.1037/0022-3514.79.5.748>

Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T. P., & Sahebkar, A. (2017). The impact of stress on body function: A review. *EXCLI Journal*, 16, 1057–1072.

<https://doi.org/10.17179/excli2017-480>

Zwolinski, J. (2012). Psychological and Neuroendocrine Reactivity to Ostracism. *Aggressive Behavior*, 38(2), 108–125. <https://doi.org/10.1002/ab.21411>

5.8. Supplemental Materials

5.8.1. Acknowledgments

We thank Ingemarie Coquyt for their help in collecting the data.

5.8.2. Funding

This research was supported by a grant for research at Ghent University (BOFSTA2017002501), a grant from the King Baudouin Foundation (KBS 2018-J1130650-209563), and the imec.AAA Context-aware health monitoring project. Jonas Van Der Donckt is funded by a doctoral fellowship of the Research Foundation – Flanders (FWO 1S56322N).

5.8.3. Supplemental Materials

All data, corresponding code, and supplementary information are available at <https://osf.io/qf6ck/> and https://github.com/mitchelkappen/stress_cyberball-mist.

Contents:

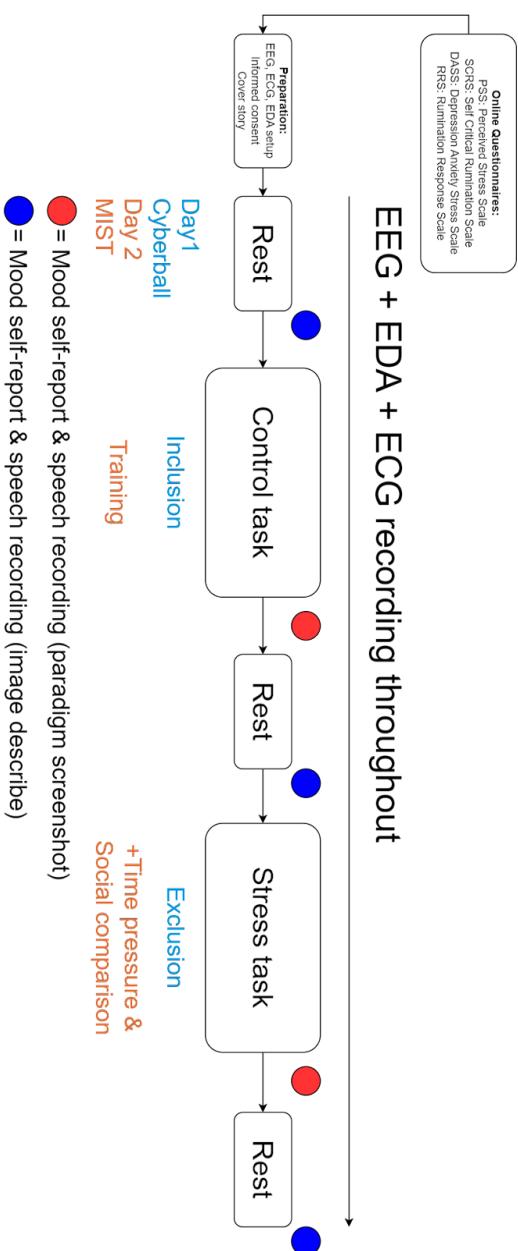
- 1) Exclusion Criteria
- 2) Complete study flowchart
- 3) Speech data collection screenshots
- 4) Self-reports
 - 4.1) Positive activating affect
 - 4.2) Positive soothing affect
- 5) ECG/HRV data
- 6) Software and packages used
 - 6.1) R
 - 6.2) Python
- 7) Full models & Anova results
 - 7.1) Skin Conductance Response Rate (SCRR)
 - 7.2) Negative Affect
 - 7.3) Self-Reported Stress
 - 7.4) Fundamental Frequency (F0)
 - 7.5) Voiced segments per second
 - 7.6) Voiced segment length
 - 7.7) Harmonics-to-noise ratio (HNR)
 - 7.8) Shimmer
 - 7.9) Jitter

5.8.3.1. Exclusion Criteria

Participants were not allowed to enroll in the study if they met any of the following criteria. Some of the criteria were directly related to the collection of EEG:

- Left-handed
- Born before 1977
- Born after 2003
- Personal or family history of epilepsy
- Recent neurosurgical procedures
- Pacemaker or other electronic implants
- Inner ear prosthesis
- Metal objects or magnetic objects in the brain or around the head (only removable earrings & piercings are allowed)
- Pregnancy
- A current depressive episode
- Other psychiatric disorders
- Skin condition on the head
- Current addiction
- Current substance abuse
- Current use of psychotropic medication
- Eye disease(s)
- Heart, respiratory, or neurological problems
- Participated in the EEG study "predicting future success"
- Psychology students
- Dreadlocks
- Tightly curled hair

5.8.3.2. Complete study flowchart



5.8.3.3. Speech data collection screenshots

All speech trials were preceded by the following screen:

Nu zal u opnieuw een afbeelding beschrijven. De afbeelding zal een weergave zijn van de taak die u zojuist hebt uitgevoerd.
Omschrijf hierbij alles wat u ziet, wat er in u opkomt en hoe u zich erbij voelt of voelde.

De afbeelding verschijnt wanneer u op de spatiebalk drukt. Dit zal er tevens ook voor zorgen dat de audio opname begint.
Na op de spatiebalk gedrukt te hebben begint u te omschrijven wat u ziet.

Maak u niet te veel druk als u vast loopt, probeer het natuurlijk te doen alsof u de afbeelding omschrijft aan iemand die
de afbeelding niet kan zien. Het doel is om minimaal 60 seconden per afbeelding te omschrijven.
De afbeelding zal kort flitsen wanneer er 60 seconden gepasseerd zijn, zodat u weet dat u mag afronden.

Als u klaar bent met uw omschrijving drukt u opnieuw op de spatiebalk en gaat u door naar het volgende scherm.

Druk op de spatiebalk om de afbeelding te zien en te starten met hem luidop te beschrijven."

Translation in English:

"Now you will describe an image again. The image will be a representation of the task you just performed. Describe everything you see, what comes to mind, and how you feel or felt about it.

The image appears when you press the spacebar. This will also start the audio recording. After pressing the spacebar, start describing what you see.

Don't worry too much if you get stuck, try to do it naturally as if you are describing the image to someone who cannot see it. The goal is to describe for at least 60 seconds per image. The image will briefly flash when 60 seconds have passed so that you know you can finish up.

When you have finished your description, press the spacebar again and proceed to the next screen.

Press the spacebar to see the image and start describing it aloud."

All screenshot description trials looked identical per participant. After describing for 60 seconds, a prompt appeared saying "You can now press the space bar to stop the audio recording".

Cyberball

Inclusion:



Exclusion:



MIST

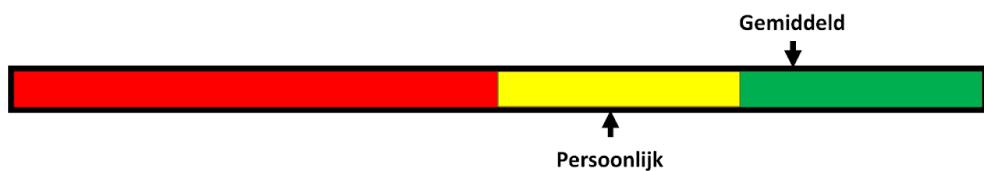
Control:

$$(72/2) - 27 - 4 = ?$$

Stress:



$$(4/2) \times (5/2) \times (80/16) - 19 = ?$$



5.8.3.4. Self-reports

5.8.3.4.1. Positive activating affect

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

Response: VAS_PAA

	Chisq	Df	Pr(>Chisq)
(Intercept)	318.0398	1	< 2.2e-16 ***
fileNum	12.6142	1	0.0003828 ***
taskType	1.7137	1	0.1905012
fileNum:taskType	12.5367	1	0.0003990 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Emmeans Contrasts:

taskType = Cyberball:

contrast	estimate	SE	df	t.ratio	p.value
Control Task - Stress Task	11.672	2.24	170	5.207	<.0001

taskType = MIST:

contrast	estimate	SE	df	t.ratio	p.value
Control Task - Stress Task	0.018	2.41	170	0.007	0.9941

Degrees-of-freedom method: kenward-roger

Effect Sizes:

taskType = Cyberball:

contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	0.92052	0.182	170	0.562	1.279

taskType = MIST:

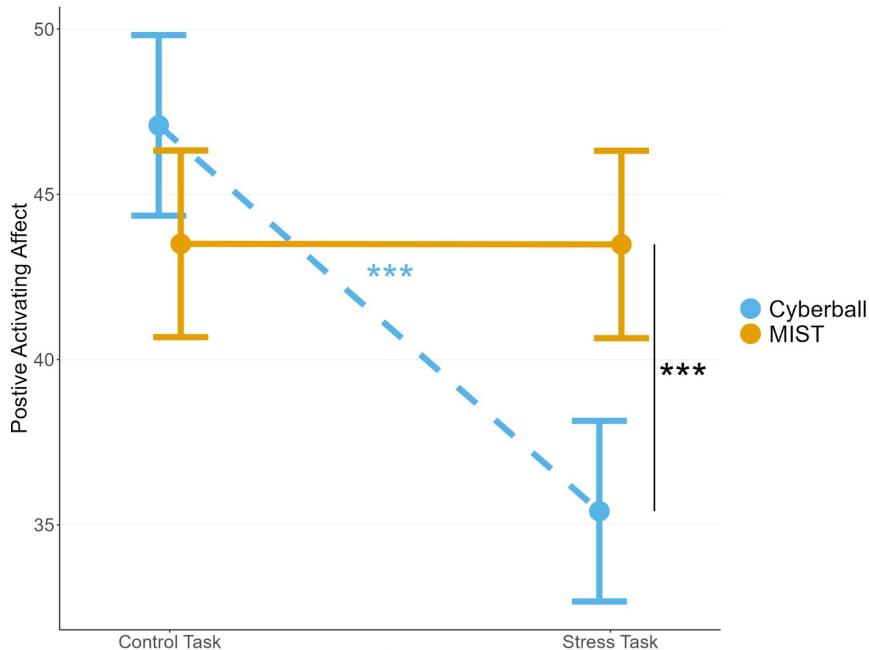
contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	0.00142	0.190	170	-0.374	0.377

sigma used for effect sizes: 12.68

Degrees-of-freedom method: inherited from kenward-roger when re-gridding

Confidence level used: 0.95

Figure:



5.8.3.4.2. Positive soothing affect

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

Response: VAS_PSA

```

Chisq Df Pr(>Chisq)
(Intercept) 925.4036 1 < 2.2e-16 ***
fileNum      35.1251 1 3.092e-09 ***
taskType     73.6405 1 < 2.2e-16 ***
fileNum:taskType 1.5812 1  0.2086
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Emmeans Contrasts:

taskType = Cyberball:

```

contrast      estimate SE df t.ratio p.value
Control Task - Stress Task  9.42 2.75 170  3.428 0.0008
```

taskType = MIST:

```

contrast      estimate SE df t.ratio p.value
Control Task - Stress Task 14.50 2.96 171  4.906 <.0001
```

Degrees-of-freedom method: kenward-roger

Effect Sizes:

taskType = Cyberball:

contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	0.606	0.179	170	0.253	0.959

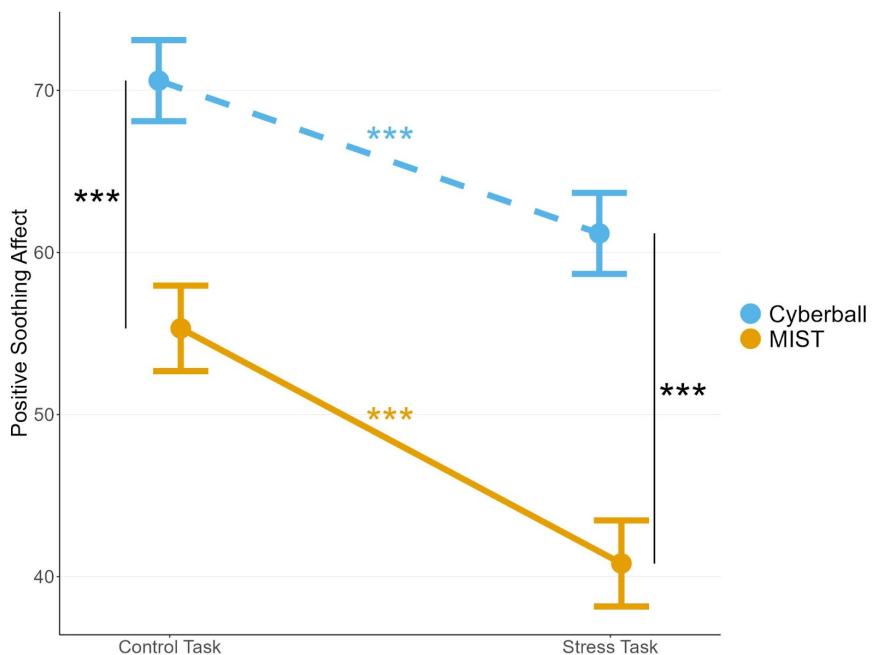
taskType = MIST:

contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	0.932	0.195	171	0.548	1.317

sigma used for effect sizes: 15.55

Degrees-of-freedom method: inherited from kenward-roger when re-gridding

Confidence level used: 0.95

Figure:**5.8.3.5. ECG/HRV data****Anova:**

Analysis of Deviance Table (Type III Wald chisquare tests)

Response: rmssd

	Chisq	Df	Pr(>Chisq)
(Intercept)	181.7315	1	<2e-16 ***
fileNum	2.7029	1	0.1002
taskType	0.9538	1	0.3288
fileNum:taskType	0.0224	1	0.8810

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Emmeans Contrasts:

taskType = Cyberball:

contrast	estimate	SE	df	t.ratio	p.value
Control Task - Stress Task	-3.78	2.90	158	-1.302	0.1947

taskType = MIST:

contrast	estimate	SE	df	t.ratio	p.value
Control Task - Stress Task	-3.15	3.05	158	-1.034	0.3028

Degrees-of-freedom method: kenward-roger

Effect Sizes:

taskType = Cyberball:

contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	-0.241	0.185	158	-0.607	0.125

taskType = MIST:

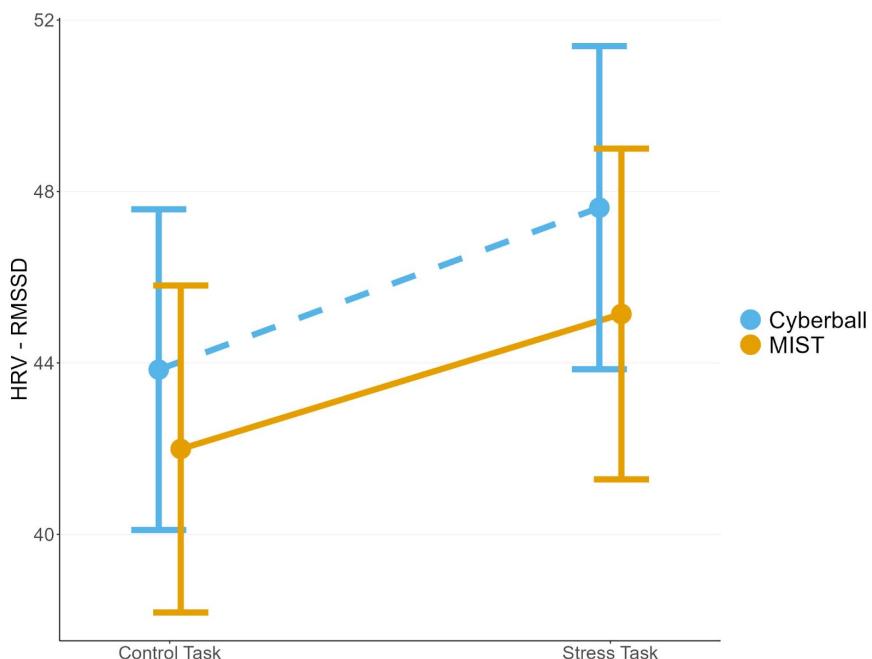
contrast	effect.size	SE	df	lower.CL	upper.CL
(Control Task - Stress Task)	-0.201	0.195	158	-0.585	0.183

sigma used for effect sizes: 15.68

Degrees-of-freedom method: inherited from kenward-roger when re-gridding

Confidence level used: 0.95

Figure:



5.8.3.6. Software and packages used

For full package and version info see:

5.8.3.6.1. R

https://github.com/mitchelkappen/stress_cyberball-mist/blob/main supplemental%20material/ package%20and%20version%20info/Rsession_info.txt

5.8.3.6.2. Python

https://github.com/mitchelkappen/stress_cyberball-mist/blob/main supplemental%20material/ package%20and%20version%20info/poetry.lock

5.8.3.7. Full models & Anova results

Here you will find each model specification with the Anova results. For more details, feel free to run our out-of-the-box code in allAnalysis.R

We display all comparisons, both within-paradigm as between-paradigm for completeness purposes. However, it should be noted that no weight should be given to the between-paradigm comparisons due to the inherent differences in the paradigms and the images that were described.

5.8.3.7.1. Skin Conductance Response Rate (SCRR)

Formula: SCRR ~ fileNum * taskType + (1|participantNum)

Anova:

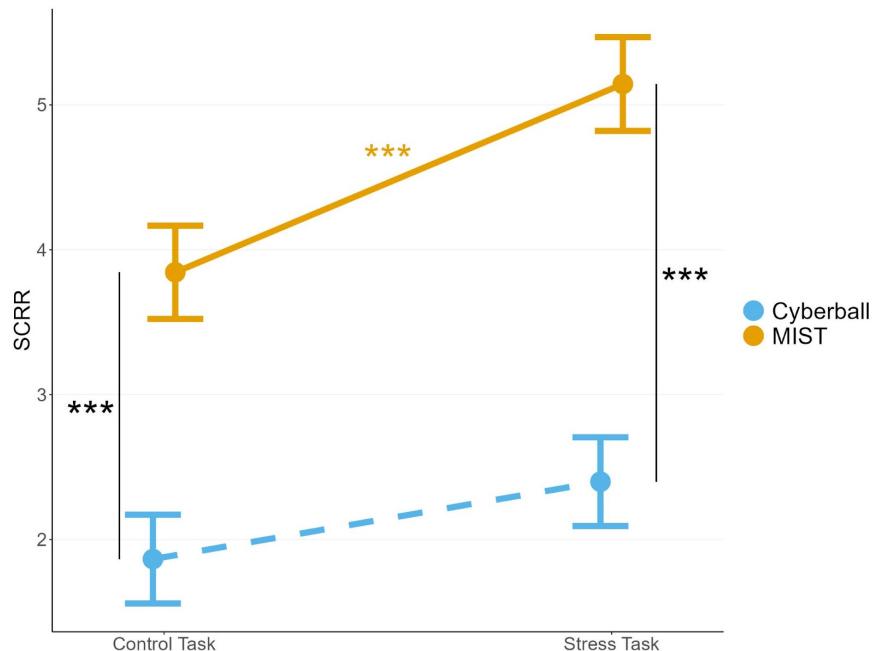
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: SCRR

	Chisq	Df	Pr(>Chisq)
(Intercept)	194.5865	1	< 2.2e-16 ***
fileNum	15.1337	1	0.0001002 ***
taskType	94.3411	1	< 2.2e-16 ***
fileNum:taskType	2.6339	1	0.1046055

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.2. Negative Affect

Formula: VAS_NA ~ fileNum * taskType + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

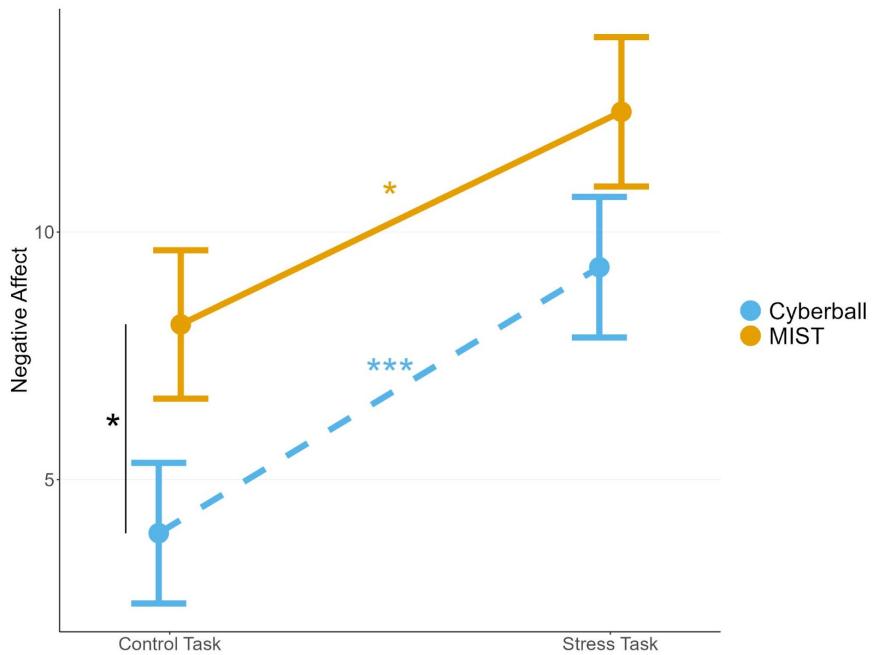
Response: VAS_NA

Chisq Df Pr(>Chisq)

(Intercept)	63.7027	1	1.447e-15 ***
fileNum	17.5548	1	2.792e-05 ***
taskType	9.6001	1	0.001946 **
fileNum:taskType	0.2176	1	0.640868

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.3. Self-Reported Stress

Formula: VAS_Stress ~ fileNum * taskType + Sex + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

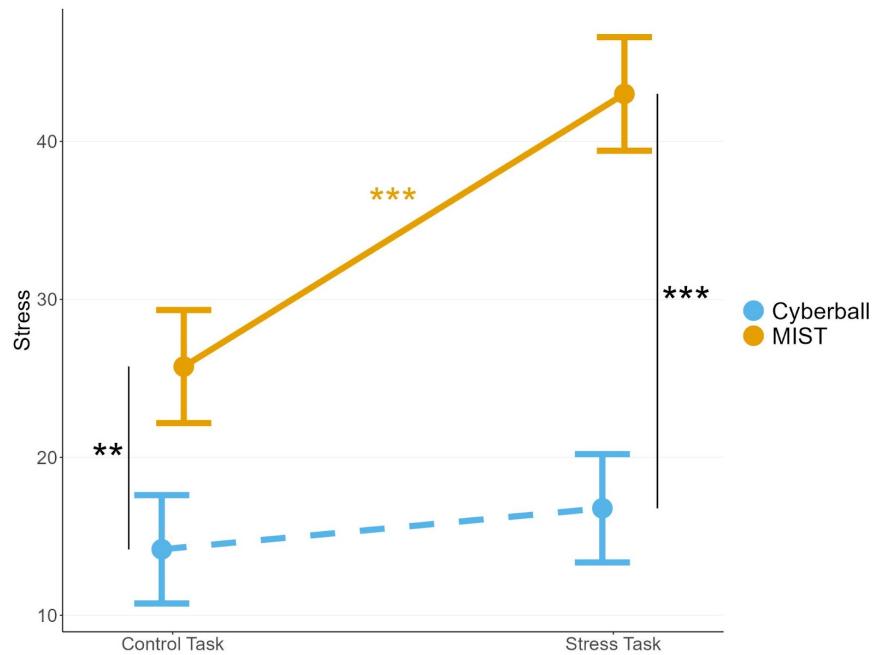
Response: VAS_Stress

Chisq Df Pr(>Chisq)

(Intercept)	88.0639	1	< 2.2e-16 ***
fileNum	14.3127	1	0.0001548 ***
taskType	49.5210	1	1.963e-12 ***
Sex	7.7876	1	0.0052607 **
fileNum:taskType	7.8061	1	0.0052071 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.4. Fundamental Frequency (F0)

Formula: F0semitoneFrom27.5Hz_sma3nz_amean ~ fileNum * taskType + Sex + (1|participantNum)

Anova:

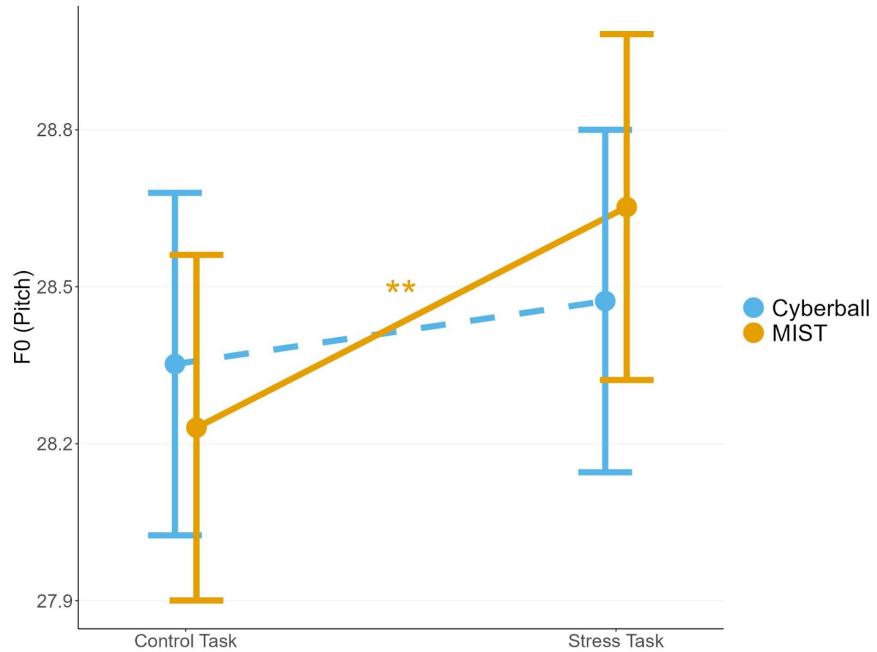
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: F0semitoneFrom27.5Hz_sma3nz_amean

	Chisq	Df	Pr(>Chisq)
(Intercept)	8108.1415	1	< 2.2e-16 ***
fileNum	6.7286	1	0.009488 **
taskType	0.0704	1	0.790737
Sex	196.7378	1	< 2.2e-16 ***
fileNum:taskType	2.0777	1	0.149464

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.5. Voiced segments per second

Formula: VoicedSegmentsPerSec ~ fileNum * taskType + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

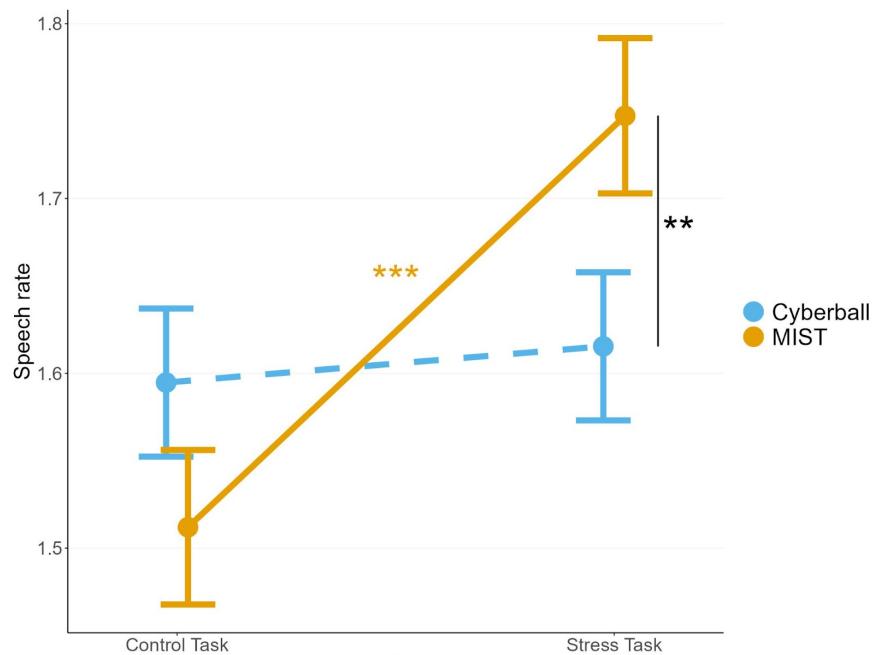
Response: VoicedSegmentsPerSec

Chisq Df Pr(>Chisq)

(Intercept)	2167.2753	1	< 2.2e-16 ***
fileNum	18.7850	1	1.463e-05 ***
taskType	0.6441	1	0.4222239
fileNum:taskType	13.1971	1	0.0002804 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure:



5.8.3.7.6. Voiced segment length

Formula: MeanVoicedSegmentLengthSec ~ fileNum * taskType + Sex + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

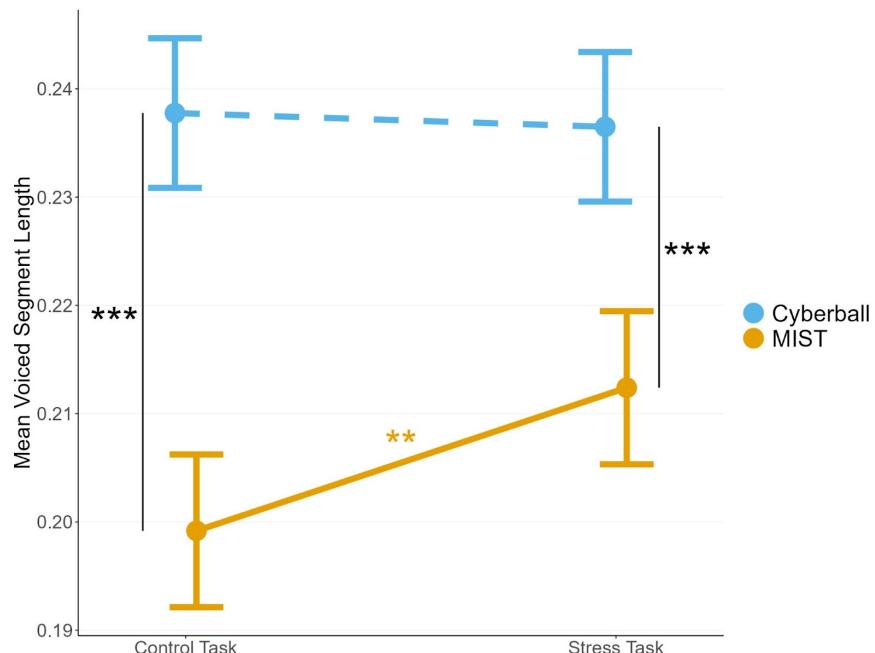
Response: MeanVoicedSegmentLengthSec

Chisq Df Pr(>Chisq)

(Intercept)	1237.4735	1	< 2e-16 ***
fileNum	3.0059	1	0.08296 .
taskType	76.3133	1	< 2e-16 ***
Sex	0.0049	1	0.94408
fileNum:taskType	4.4213	1	0.03549 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.7. Harmonics-to-noise ratio (HNR)

Formula: HNRdBACF_sma3nz_amean ~ fileNum * taskType + Sex + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

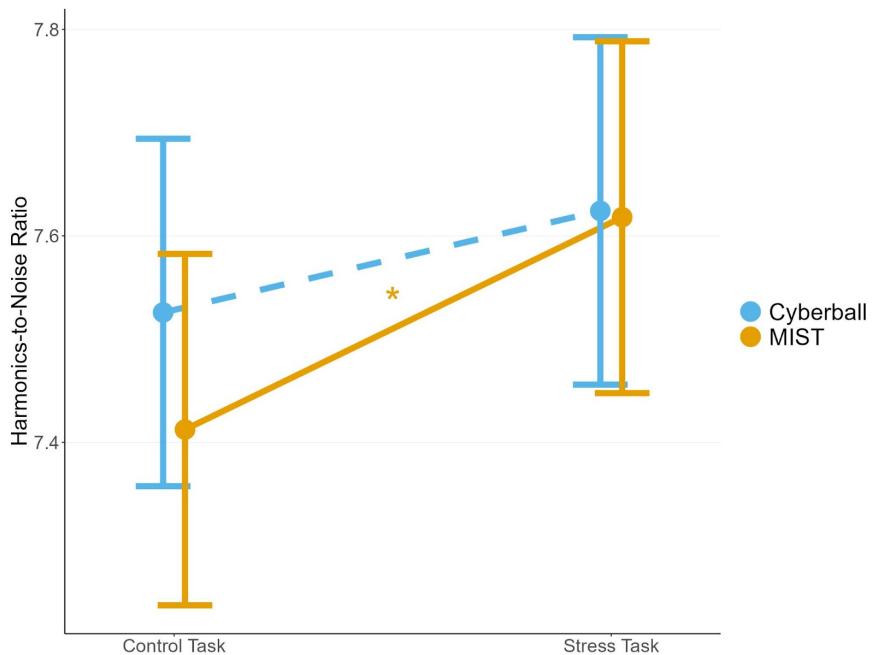
Response: HNRdBACF_sma3nz_amean

Chisq Df Pr(>Chisq)

(Intercept)	2204.9603	1	< 2e-16 ***
fileNum	6.2719	1	0.01227 *
taskType	0.8847	1	0.34691
Sex	127.7845	1	< 2e-16 ***
fileNum:taskType	0.7817	1	0.37663

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.8. Shimmer

Formula: shimmerLocaldB_sma3nz_amean ~ fileNum * taskType + Sex + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

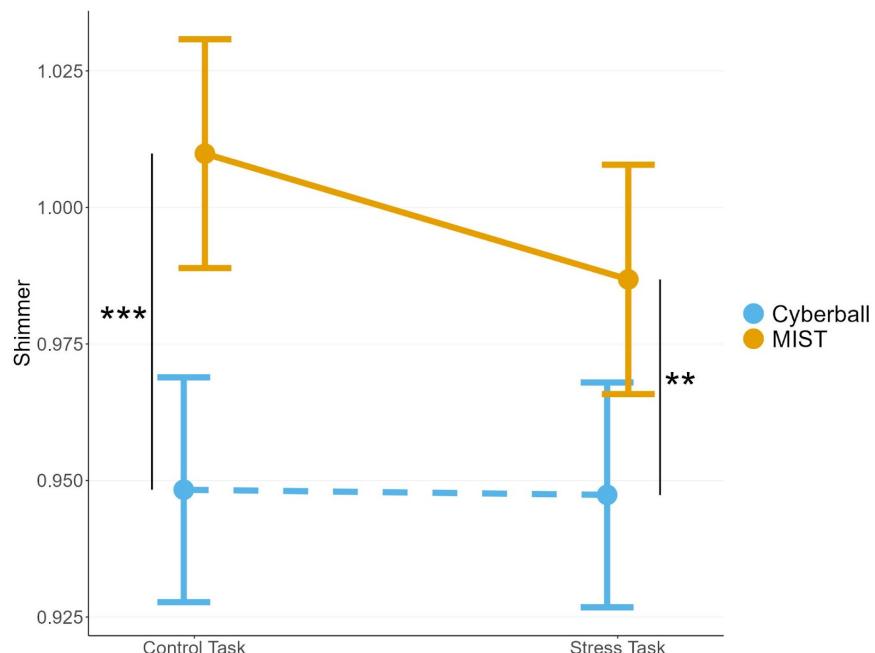
Response: shimmerLocaldB_sma3nz_amean

Chisq Df Pr(>Chisq)

(Intercept)	2617.9744	1	< 2.2e-16 ***
fileNum	1.5846	1	0.2081
taskType	25.8647	1	3.662e-07 ***
Sex	0.2809	1	0.5961
fileNum:taskType	1.3453	1	0.2461

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



5.8.3.7.9. Jitter

Formula: jitterLocal_sma3nz_amean ~ fileNum * taskType + (1|participantNum)

Anova:

Analysis of Deviance Table (Type III Wald chisquare tests)

Response: jitterLocal_sma3nz_amean

Chisq Df Pr(>Chisq)

(Intercept) 1012.8052 1 < 2.2e-16 ***

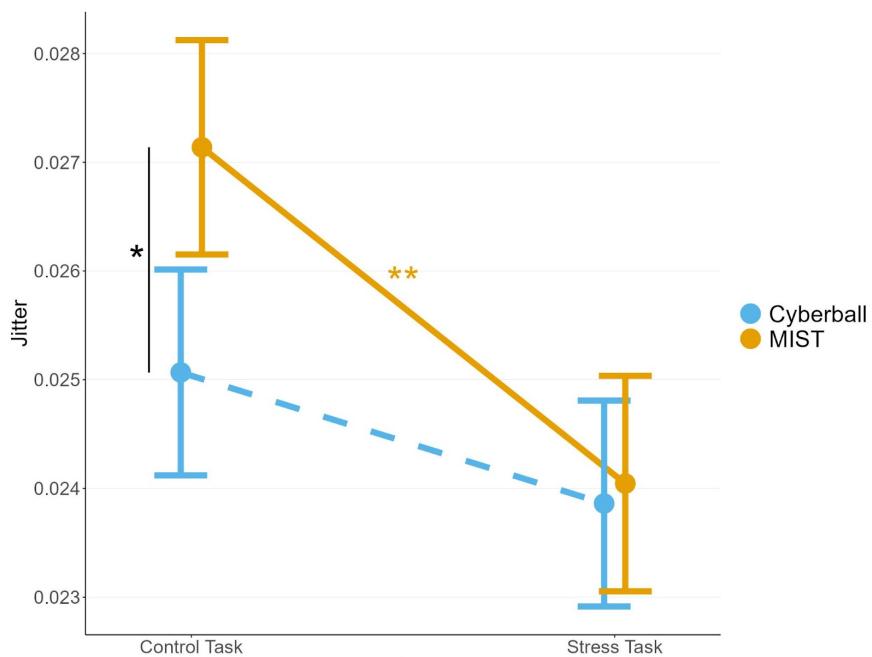
fileNum 11.1673 1 0.0008325 ***

taskType 2.8528 1 0.0912171 .

fileNum:taskType 2.1522 1 0.1423676

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure:



Speech as a Promising Biosignal in Precision Psychiatry

Mitchel Kappen¹, Marie-Anne Vanderhasselt¹, & George M. Slavich²

¹ Department of Head and Skin, Ghent University, University Hospital Ghent (UZ Ghent), Department of Psychiatry and Medical Psychology, Ghent, Belgium

² Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, USA

Published as:

Kappen, M., Vanderhasselt, M. A., & Slavich, G. M. (2023). Speech as a Promising Biosignal in Precision Psychiatry. *Neuroscience & Biobehavioral Reviews*, 105121.

6.1. Abstract

Health research and health care alike are presently based on infrequent assessments that provide an incomplete picture of clinical functioning. Consequently, opportunities to identify and prevent health events before they occur are missed. New health technologies are addressing these critical issues by enabling the continual monitoring of health-related processes using speech. These technologies are a great match for the healthcare environment because they make high-frequency assessments non-invasive and highly scalable. Indeed, existing tools can now extract a wide variety of health-relevant biosignals from smartphones by analyzing a person's voice and speech. These biosignals are linked to health-relevant biological pathways and have shown promise in detecting several disorders, including depression and schizophrenia. However, more research is needed to identify the speech signals that matter most, validate these signals against ground-truth outcomes, and translate these data into biomarkers and just-in-time adaptive interventions. We discuss these issues herein by describing how assessing everyday psychological stress through speech can help both researchers and health care providers monitor the impact that stress has on a wide variety of mental and physical health outcomes, such as self-harm, suicide, substance abuse, depression, and disease recurrence. If done appropriately and securely, speech is a novel digital biosignal that could play a key role in predicting high-priority clinical outcomes and delivering tailored interventions that help people when they need it most.

6.2. Introduction

Many patients experiencing mental health problems today do not receive adequate treatment (Thornicroft et al., 2017; WHO, 2021). Moreover, among those receiving treatment, the modal number of sessions attended for psychotherapy and medication treatment is one, with little follow-up thereafter (e.g., Connolly Gibbons et al., 2011). As a result, patients are not presently being followed in a way that could help detect increases in symptoms or prevent relapse. To address these issues, providers are increasingly using ecological momentary assessments, phone check-ins, and automatic chatbots to monitor patients to monitor symptoms and prevent health emergencies from occurring. These monitoring practices are invasive and burdensome for both patients and providers, though, and they are also subject to self-report biases caused by social desirability, unawareness, and stigma, thus limiting their precision and utility.

Passive data collection bypasses these challenges as it uses technology to monitor patient progress. These technological devices include smartphone pedometers to track physical activity and phone apps to track sleep quality, social activity and engagement, and emotional words typed. Passive data collection modalities that measure health-relevant biological activity, or *biosignals*, are particularly informative as they index the activity of systems that are directly relevant for health and wellbeing. Consequently, they hold great promise for helping both patients and health care providers catch mental health emergencies before they occur. Below, we describe what we believe is one of the most promising health-relevant biosignals to monitor: speech.

6.3. Speech Contains Critical Psychosocial Information

When people talk, their voice and language contain more information than the mere messages they are conveying. Although word selection is important, it can be limited or biased by a lack of awareness or trust, inability to access emotions, shame, and more. In contrast, from the speech signal itself, we can now get metrics of emotion and human functioning that are not subject to these biases (Slavich et al., 2019). Examples include phonetic markers due to physiological changes such as muscle tension and semantic markers due to psychological changes such as increased use of first-person singular pronouns. These markers contain valuable, health-relevant information. Moreover, the collection and analysis of these data come with unique advantages over and above self-reported data. For example, high-quality microphones are now present in all smartphones, making data collection easy, and smartphones can contain apps that make data storage, analysis, and transmission fast, immediate, automated, and secure. Consequently, assessing speech could provide a highly scalable, accessible, and non-invasive strategy for monitoring psychosocial functioning and delivering just-in-time adaptive interventions (JITAs) that help people when they need it most.

Once biosignals are translated into health-relevant indices, they have the potential to become biomarkers of stress and health that can be collected passively and analyzed automatically, and thus help predict the emergence and/or recurrence of high-priority clinical outcomes. In addition to improving human health and resilience, therefore, speech can help

relieve pressure on the healthcare system. To this extent, high-frequency patient monitoring has shown to be promising in predicting self-harm, suicide, depression, bipolar disorder, schizophrenia, alcohol and substance abuse, and more (Colombo et al., 2019; Faurholt-Jepsen et al., 2018; Gee et al., 2020; Mote & Fulford, 2020; Serre et al., 2015). Moreover, JITAls that use smartphones and wearables to provide tailored support to people at just the right time have been found to be useful in managing mental illness, alcohol use, smoking, obesity, and suicide (Nahum-Shani et al., 2018). To date, JITAls have not used speech data despite it being readily available, but making this connection is not difficult. Ultimately, expanding high-frequency patient monitoring and risk detection to include voice recordings could help enable the delivery of life-saving interventions.

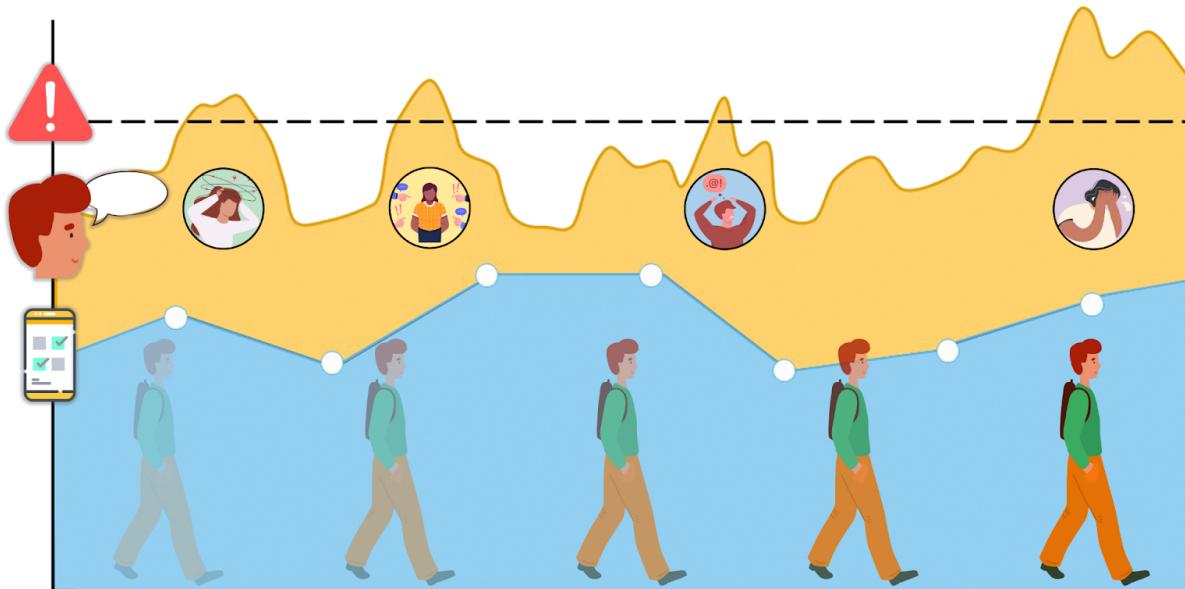


Figure 1. A visual representation of the added value of speech analysis versus ecological momentary assessment (EMA) in tracking stress levels. Although peoples' stress levels change continuously over time, tracking these levels using EMA by phone (blue line) is intermittent, burdensome, and subject to several self-report biases, thus limiting its precision and clinical utility. In contrast, tracking stress and

other health-relevant psychosocial processes using speech analysis (orange line) can be done non-invasively and continuously, and can in turn be used to both detect stressful life events (black circles) and deliver just-in-time adaptive interventions (JITAs) precisely when individuals need them most. This figure has been designed using assets from Freepik.com.

6.4. Stress and Speech

In terms of health-relevant processes to monitor, several exist, but the highest priority is probably stress (Slavich, 2016). Stress is a strong predictor of morbidity across a variety of diseases and is associated with 9 out of 10 leading causes of death in the U.S. today (Bhushan et al., 2020; See Table 1). Measuring stress using speech overcomes several limitations in the assessment of stress, including self-reporting biases and the need for real-time assessment. Moreover, these data can prompt the delivery of evidence-based JITAs for stress, of which there are several (Sarker et al., 2017).

Early research on stress and speech is promising. For example, this work has identified key acoustic features of speech (Van Puyvelde et al., 2018; see also Kappen, Van der Donckt, et al., 2022) and how their associations are modulated by stress (Kappen, Hoorelbeke, et al., 2022). In turn, these biosignals have been found to reliably predict emotion, heart rate, respiration, and cortisol responses (Baird et al., 2021), and peoples' experience of everyday work stressors (Langer et al., 2022; see also Lu et al., 2012). Likewise, a recent systematic review of 127 speech acoustic studies synthesized research describing the use of speech for detecting a variety of psychiatric disorders, including depression, schizophrenia, bipolar disorder, posttraumatic stress disorder, anxiety, anorexia, obsessive-compulsive disorder, and bulimia (Low et al., 2020). Despite promising results with regard to the identification of stress,

limited research has translated this information into JITAs for managing everyday stress, where it could have huge benefits for both patients and caretakers.

Table 1

Role of stress in the top 10 causes of death in the United States.

Disease	Role of stress
Heart disease	Stress causes cardiovascular hyperreactivity and stress-induced hyperactivation of the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system (SNS), both related to coronary heart disease as acute coronary syndromes. Stress causes inflammation, increases in cholesterol, depression, risk of smoking, and metabolic syndrome, all strongly related to heart disease (Bhushan et al., 2020; Bunker et al., 2003; Wirtz & von Känel, 2017).
Cancer	Persistent activation of the HPA axis impairs the immune response and, together with chronic inflammation, heightens the risk of cancer and promotes the spread of cancer after development (Bhushan et al., 2020; Grivennikov et al., 2010; Reiche et al., 2004).
Accidents (unintentional injuries)	Stress can impair cognitive function, reaction time, and decision-making, which can increase the risk of accidents. Moreover, chronic and early life stress leads to more risk seeking behavior, further increasing the risk of accidents.
Chronic lower respiratory diseases	Stress can weaken the immune system and make it more difficult for the body to fight off infections, which can increase the risk of respiratory diseases. The exact associations with stress need to be further studied, but are strongly present (Hughes et al., 2017; Petrucelli et al., 2019).
Stroke	Acute stress can increase the risk of blood clots, leading to a higher likelihood of heart attacks and strokes due to changes in endothelial cell function, arterial stiffness, vessel wall damage, elevated blood viscosity, and hypercoagulability. Stress also increases the risk of strokes through the metabolic syndrome (Bhushan et al., 2020).
Alzheimer's disease or dementia	Although the reasons for neuropsychiatric disorders are multifaceted and intricate, alterations to the brain's threat response, pain perception mechanisms, motivation and reward pathways, and impulse control are linked to toxic stress and are considered to play a part in increasing the likelihood of these disorders. Also, accelerated cellular aging as a component of toxic stress physiology may lead to higher rates of Alzheimer's disease and other types of dementia (Bhushan et al., 2020).
Diabetes	Similarly to heart disease and stroke, diabetes is a risk factor of stress through the metabolic syndrome. Moreover, stress affects glucose regulation, insulin resistance, and insulin secretion, with effects even occurring intergenerationally (Bhushan et al., 2020; Lloyd et al., 2005).
Influenza and pneumonia	Unknown
Kidney disease	Stress increases the risk of kidney disease and is believed to be increased by other factors, such as heart disease, obesity, diabetes, and high blood pressure, which can also cause damage to the kidneys. It is suggested that dysregulation of endothelin-1, a molecule involved in regulating blood pressure and arterial stiffness, may be an underlying mechanism through which stress and other risk factors contribute to the development of both cardiovascular and kidney disease (Bhushan et al., 2020; Bruce et al., 2016).
Suicide (attempts)	In addition of stress functioning through depression, anxiety, and other mental health problems to affect risk for suicide, multiple models of suicide include stress directly. It is proposed that suicidal behavior is a result of an interaction between acutely stressful life events and a susceptibility to suicidal behavior (Bhushan et al., 2020; Van Heeringen, 2012).

Note. The association between stress and these causes of death is complex and multifactorial. Moreover, stress may not be the only or even the primary factor contributing to these causes of death. However, reducing stress through lifestyle changes, stress management techniques, and seeking support from family and friends can have numerous health benefits, including reducing the risk of these and other serious health problems.

6.5. Real-world Validation and Application

Looking forward, more research is needed to validate speech against clinical outcomes, and in this context, focusing on disease recurrence should be a top priority for a few reasons. First, these patients already have a demonstrated disease risk (often accompanied by decreased stress resilience) and, therefore, a reason to be followed. Second, early detection

and intervention in these patients would have substantial cost savings, help prevent complete relapse from occurring, and lead to less time-intensive treatments. Finally, these patients are already connected to clinical care, making immediate intervention easier and more likely.

6.6. Ethical and Legal Issues

Ultimately, the real-world implementation of speech data collection comes with multiple privacy and ethical concerns, and these issues must be taken seriously to maximize the potential benefits and minimize the risks associated with this emerging technology. As described by Slavich et al. (2019), risk minimization should include, at minimum: (a) telling users what devices are sampling, assessing, and/or transmitting speech, and providing examples of possible risks; (b) enabling users to digitally control the listening function of devices; (c) enabling users to physically control the listening function of devices (e.g., using the audio equivalent of a physical lens cap); (d) allowing users to manage access to their speech data; (e) permitting users to *opt in* to having devices in their environment; and (f) allowing users to *opt out* of having their speech logged or analyzed. Several high-profile breaches of speech data have occurred (see Slavich et al., 2019), and when it comes to these issues, we believe user data protection must come first.

6.7. Conclusion

In conclusion, speech can drive the next frontier in monitoring health and delivering JITAIs to prevent disease recurrence and foster resilience. Looking forward, more research is

needed to validate speech against ground-truth outcomes, including clinical functioning, diagnoses, biomarkers, and subjective self-reported responses (Sarker et al., 2017). Focusing on stress makes a lot of sense in this context, as stress plays a crucial role in the development and recurrence of numerous mental and physical health conditions, and can be treated using existing evidence-based strategies (Slavich & Auerbach, 2018). We already have the technological devices needed to realize the promise of JITAIs in our hands. To solve some of the world's biggest health challenges, all we need to do is empower these devices with the right diagnostic and therapeutic programs.

6.8. References

- Baird, A., Triantafyllopoulos, A., Zänkert, S., Ottl, S., Christ, L., Stappen, L., Konzok, J., Sturmbauer, S., Meßner, E.-M., Kudielka, B. M., Rohleder, N., Baumeister, H., & Schuller, B. W. (2021). An evaluation of speech-based recognition of emotional and physiological markers of stress. *Frontiers in Computer Science*, 3, 750284. <https://doi.org/10.3389/fcomp.2021.750284>
- Bhushan, D., Kotz, K., McCall, J., Wirtz, S., Gilgoff, R., Dube, S.R., Powers, C., Olson-Morgan, J., Galeste, M., Patterson, K., Harris, L., Mills, A., Bethell, C., Burke Harris, N., & Office of the California Surgeon General. (2020). *Roadmap for Resilience: The California Surgeon General's Report on Adverse Childhood Experiences, Toxic Stress, and Health*. Office of the California Surgeon General. <https://doi.org/10.48019/PEAM8812>
- Bruce, M. A., Griffith, D. M., & Thorpe, R. J. (2015). Stress and the Kidney. *Advances in Chronic Kidney Disease*, 22(1), 46–53. <https://doi.org/10.1053/j.ackd.2014.06.008>
- Bunker, S. J., Colquhoun, D. M., Esler, M. D., Hickie, I. B., Hunt, D., Jelinek, V. M., Oldenburg, B. F., Peach, H. G., Ruth, D., Tennant, C. C., & Tonkin, A. M. (2003). “Stress” and coronary heart disease: Psychosocial risk factors. *Medical Journal of Australia*, 178(6), 272–276. <https://doi.org/10.5694/j.1326-5377.2003.tb05193.x>
- Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., Cipresso, P., Riva, G., & Botella, C. (2019). Current State and Future Directions of Technology-Based Ecological Momentary Assessment and Intervention for

Major Depressive Disorder: A Systematic Review. *Journal of Clinical Medicine*, 8(4), 465.

<https://doi.org/10.3390/jcm8040465>

Connolly Gibbons, M. B., Rothbard, A., Farris, K. D., Wiltsey Stirman, S., Thompson, S. M.,

Scott, K., Heintz, L. E., Gallop, R., & Crits-Christoph, P. (2011). Changes in psychotherapy utilization among consumers of services for major depressive disorder in the community mental health system. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(6), 495-503.

<https://doi.org/10.1007/s10488-011-0336-1>

Faurholt-Jepsen, M., Bauer, M., & Kessing, L. V. (2018). Smartphone-based objective monitoring in bipolar disorder: Status and considerations. *International Journal of Bipolar Disorders*, 6(1), 6. <https://doi.org/10.1186/s40345-017-0110-8>

Gee, B. L., Han, J., Benassi, H., & Batterham, P. J. (2020). Suicidal thoughts, suicidal behaviours and self-harm in daily life: A systematic review of ecological momentary assessment studies. *Digital Health*, 6, 2055207620963958.

<https://doi.org/10.1177/2055207620963958>

Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, Inflammation, and Cancer. *Cell*, 140(6), 883–899. <https://doi.org/10.1016/j.cell.2010.01.025>

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: A systematic review and meta-analysis. *The Lancet Public Health*, 2(8), e356–e366. [https://doi.org/10.1016/S2468-2667\(17\)30118-4](https://doi.org/10.1016/S2468-2667(17)30118-4)

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54, 910-921. <https://doi.org/10.3758/s13428-021-01670-x>

Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van Hoecke, S., & Vanderhasselt, M.-A. (2022). Acoustic speech features in social comparison: How stress impacts the way you sound. *Scientific Reports*, 12(1), Article 1. <https://doi.org/10.1038/s41598-022-26375-9>

Langer, M., König, C. J., Siegel, R., Fredenhagen, T., Schunck, A. G., Hähne, V., & Baur, T. (2022). Vocal-stress diary: A longitudinal investigation of the association of everyday work stressors and human voice features. *Psychological Science*, 33, 1027-1039. <https://doi.org/10.1177/09567976211068110>

Lloyd, C., Smith, J., & Weinger, K. (2005). Stress and Diabetes: A Review of the Links. *Diabetes Spectrum*, 18(2), 121–127. <https://doi.org/10.2337/diaspect.18.2.121>

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1), 96-116. <https://doi.org/10.1002/lio2.354>.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., & Choudhury, T. (2012). StressSense: Detecting stress in unconstrained acoustic environments using smartphones. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, 351-360. <https://doi.org/10.1145/2370216.2370270>

Mote, J., & Fulford, D. (2020). Ecological momentary assessment of everyday social experiences of people with schizophrenia: A systematic review. *Schizophrenia Research*, 216, 56–68. <https://doi.org/10.1016/j.schres.2019.10.021>

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52, 446-462. <https://doi.org/10.1007/s12160-016-9830-8>

Petrucelli, K., Davis, J., & Berman, T. (2019). Adverse childhood experiences and associated health outcomes: A systematic review and meta-analysis. *Child Abuse & Neglect*, 97, 104127. <https://doi.org/10.1016/j.chabu.2019.104127>

Reiche, E. M. V., Nunes, S. O. V., & Morimoto, H. K. (2004). Stress, depression, the immune system, and cancer. *The Lancet Oncology*, 5(10), 617–625. [https://doi.org/10.1016/S1470-2045\(04\)01597-9](https://doi.org/10.1016/S1470-2045(04)01597-9)

Sarker, H. et al. (2017). From markers to interventions: The case of just-in-time stress intervention. In: J. Rehg, S. Murphy, S. Kumar, (Eds), *Mobile Health*. Springer. https://doi.org/10.1007/978-3-319-51394-2_21

Serre, F., Fatseas, M., Swendsen, J., & Auriacombe, M. (2015). Ecological momentary assessment in the investigation of craving and substance use in daily life: A systematic review. *Drug and Alcohol Dependence*, 148, 1–20. <https://doi.org/10.1016/j.drugalcdep.2014.12.024>

Slavich, G. M. (2016). Life stress and health: A review of conceptual issues and recent findings. *Teaching of Psychology*, 43, 346-355. <https://doi.org/10.1177/0098628316662768>

Slavich, G. M., & Auerbach, R. P. (2018). Stress and its sequelae: Depression, suicide, inflammation, and physical illness. In J. N. Butcher & J. M. Hooley (Eds.), *APA handbook of psychopathology: Vol. 1. Psychopathology: Understanding, assessing, and treating adult mental disorders* (pp. 375-402). Washington, DC: American Psychological Association. <https://doi.org/10.1037/0000064-016>

Slavich, G. M., Taylor, S., & Picard, R. W. (2019). Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22, 408-413. <https://doi.org/10.1080/10253890.2019.1584180>

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L., Borges, G., Bruffaerts, R., Bunting, B., de Almeida, J. M. C., Florescu, S., de Girolamo, G., Gureje, O., Haro, J. M., He, Y., Hinkov, H., ... Kessler, R. C. (2017). Undertreatment of people with major depressive disorder in 21 countries. *British Journal of Psychiatry*, 210(2), 119–124. <https://doi.org/10.1192/bjp.bp.116.188078>

Van Heeringen, K. (2012). Stress-diathesis model of suicidal behavior. *The neurobiological basis of suicide*, 51, 113.

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>

Wirtz, P. H., & von Känel, R. (2017). Psychological Stress, Inflammation, and Coronary Heart Disease. *Current Cardiology Reports*, 19(11), 111.
<https://doi.org/10.1007/s11886-017-0919-x>

World Health Organization. (2021). *Mental health atlas 2020*. World Health Organization. ISBN 978-92-4-003670-3.

6.9. Supplemental Materials

6.9.1. Funding

M.K. was supported by grant #KBS 2018-J1130650-209563 from the King Baudouin Foundation. M.A.V. was supported by grant #BOFSTA2017002501 from Ghent University. G.M.S. was supported by grant #OPR21101 from the California Governor's Office of Planning and Research/California Initiative to Advance Precision Medicine. These organizations had no role in planning, writing, editing, or reviewing this article, or in deciding to submit this article for publication.

General Discussion

Stress is an everyday experience, encountered by everyone on a regular basis. It emerges as the body's reaction to various demands or threats, referred to as *stressors*, which originate from a multitude of sources such as financial issues, childcare, interpersonal relationships, social gatherings, or significant life events. The growing prevalence of stress in today's fast-moving world is closely associated with its substantial impact on both physical and mental well-being.

Although stress can serve as a normal and adaptive aspect of life, helping individuals cope with challenges and motivating the pursuit of goals, it is essential to identify when stress becomes harmful. Specifically, chronic stress can result in a wide range of health complications, including cardiovascular disease, coronary heart disease, anxiety disorders, depression, autoimmune disease, and neurodegenerative disorders. (Bhushan et al., 2020; Brosschot et al., 2017; S. Cohen et al., 2007; Juster et al., 2010; Kappen et al., 2023; Slavich & Irwin, 2014). It has also been connected to most leading causes of death in the US, as shown in Table 1 (Kappen et al., 2023).

Given the significant impact of stress on our well-being, accurately and frequently measuring stress levels is vital. As such, many different stress measurement methods have been developed, each with its own strengths and weaknesses. Therefore, the pressing need for developing innovative and accessible stress measurement methods persists. Increasingly, measuring stress from one's speech has gained attention. Specifically, the production of speech involves multiple physiological systems of the body that also play a crucial role in physical stress reactions (Slavich et al., 2019; Van Puyvelde et al., 2018), including multiple cranial and spinal nerves (Duffy, 2000), multiple subcortical and cortical brain regions (Carlson & Birkett, 2017; Jürgens, 2002), and cardiorespiratory processes (Câmara & Griessenauer, 2015; Monkhouse, 2005).

Measuring stress through speech would come with many perks, such as it being affordable, accessible, non-intrusive, swift, applicable in natural settings, and low effort for participants, which makes it a good option for high-frequency and passive monitoring of stressors in daily life. Therefore, researchers have tried to unveil which features of one's speech respond to stress specifically and in what context. However, there has been considerable heterogeneity in the observed results, which can be attributed to various limitations in their designs as described in **Chapter 1** (Giddens et al., 2013; Slavich et al., 2019; Van Puyvelde et al., 2018).

Table 1

Role of stress in the top 10 causes of death in the United States.

Disease	Role of stress
Heart disease	Stress causes cardiovascular hyperreactivity and stress-induced hyperactivation of the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system (SNS), both related to coronary heart disease as acute coronary syndromes. Stress causes inflammation, increases in cholesterol, depression, risk of smoking, and metabolic syndrome, all strongly related to heart disease (Bhushan et al., 2020; Bunker et al., 2003; Wirtz & von Känel, 2017).
Cancer	Persistent activation of the HPA axis impairs the immune response and, together with chronic inflammation, heightens the risk of cancer and promotes the spread of cancer after development (Bhushan et al., 2020; Grivennikov et al., 2010; Reiche et al., 2004).
Accidents (unintentional injuries)	Stress can impair cognitive function, reaction time, and decision-making, which can increase the risk of accidents. Moreover, chronic and early life stress leads to more risk seeking behavior, further increasing the risk of accidents.
Chronic lower respiratory diseases	Stress can weaken the immune system and make it more difficult for the body to fight off infections, which can increase the risk of respiratory diseases. The exact associations with stress need to be further studied, but are strongly present (Hughes et al., 2017; Petruccelli et al., 2019).
Stroke	Acute stress can increase the risk of blood clots, leading to a higher likelihood of heart attacks and strokes due to changes in endothelial cell function, arterial stiffness, vessel wall damage, elevated blood viscosity, and hypercoagulability. Stress also increases the risk of strokes through the metabolic syndrome (Bhushan et al., 2020)
Alzheimer's disease or dementia	Although the reasons for neuropsychiatric disorders are multifaceted and intricate, alterations to the brain's threat response, pain perception mechanisms, motivation and reward pathways, and impulse control are linked to toxic stress and are considered to play a part in increasing the likelihood of these disorders. Also, accelerated cellular aging as a component of toxic stress physiology may lead to higher rates of Alzheimer's disease and other types of dementia (Bhushan et al., 2020).
Diabetes	Similarly to heart disease and stroke, diabetes is a risk factor of stress through the metabolic syndrome. Moreover, stress affects glucose regulation, insulin resistance, and insulin secretion, with effects even occurring intergenerationally (Bhushan et al., 2020; Lloyd et al., 2005).
Influenza and pneumonia	Unknown
Kidney disease	Stress increases the risk of kidney disease and is believed to be increased by other factors, such as heart disease, obesity, diabetes, and high blood pressure, which can also cause damage to the kidneys. It is suggested that dysregulation of endothelin-1, a molecule involved in regulating blood pressure and arterial stiffness, may be an underlying mechanism through which stress and other risk factors contribute to the development of both cardiovascular and kidney disease (Bhushan et al., 2020; Bruce et al., 2016).
Suicide (attempts)	In addition of stress functioning through depression, anxiety, and other mental health problems to affect risk for suicide, multiple models of suicide include stress directly. It is proposed that suicidal behavior is a result of an interaction between acutely stressful life events and a susceptibility to suicidal behavior (Bhushan et al., 2020; Van Heeringen, 2012).

The primary aim of this dissertation was to explore the potential of speech as a measure of stress by addressing these limitations in the current literature. This is achieved through a

series of studies that employ large, statistically-powered studies, where we elicited stress using psychosocial stressors, in non-actor participants, validate stress inductions using self-reports and/or physiological responses, use within-participant designs, including neutral or active control condition speech recordings, ensure consistency in the testing environment and microphone quality, and focus on speech features with a scientific basis.

In the following sections, I will first briefly summarize the main findings of the performed studies ([section 7.1.](#)). Next, I will discuss the theoretical implications by framing these results within the context of the present literature ([section 7.2.](#)). Following, I will discuss the potential practical and clinical implications of the current findings ([section 7.3.](#)), including specifically discussing implications for continuous monitoring settings ([section 7.4.](#)). Then, the limitations of the studies from this dissertation as well as the field of speech as a whole will be discussed ([section 7.5.](#)). Lastly, suggestions for future research are proposed ([section 7.6.](#)) as well as stating the dissertation's general conclusions ([section 7.7.](#)).

7.1. General overview of the findings

In the first study, **Chapter 2**, we collected speech before and after exposing participants to stress (using the MIST), and extracted core speech features as they are described in the literature. By applying network analysis, we found that the networks of how these speech features are connected remain identical before and after the stress induction. When analyzing the delta (change; post- vs pre-stressor) network, we discovered a comprehensive network that shows that changes in any feature were related to changes in

self-reported negative affect—an actual index of how stressed someone was—through changes in vocal jitter. This study provides insights into the complex relationships between different speech parameters in the context of psychosocial stress, highlighting the central role of harmonic-to-noise ratio (HNR) in the network and the potential importance of vocal jitter in its relation to self-reported negative affect. These results necessitate and enable the investigation of stress effects on these speech features in a confirmatory manner. Moreover, the following chapters should incorporate the use of active control blocks to further isolate the impact of stress on speech.

In the second experimental study, **Chapter 3**, participants engaged in a cognitively challenging task and received neutral or negative comparative feedback on their performance. We used a within-subject design and validated a successful stress induction with self-reports and physiological measures. Our analysis of acoustic speech features, from read-out-loud speech, revealed a significant increase in Fundamental Frequency (F0) and harmonics-to-noise ratio (HNR), and a significant decrease in shimmer during the negative feedback condition. These strong results, generated using a validated stress paradigm, an active control condition (i.e., neutral vs negative socially comparative feedback), a high-quality microphone, and validated self-reports and physiological measures, contribute to the understanding of stress effects on specific acoustic speech features in a well-controlled but ecologically-valid stress setting. This study is a solid step toward the generalization of these findings to real-life settings. Subsequent chapters should move towards more realistic settings by incorporating a greater diversity of stressors to examine the robustness and sensitivity of the speech features and stress, as well as utilizing speech samples that better resemble everyday speech.

In **Chapter 4**, we developed the Ghent Semi-spontaneous Speech Paradigm (GSSP) to enable researchers to capture unscripted speech data in affective-behavioral research. The GSSP allows for flexible speech acquisition durations, non-interfering tasks, experimental control, prosodic richness, and minimal human interference for scalability. We validated the GSSP through an online task, comparing it to a fixed-text read-out-loud task. Acoustic analysis revealed trends consistent with the targeted speech styles (unscripted spontaneous speech vs. scripted read-aloud speech), and a speech style classification model achieved a balanced accuracy of 83% on within-dataset validation, indicating separability between the GSSP and read-out-loud speech task. Therefore, the GSSP is a valuable tool for capturing spontaneous speech in longitudinal ambulatory behavioral studies and laboratory studies, advancing the field toward utilizing speech as a biomarker in everyday settings. Specifically, we propose that the GSSP should be the advised method of collecting speech in experimental settings, generating results that would be directly implementable to real-world scenarios, as opposed to read-out-loud speech, which is less natural. Thus, this paradigm allows us to obtain more ecologically valid speech samples in our subsequent chapters and studies.

Chapter 5 investigated the effects of two distinct psychosocial stress paradigms (Cyberball and MIST) on semi-guided speech features. We observed that only negative affect increased during Cyberball, while self-reported stress, and skin conductance response rate, in addition to negative affect increased during MIST. This is the first study using a multi-day, multi-paradigm, within-participant experimental setup, and collected speech through a semi-guided picture description paradigm, similar to the GSSP. Fundamental frequency (F0), speech rate, and jitter significantly changed during MIST, but not Cyberball, while HNR and

shimmer showed no expected changes. These observed changes in most speech features follow the self-reported and physiological reactions (i.e., responsive during MIST, but not Cyberball). The results indicate that observed speech features are robust in semi-guided speech (as found in previous studies using read-out-loud speech) and sensitive to stressors eliciting additional physiological stress responses, rather than solely decreases in negative affect. We can relate these results to the fact that the production of speech involves multiple physiological systems of the body that are specifically relevant to the physical stress component (Câmara & Griessnauer, 2015; Monkhouse, 2005; Slavich et al., 2019; Van Puyvelde et al., 2018). These differences between stressors may explain the heterogeneity in the literature and further support the potential of speech as a biomarker for stress. Because changes can be 1) observed in freely spoken speech, 2) respond to physiological stress reactions specifically, and 3) are as responsive as other used methods. This highlights the promise of speech as a tool for measuring stress in everyday settings, considering its affordability, non-intrusiveness, and ease of collection.

In **Chapter 6**, our perspective, we discuss the potential impact of speech as a biosignal or biomarker in precision psychiatry. We explore the current use of experience sampling methods and other innovative approaches for assessing well-being throughout the day, as well as their application in just-in-time interventions (JITAs) for people at risk, such as those experiencing suicidal behavior. We propose that speech, as a marker for stress as a transdiagnostic risk factor, could be the missing link in fine-tuning these systems, offering an easily accessible and affordable method. Furthermore, we address practical and ethical implications. Speech as a novel digital biosignal could play a key role in predicting high-priority

clinical outcomes and delivering tailored interventions to help people when they need it the most if appropriately and securely implemented. We urge for more research to identify the most relevant speech signals and to validate them against ground-truth outcomes, to then translate speech data into biomarkers and JITAs.

Table 2

Schematic overview of current findings

Speech-style	Negative		Stressor type		<i>Ostracism</i>
	F0	HNR	Cognitive Load	?	
Read-out-loud	Shimmer	Jitter	?	?	?
	MVSPS	MVSL			
	<i>Chapter 3</i>				
	<i>Network-Analysis</i>				
	<i>Chapter 2</i>				
<i>Semi-guided</i>	?		F0	HNR	F0
	Shimmer	Jitter	Shimmer	Jitter	Shimmer
	MVSPS	MVSL	MVSPS	MVSL	MVSPS
	<i>Chapter 5</i>				<i>Chapter 5</i>

Note. Findings per chapter are categorized per speech style and stressor type. Green indicates an increase and red indicates a decrease under stress. White indicates no observed effect. Question marks are used to indicate missing pieces in the literature. F0; Fundamental frequency, HNR; Harmonics-to-Noise Ratio, MVSPS; Mean Voiced Segments Per Second, MVSL; Mean Voiced Segment Length.

7.2. Theoretical implications

7.2.1. Combined patterns of speech features

A recent review highlighted a limitation in stress in speech research, where individual features are often considered in isolation (Van Puyvelde et al., 2018). Researchers have emphasized the need to examine combined patterns of multiple voice parameters that might respond in a meaningful way, rather than exclusively focusing on the expression of each voice parameter separately (Godin & Hansen, 2015; Van Puyvelde et al., 2018). To address this issue, we conducted network analyses in **Chapter 2**, which revealed the interconnectedness of speech features remained stable when comparing pre-stress and post-stress recordings.

Our analysis extended beyond this initial finding by examining change networks that directly related each individual's speech feature changes to their changes in negative affect. We discovered a comprehensive network with only one direct connection between a speech feature (jitter) and mood, indicating that changes in other speech features due to mood alterations function through jitter.

7.2.2. Current heterogeneity

Throughout the existing literature on the effects of stress on speech, there has been considerable heterogeneity in the observed results, which can be attributed to various limitations (Giddens et al., 2013; Slavich et al., 2019; Van Puyvelde et al., 2018). Heterogeneity

refers to the variability or diversity of findings across different studies or analyses and it can arise due to differences in study design, methodology, populations, or other factors that contribute to the dispersion of outcomes (Fletcher, 2007). The disparity in outcomes is not necessarily bad, as it gives us a comprehensive understanding of the concept's complexity and its sensitivity to contextual factors, and it encourages further research. However, it also means there is a certain difficulty in the interpretation of the results and puts us at risk of false conclusions when considering the results from just one study.

Considering the novelty of the field of measuring stress (and many other psychological and physiological phenomena) through speech, it is of utmost importance to target this heterogeneity early to have a solid foundation to build upon. Therefore, this dissertation makes a first step towards finding the *true* effects (and their complexity) of psychosocial stress on speech, whilst building on the known knowledge. We exclusively ran statistically powered studies, using within-participant designs, eliciting actual stress in real participants (rather than actors portraying stress), and a controlled environment to ensure conditions are stable within the sample. These standards ensure the minimization of observing false positives in our results and supply trustworthy effects that could translate to real-world scenarios.

However, as mentioned before, heterogeneity itself is not inherently negative, as it reveals the intricacy of an effect. To uncover true effects and unravel this complexity, consistent and systematic research is necessary, which involves studying specific features in various contexts. With regard to speaking style and stressor type, we do not claim that there is a single correct answer. As such, the following sections will outline some considerations

concerning these aspects and explain how this dissertation addresses these facets of heterogeneity.

7.2.2.1. Speech styles

In our first studies, we used read-out-loud paradigms to generate results. This was purposely done, because scripted lab speech more conveniently allows for systematic experimental control, thus limiting the implicit inclusion of unwanted latent variables (Xu, 2010). As such, we conducted our study described in **Chapter 3**, where multiple scripted speech recordings were done throughout both a stress task (negative social comparison) and an active control. This was the first study to do so, and as such, the results show how speech behaves under stress when all other factors are controlled. This could be described as a rather fundamental approach since it minimizes the background and situational noise that would occur in more naturalistic settings, such as background sounds (e.g., traffic, wind), conversational overlap, and acoustic variability (e.g., variable distance to microphone). The observed results were partially in line with former literature, such as increases in HNR and F0, and a decrease in Shimmer, yet, no results were observed for jitter, MVSL, and MVSPS (Giddens et al., 2013; Kappen, Van Der Donckt, et al., 2022; Van Puyvelde et al., 2018).

However, acoustic properties found in one speech style can be style-specific, which limits the explanatory power of the speech results to other settings (e.g., the real world). Unscripted speech (as found in daily life), which requires larger planning units such as sentences, clauses, and temporal structure, can lead to changes in wording, grammar, and timing of speech under these affective states (Fromkin, 1973; Paulmann et al., 2016; Slavich et

al., 2019). These prosodic markers are less pronounced in scripted speech, as fewer planning units are needed (Barik, 1977; Xu, 2010). Therefore, a promising research direction would be to investigate the influence of stress on speech in different speech styles that could be implemented in real-world settings (Wagner et al., 2015). On top of this, the scalability of speech acquisition methods should be considered, given that the long-term objective of affective sensing experiments is to facilitate widespread, real-world affect monitoring (Slavich et al., 2019), which is challenging and perhaps unrealistic using exclusively scripted speech acquisition.

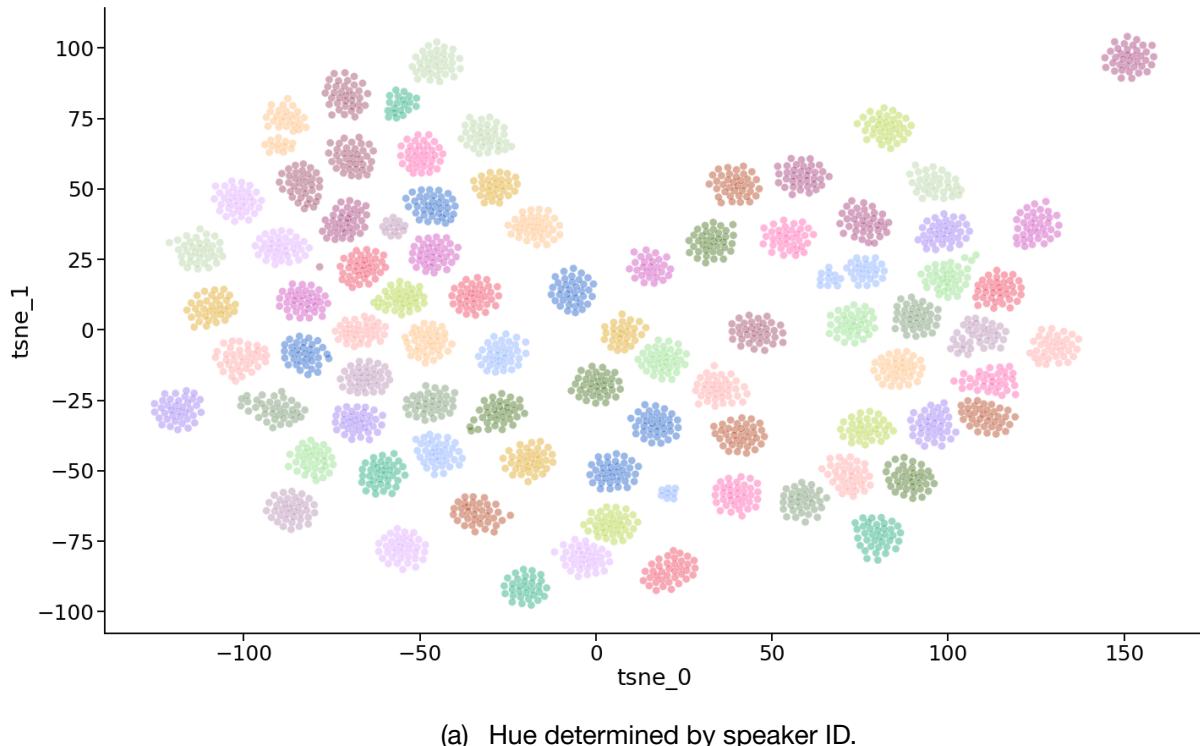
In **Chapter 4**, we outline considerations with regard to the differences in speech style (i.e., unscripted spontaneous speech versus scripted read-aloud speech) and propose what would be needed to generate results from speech research both in the lab and in the real world and how they no longer have to be two different fields, but could actually support each other. Speech style As such, we developed the Ghent Semi-spontaneous Speech Paradigm (GSSP) which entails a picture-description paradigm to collect speech data. To validate whether the GSSP would be able to cater to this need, we validated not only whether the collected speech from the GSSP was different from the scripted speech, but also to what extent it was similar to naturalistic speech. We first demonstrated that indeed, each person has their own speaking style, as illustrated by a different color for every speaker in Figure 1a¹, but we also showed that for each person there was a clear distinction in speaking style between the two paradigms as shown in Figure 1b¹³. Moreover, when we validated the model using an external dataset

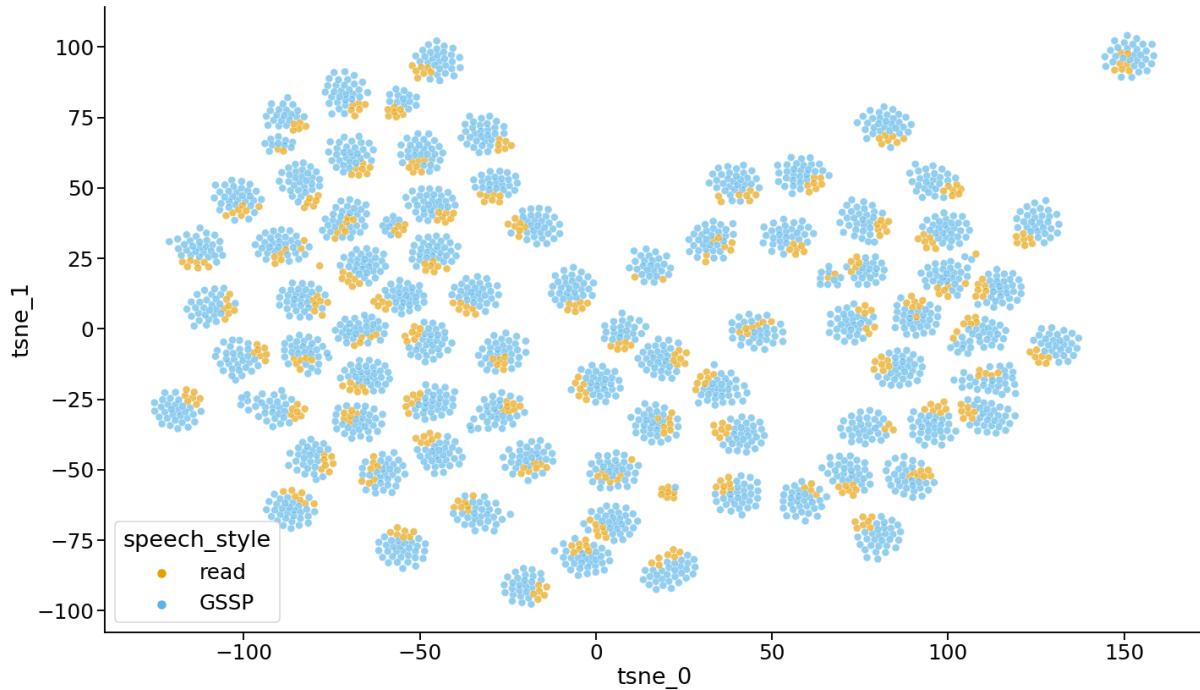
¹³ In the provided figures (t-SNE visualization) for speaker identification, it is important to note that the x and y axes do not represent specific, interpretable values or measurements. Rather, the primary focus of this visualization is the clustering of data points, which correspond to speakers exhibiting similar features. The axes serve to facilitate the representation of these relationships within a two-dimensional plane, and as such, the numerical values on the axes should not be a point of concern.

containing both interview and read-aloud speech, we achieved a balanced accuracy score of 70%. This result indicates that there is an acoustic similarity between the speech collected through the GSSP and spontaneous interviewee speech. The GSSP is therefore a crucial addition and main recommendation for future speech research, both in the lab and in everyday settings.

Figure 1.

Plots showing t-SNE visualizations for speaker identification on all experiment data, collected using the GSSP and read-out-loud speech¹. Plot a shows hue based on speaker ID, plot b shows hue based on speech style (i.e., GSSP vs read-out-loud).





(b) Hue determined by speech style.

Using a picture description in **Chapter 4**, therefore adds substantially to the literature, considering the reproduction of several speech features such as an increase in F0, MVSPS, and MVSL, and a decrease in jitter, but not being able to find significant changes in HNR and shimmer. These results give us a clear direction on which speech features will play a substantial role in the development of further deterministic modeling.

7.2.2.2. Diversity in stressors

Another proposed explanation for the observed heterogeneity in the literature is related to how different studies use different stressors. Not only is there often a lack of differentiation between physiological and psychological stressors (Giddens et al., 2013) but there is also a lot of variability in stress reactions to different stressors and between individuals. Indeed, like physical stressors (e.g., electric shock, prolonged exercise), psychological stressors are indeed

capable of activating the HPA axis (Dickerson & Kemeny, 2004; Kirschbaum et al., 1993). However, the effects of psychological stressors on this physiological system are highly variable, whereas many studies have failed to find cortisol changes and inconsistencies are often described (Allen et al., 2014; Biondi & Picardi, 1999; Dickerson & Kemeny, 2004; Manuck et al., 1991). As such, we designed a study targeting this exact challenge. In **Chapter 5** we showed, in a within-participant design, that participants indeed responded differently to two different stressors (i.e., Cyberball & MIST). Not only did the responses differ in self-reports and physiological measures, but also in speech features. More specifically, the speech features showed similar responsiveness to the different stressors as the physiological and self-reported stress measures (i.e., only responding to the MIST, and not the Cyberball). These results support the claim that the observed heterogeneity is due to the variety of used stressors in the literature and show that these core speech features are exclusively responsive to stressors eliciting a physiological reaction when being assessed in semi-guided (picture describe) speech. Specifically, this shows that our endeavors to make deterministic models of everyday stress from free speech samples should focus on using F0 (~pitch), MVSPS (~speech rate), and jitter (pitch variation).

7.3. Practical and clinical implications

As discussed in the former section, this dissertation mainly adds to the literature by showing that stress indeed has significant effects on the voice and that these resemble the sensitivity of other, often-used methods. Considering the cost-effectiveness, ease of collection, and accessibility of this data modality, it holds great promise for the future of health monitoring.

However, it should be noted that the implementation of speech as a detection tool for stress is beyond our objectives, but these findings will primarily contribute to the development of such tools. Therefore, the following section will touch upon the practical and clinical implications of further developing human speech into a trustworthy biomarker for stress.

First, it is good to acknowledge that this goes beyond the mere field of stress and this dissertation's scope, but crosses into a wide variety of disorders and symptoms, including, but not exclusively, depression, post-traumatic stress disorder, eating disorders, and anxiety disorder (AI Hanai et al., 2018; J. F. Cohn et al., 2009; Koops et al., 2021; Low et al., 2020; Marmar et al., 2019; Voppel et al., 2022). Increasingly, models are proposed that are trained on datasets to predict whether someone is depressed or stressed or any of these other factors. However, many of these studies share a limitation, which is a problem with feature selection. It is often unclear which features play a role, so large feature sets are used, resulting in often overfitted models that would not function well in newly presented data (Parry et al., 2022). Therefore, many studies specifically focus on identifying particular features that respond in order to identify a set of features that should be given extra weight and can make the final stretch to predictive models that would translate well to newly presented data (Baird et al., 2019, 2021; Bhatia et al., 2021; Boyer et al., 2018; J. Li et al., 2022; N. Li et al., 2021; Rodellar-Biarge et al., 2015). This dissertation outlines, specifically in increasingly naturalistic settings, how certain speech features behave with regard to stress. This information is key to building deterministic models that translate well to the grand population, especially when expanding to and taking into consideration the individualistic inter-speaking variability of stress

reactions, both vocally and physiologically (Giddens et al., 2013; Kappen, Hoorelbeke, et al., 2022; Kurniawan et al., 2013; Van Puyvelde et al., 2018; Zhu-Zhou et al., 2022).

Being able to accurately detect stress from one's speech would be of value in a wide variety of settings, such as quick stress assessments in emergency calls (König et al., 2021), speech recognition (Bou-Ghazale & Hansen, 2000) for instance for security systems and human-computer interactions, and as a transdiagnostic marker for many different psychiatric diseases (Kappen et al., 2023).

7.4. Continuous monitoring

The chronic and relapsing nature of many mental health disorders is the rule and not the exception. Thus, the need for long-term follow-up and assessment methods become essential for patients' symptom reduction and recovery. Traditional monitoring methods often rely on retrospective reports which are subject to recall bias, lack of measurement frequency, and are time-consuming (Abbas et al., 2021; Garcia-Ceja et al., 2018; Slavich et al., 2019). This approach limits the ability to accurately characterize, understand, and change behavior in real-world settings (Garcia-Ceja et al., 2018; Shiffman et al., 2008). As such, continuous monitoring of people's mental health has gained increasing interest in recent years as a means to better understand the dynamic nature of psychological well-being. This approach focuses on the real-time assessment of an individual's thoughts, emotions, and behaviors, offering valuable insights into the fluctuating patterns of mental health (Shiffman et al., 2008). By tracking these changes over time, continuous monitoring can provide a more comprehensive

understanding of an individual's mental health and inform personalized interventions. One effective method for continuous mental health monitoring is the use of Ecological Momentary Assessments (EMAs).

The use of EMAs in psychiatric and psychological research allows for the examination of dynamic patterns in mental health symptoms and their underlying mechanisms (Garcia-Ceja et al., 2018; Shiffman et al., 2008). This can lead to a better understanding of the temporal relationships between various psychological factors and may reveal crucial information about the onset and progression of mental health issues. EMA measurements have been shown to outperform paper and pencil reports in the assessment of some mental states in terms of sensitivity to detect changes (Moore et al., 2016). However, EMAs are still often limited to high-frequency self-reports, and could greatly flourish with the use of our earlier-mentioned multimodal approach. Specifically, integrating data from smart devices all around us can vastly increase the inference we have on mental health metrics (Donker et al., 2013; Firth et al., 2017; Torous et al., 2014). EMAs enable the investigation of contextual factors that may influence mental health, such as environmental stressors, social interactions, and daily routines, which can provide valuable insights for the development of tailored interventions. However, including data from other sensors can even supply more contextual information such as physical activity (Lara & Labrador, 2013), location (Brena et al., 2017), mood (LiKamWa et al., 2011), and social relationships (Eagle & Pentland, 2006).

Currently, many researchers are focussing on individual or the combination of only a few different modalities, whereas the combination of as many different methods at hand would offer continuously increasing accuracies (Abbas et al., 2021; Garcia-Ceja et al., 2018;

Giannakakis et al., 2022). In a future of digital phenotyping, where multiple data streams are integrated for clinical decision-making, self-reports will remain an important feature of health and function, but other measures are needed to maximize its potential (Abbas et al., 2021; Garcia-Ceja et al., 2018; Kappen et al., 2023). Analyzing speech, from the ubiquitous amounts of speech data constantly being recorded around us, might be the missing link here, as presented in **Chapter 6**. Specifically, if we are able to derive an accurate index of the stress of speech signals, considering its transdiagnostic presence (Giannakakis et al., 2022; Kappen et al., 2023).

By continuously monitoring stress, we not only increase the sampling frequency but also gain insight into the highly dynamic nature of stress, which is critical for understanding individual well-being and developing tailored interventions (Cramer et al., 2016; Wichers, 2014). For example, research has shown that individual responses to stress and recovery from stress can be indicative of more serious psychopathology (Burke et al., 2005). This approach can also incorporate various aspects of mental health, such as sleep, which has been shown to play a critical role in stress response and recovery (Cramer et al., 2016; Hemmeter et al., 2010). Sleepiness can even be detected from speech, further emphasizing the potential of multimodal monitoring (Martin et al., 2019). By taking into account multiple factors, such as stress and sleep, we can develop a comprehensive understanding of the complex interplay between these factors and their role in the onset and progression of mental health issues. This focus on the dynamic nature of mental health and the integration of multiple symptoms ultimately contributes to a more effective, personalized approach to mental health care (Borsboom, 2008; Cramer et al., 2010, 2016; Robinaugh et al., 2020).

To summarize, in mental health, continuous monitoring can facilitate personalized care by tracking individuals' mental health status in real-time, helping clinicians identify early warning signs, and adjust treatment plans accordingly (Faurholt-Jepsen et al., 2018; Garcia-Ceja et al., 2018; Kappen et al., 2023; Shiffman et al., 2008). This approach can also empower individuals to become more aware of their own mental health patterns and triggers, potentially promoting self-management and fostering resilience. Overall, the incorporation of continuous, passive monitoring in psychiatric and psychological research and mental health monitoring has the potential to significantly enhance our understanding of mental health processes and contribute to the development of more effective, personalized interventions (Abbas et al., 2021; Garcia-Ceja et al., 2018; Kappen et al., 2023).

7.5. Limitations

The following sections will discuss some limitations, both of speech research in general as well as this dissertation specifically.

7.5.1. Limitations of speech research in previous work

Firstly, it is crucial to acknowledge that many early studies relied on publicly available datasets instead of collecting their own data. These datasets often lack multiple recordings of the same individual under different conditions, such as a relaxed state, or are composed of lab recordings featuring actors simulating emotions, resulting in exceptional audio quality that may

not be reflective of real-life situations (Giddens et al., 2013; Zhu-Zhou et al., 2022). Consequently, the existing literature is rather limited and its applicability is restricted.

Additionally, a significant number of studies employ different types of stress and various methods to record speech. These varying approaches make it challenging to consolidate results, leading to observed heterogeneity that might not be entirely justified (Giddens et al., 2013; Van Puyvelde et al., 2018). Both these limitations have been partially challenged in this dissertation, but more work is needed, as described in the future directions section.

7.5.2. Limitations of this dissertation

While our research makes significant steps toward the real-world application of this novel method, certain limitations must be acknowledged. Firstly, our sample was exclusively collected in Flanders and the Netherlands among predominantly young people. Research indicates that while some vocal bursts are similar across cultures (79%), differences exist in vocal expression (Brooks et al., 2023). Although these studies primarily focus on *emotional* speech contexts across cultures, it can be assumed that this translates to speech characteristics related to stress (Cowen et al., 2019). Assuming a similar pattern, our results remain highly informative for intercultural studies, but validation in other samples is necessary.

Additionally, the research in this dissertation focused specifically on acute stressors. Acute stressors are adaptive in nature to handle situations that are deemed threatening. They are, therefore, not inherently related to negative health outcomes. When stress becomes more chronic, it can lead to a wide range of health problems including cardiovascular disease,

coronary heart disease, anxiety disorders, depression, autoimmune disease, and neurodegenerative disorders, among others (Bhushan et al., 2020; Brosschot et al., 2017; S. Cohen et al., 2007; Juster et al., 2010; Kappen et al., 2023; Slavich & Irwin, 2014). Therefore, the direct impact on the long-term human well-being of the presented research is limited in itself until it is expanded to include a focus on chronic and lifetime stressors. Moreover, our research focused on group-level stress-induction, without taking individual emotion regulation strategies and coping mechanisms into account. Whilst the speech recordings were administered directly after or during the tasks, limiting the space for conscious reappraisal due to occupancy with the task, it is likely that participants showed emotion regulation and coping strategies during the task (Wang & Saudino, 2011). The occurrence of such techniques would have minimized the observed effect sizes in stress responses and, therefore, speech reactivity. As such, future research should gauge participants' emotion regulation measures.

Lastly, an ideal addition to this dissertation would have been the development of a large, openly available dataset, as the field currently lacks high-quality, high-dimensional ones. However, due to recent changes in ethical guidelines and the sensitive nature of vocal recordings, we have not yet been able to create a dataset that fully meets the current needs. We have, however, made all our individual datasets as openly accessible as possible, complete with detailed guidelines and usage instructions.

7.6. General suggestions for future research

In order to further progress the development of speech as a biomarker for stress, we propose some suggestions for future research taking on the aforementioned limitations. Firstly, it is necessary to keep expanding sample sizes as well as the diversity of sample characteristics. Large samples will both enable us to detect smaller effects and differences, for instance between stressors, but could also enable us to identify individual differences in vocal responses to stress. More specifically, it is currently often argued that the individual's vocal and stress responses are highly individual, but it could also be proposed that there are in fact highly dynamic underlying latent groups (Abbas et al., 2021; Garcia-Ceja et al., 2018; Giddens et al., 2013; Van Puyvelde et al., 2018).

Moreover, we should increase the diversity of samples used in these studies. It will yield unique insights into cultural differences in vocal and stress responses, but could also help develop this method to detect stress in groups that would profit the most from it, such as ethnic, sexual, and socio-economic minorities (Bhushan et al., 2020; Diamond & Alley, 2022; Frisell et al., 2010; Meyer, 2003; Slavich et al., 2023; Walton & Cohen, 2011).

Furthermore, researchers often limit their endeavors to one specific dependent variable, such as stress, depression, schizophrenia, sleepiness, etc. (Cho et al., 2022; Koops et al., 2021; Langer et al., 2022; Low et al., 2020; Lu et al., 2012; Martin et al., 2021; Voppel et al., 2022). However, there is an increasing network approach when discussing psychopathology, that suggests that mental disorders can be conceptualized and studied as causal systems of mutually reinforcing symptoms, rather than the distinct groups as they are identified by the

DSM (Cramer et al., 2010; de Boer et al., 2021; Robinaugh et al., 2020). While staying up to date with these latest developments, we should also acknowledge how the identification of psychopathology and symptoms in a real-world setting would look as it is rarely a matter of dichotomously approaching each individual label (e.g., depression; yes/no), but more trying to make a probability estimate of each individual label at the same time when considering a multitude of symptoms. Therefore, future research should try and face this challenge too by considering a network of dependent variables instead (Borsboom, 2008; Cramer et al., 2010; de Boer et al., 2021; Robinaugh et al., 2020).

Before reaching the stage of real-world applications, speech analysis must be tested and validated in various settings, as the real world offers diverse conditions, and ideally, this technique should perform well across a wide range of situations. Future studies should take incremental steps towards increasingly naturalistic settings, including the introduction of multiple speakers in recordings, enhancing noise in the signal, expanding the variety and intensity of stressors, testing with real-world microphones such as those found in smartphones, and gradually moving towards a completely natural, real-world testing environment, see Table 2 (Kappen et al., 2023; Paulmann et al., 2016; Slavich et al., 2019; Zhu-Zhou et al., 2022).

When we reach the stage of real-world applications, we can begin developing models tailored to each individual's specific responses and environments. These applications should integrate and combine a wide variety of methodologies that could learn from each other and improve collectively (Abbas et al., 2021; Garcia-Ceja et al., 2018; Van Puyvelde et al., 2018).

More specifically, speech measurements could be both improved by and incorporated into current high-frequency patient monitoring systems. These systems have demonstrated promising results in the fields of self-harm, suicide, depression, bipolar disorder, schizophrenia, alcohol and substance abuse, and more, and would likely greatly benefit from the integration of speech measures (Colombo et al., 2019; Faurholt-Jepsen et al., 2018; Gee et al., 2020; Kappen et al., 2023; Mote & Fulford, 2020; Serre et al., 2015).

We specifically designed Table 2 to illustrate which aspects of the puzzle this dissertation has addressed, and which areas still require further investigation. By filling in the empty spaces, inconsistencies that arise due to factors such as individual differences, settings, and speech styles will become more evident, allowing us to work towards the ultimate goal of understanding how speech, recorded in a wide variety of settings, responds to a diverse range of stressors, making it an applicable biomarker for stress.

7.6.1. Language-based markers

The studies presented in this dissertation predominantly focused on acoustic and prosodic speech features. These were chosen because they are easy to extract and less language and culturally dependent. However, future research could expand into investigating the effects of stress on language-based features. Currently, these methods are highly time intensive, as the main method of acquiring accurate transcriptions of speech samples is by hand, but with recent developments in automated speech-to-text algorithms, these approaches will be increasingly easily accessible. Including semantic features could reveal unique insights

into underlying processes, such as cognitive capacity (A. S. Cohen et al., 2014; Hansen & Patil, 2007; Parry et al., 2022).

The potential of language-based features for stress and mental health monitoring has been supported by a range of studies. For instance, Rude et al. (2004) and Tausczik and Pennebaker (2010) showed that individuals experiencing depression tend to use more first-person singular pronouns, which indicates that pronoun usage should be explored in the context of stress detection too. Moreover, research analyzing social media data has demonstrated that certain words, phrases, and linguistic patterns can be associated with mental health conditions (Coppersmith et al., 2014). This highlights the potential of text analysis as a tool for detecting mental health-related linguistic features.

Additionally, studies have found associations between daily word use and psychological states, with certain linguistic features, like the use of function words, relating to emotional well-being (Pennebaker et al., 2003). As well as in the context of trauma, research has shown that the use of specific words, such as negative emotion words, can be associated with increased psychological distress (M. A. Cohn et al., 2004). Furthermore, self-reported emotions have been found to be related to the use of certain emotional words and vocabulary richness, reinforcing the potential of language-based features for stress detection (Shuman et al., 2015).

Taken together, these findings suggest that language-based features, including pronoun usage and word choice, can be indicative of an individual's psychological state, such as stress. Further investigation of language patterns in stress research could further increase accuracy

from speech samples and pave the way for innovative approaches to understanding, monitoring, and managing stress, ultimately enhancing overall mental health and well-being.

7.7. General Conclusions

The primary goal of the research presented in this dissertation was to develop and validate speech as a biomarker for acute stress using state-of-the-art speech analysis techniques. Our findings contribute to the understanding of how speech characteristics can be used to detect and monitor stress levels, with potential real-world applications in various fields such as mental health care, occupational health, and personal well-being.

In summary, our research identified specific speech characteristics related to acute stress and how they interact under stress (Chapter 2), validated these features using read-out-loud speech and negative social feedback (Chapter 3), developed a new methodology to collect naturalistic speech samples (Chapter 4), validated the speech features in freely spoken speech and in different stressor paradigms (Chapter 5), and presented a context of future implementations for speech as a biomarker for stress (Chapter 6). These findings provide a foundation for future research aimed at refining and expanding the use of speech analysis as a biomarker for stress and mental health.

The clinical relevance and potential applications of this research are significant, as the developed methodology can be translated into tools for monitoring stress and mental health in various populations. Integrating speech analysis with existing high-frequency patient

monitoring systems could further enhance our ability to detect and manage stress-related conditions and improve mental health outcomes.

In conclusion, our findings support the potential of speech analysis as a non-invasive, affordable, and easily accessible tool for detecting and monitoring acute stress. We hope that the developed methodology can contribute to the improvement of mental health care and overall well-being by providing new insights and tools for monitoring stress and related conditions.

7.8. References

- Abbas, A., Schultebraucks, K., & Galatzer-Levy, I. R. (2021). Digital Measurement of Mental Health: Challenges, Promises, and Future Directions. *Psychiatric Annals*, 51(1), 14–20.
<https://doi.org/10.3928/00485713-20201207-01>
- Al Hanai, T., Ghassemi, M., & Glass, J. (2018). Detecting Depression with Audio/Text Sequence Modeling of Interviews. *Interspeech 2018*, 1716–1720.
<https://doi.org/10.21437/Interspeech.2018-2522>
- Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., & Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test. *Neuroscience & Biobehavioral Reviews*, 38, 94–124.
<https://doi.org/10.1016/j.neubiorev.2013.11.005>
- Baird, A., Amiriparian, S., Cummins, N., Sturmbauer, S., Janson, J., Messner, E., Baumeister, H., & Rohleder, N. (2019). *Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test Chair of Clinical Psychology and Psychotherapy , University of Ulm , Germany*. 534–538.
- Baird, A., Triantafyllopoulos, A., Zänkert, S., Ottl, S., Christ, L., Stappen, L., Konzok, J., Sturmbauer, S., Meßner, E.-M., Kudielka, B. M., Rohleder, N., Baumeister, H., & Schuller, B. W. (2021). An evaluation of speech-based recognition of emotional and physiological markers of stress. *Frontiers in Computer Science*, 3, 750284.
<https://doi.org/10.3389/fcomp.2021.750284>
- Barik, H. C. (1977). Cross-Linguistic Study of Temporal Characteristics of Different Types of Speech Materials. *Language and Speech*, 20(2), 116–126.

<https://doi.org/10.1177/002383097702000203>

Bhatia, A., Miyatsu, T., & Pirolli, P. (2021). Towards the Development of Speech-Based Measures of Stress Response in Individuals. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 192–203.

<https://doi.org/10.18653/v1/2021.clpsych-1.21>

Bhushan, D., Kotz, K., McCall, J., Wirtz, S., Gilgoff, R., Rishi Dube, S., Powers, C., Olson-Morgan, J., Galeste, M., Patterson, K., Harris, L., Mills, A., Bethell, C., & Burke Harris, N. (2020). *The Roadmap for Resilience: The California Surgeon General's Report on Adverse Childhood Experiences, Toxic Stress, and Health*. Office of the California Surgeon General. <https://doi.org/10.48019/PEAM8812>

Biondi, M., & Picardi, A. (1999). Psychological Stress and Neuroendocrine Function in Humans: The Last Two Decades of Research. *Psychotherapy and Psychosomatics*, 68(3), 114–150. <https://doi.org/10.1159/000012323>

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, 64(9), 1089–1108. <https://doi.org/10.1002/jclp.20503>

Bou-Ghazale, S. E., & Hansen, J. H. L. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4), 429–442. <https://doi.org/10.1109/89.848224>

Boyer, S., Paubel, P.-V., Ruiz, R., El Yagoubi, R., & Daurat, A. (2018). Human Voice as a Measure of Mental Load Level. *Journal of Speech, Language & Hearing Research*, 61(11), 2722–2734. https://doi.org/10.1044/2018_JSLHR-S-18-0066

Brena, R. F., García-Vázquez, J. P., Galván-Tejada, C. E., Muñoz-Rodriguez, D., Vargas-Rosales, C., & Fangmeyer, J. (2017). Evolution of Indoor Positioning

Technologies: A Survey. *Journal of Sensors*, 2017, 1–21.

<https://doi.org/10.1155/2017/2630413>

Brooks, J. A., Tzirakis, P., Baird, A., Kim, L., Opara, M., Fang, X., Keltner, D., Monroy, M.,

Corona, R., Metrick, J., & Cowen, A. S. (2023). Deep learning reveals what vocal bursts express in different cultures. *Nature Human Behaviour*, 7(2), Article 2.

<https://doi.org/10.1038/s41562-022-01489-2>

Brosschot, J. F., Verkuil, B., & Thayer, J. F. (2017). Exposed to events that never happen:

Generalized unsafety, the default stress response, and prolonged autonomic activity.

Neuroscience & Biobehavioral Reviews, 74, 287–296.

<https://doi.org/10.1016/j.neubiorev.2016.07.019>

Burke, H. M., Davis, M. C., Otte, C., & Mohr, D. C. (2005). Depression and cortisol responses to psychological stress: A meta-analysis. *Psychoneuroendocrinology*, 30(9), 846–856.

<https://doi.org/10.1016/j.psyneuen.2005.02.010>

Câmara, R., & Griessenauer, C. J. (2015). Chapter 27—Anatomy of the Vagus Nerve. In R. S.

Tubbs, E. Rizk, M. M. Shoja, M. Loukas, N. Barbaro, & R. J. Spinner (Eds.), *Nerves and Nerve Injuries* (pp. 385–397). Academic Press.

<https://doi.org/10.1016/B978-0-12-410390-0.00028-7>

Carlson, N. R., & Birkett, M. A. (2017). *Physiology of Behavior* (12th ed.). Pearson Education Limited.

Cho, S., Fusaroli, R., Pelella, M. R., Tena, K., Knox, A., Hauptmann, A., Covello, M., Russell, A., Miller, J., Hulink, A., Uzokwe, J., Walker, K., Fiumara, J., Pandey, J., Chatham, C., Cieri, C., Schultz, R., Liberman, M., & Parish-morris, J. (2022). Identifying stable speech-language markers of autism in children: Preliminary evidence from a longitudinal

telephony-based study. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 40–46.

<https://doi.org/10.18653/v1/2022.clpsych-1.4>

Cohen, A. S., McGovern, J. E., Dinzeo, T. J., & Covington, M. A. (2014). Speech deficits in serious mental illness: A cognitive resource issue? *Schizophrenia Research*, 160(1), 173–179. <https://doi.org/10.1016/j.schres.2014.10.032>

Cohen, S., Janicki-Deverts, D., & Miller, G. E. (2007). Psychological stress and disease. *Journal of the American Medical Association*, 298(14), 1685–1687.

<https://doi.org/10.1001/jama.298.14.1685>

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., & Torre, F. D. (2009). Detecting Depression from Facial Actions and Vocal Prosody. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. <https://doi.org/10.1109/ACII.2009.5349358>

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychological Science*, 15(10), 687–693. <https://doi.org/10.1111/j.0956-7976.2004.00741.x>

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., Cipresso, P., Riva, G., & Botella, C. (2019). Current State and Future Directions of Technology-Based Ecological Momentary Assessment and Intervention for Major Depressive Disorder: A Systematic Review. *Journal of Clinical Medicine*, 8(4), 465. <https://doi.org/10.3390/jcm8040465>

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*:

From Linguistic Signal to Clinical Reality, 51–60. <https://doi.org/10.3115/v1/W14-3207>

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures.

Nature Human Behaviour, 3(4), Article 4. <https://doi.org/10.1038/s41562-019-0533-6>

Cramer, A. O. J., Borkulo, C. D. van, Giltay, E. J., Maas, H. L. J. van der, Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major Depression as a Complex Dynamic System. *PLOS ONE*, 11(12), e0167490. <https://doi.org/10.1371/journal.pone.0167490>

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, 33(2–3), 137–150. <https://doi.org/10.1017/S0140525X09991567>

de Boer, N. S., de Bruin, L. C., Geurts, J. J. G., & Glas, G. (2021). The Network Theory of Psychiatric Disorders: A Critical Assessment of the Inclusion of Environmental Factors. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.623970>

Diamond, L. M., & Alley, J. (2022). Rethinking minority stress: A social safety perspective on the health effects of stigma in sexually-diverse and gender-diverse populations.

Neuroscience & Biobehavioral Reviews, 138, 104720. <https://doi.org/10.1016/j.neubiorev.2022.104720>

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), 355–391. <https://doi.org/10.1037/0033-2909.130.3.355>

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., & Christensen, H. (2013). Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review.

Journal of Medical Internet Research, 15(11), e2791. <https://doi.org/10.2196/jmir.2791>

Duffy, J. R. (2000). Motor Speech Disorders: Clues to Neurologic Diagnosis. In C. H. Adler & J. E. Ahlskog (Eds.), *Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physician* (pp. 35–53). Humana Press.
https://doi.org/10.1007/978-1-59259-410-8_2

Eagle, N., & Pentland, A. (Sandy). (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), 255–268.
<https://doi.org/10.1007/s00779-005-0046-3>

Faurholt-Jepsen, M., Bauer, M., & Kessing, L. V. (2018). Smartphone-based objective monitoring in bipolar disorder: Status and considerations. *International Journal of Bipolar Disorders*, 6(1), 6. <https://doi.org/10.1186/s40345-017-0110-8>

Firth, J., Torous, J., Nicholas, J., Carney, R., Rosenbaum, S., & Sarris, J. (2017). Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, 218, 15–22.
<https://doi.org/10.1016/j.jad.2017.04.046>

Fletcher, J. (2007). What is heterogeneity and is it important? *BMJ: British Medical Journal*, 334(7584), 94–96. <https://doi.org/10.1136/bmj.39057.406644.68>

Frisell, T., Lichtenstein, P., Rahman, Q., & Långström, N. (2010). Psychiatric morbidity associated with same-sex sexual behaviour: Influence of minority stress and familial factors. *Psychological Medicine*, 40(2), 315–324.
<https://doi.org/10.1017/S0033291709005996>

Fromkin, V., A. (1973). Speech Errors as Linguistic Evidence. In *Speech Errors as Linguistic Evidence*. De Gruyter Mouton. <https://doi.org/10.1515/9783110888423>

- Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51, 1–26.
<https://doi.org/10.1016/j.pmcj.2018.09.003>
- Gee, B. L., Han, J., Benassi, H., & Batterham, P. J. (2020). Suicidal thoughts, suicidal behaviours and self-harm in daily life: A systematic review of ecological momentary assessment studies. *DIGITAL HEALTH*, 6, 2055207620963958.
<https://doi.org/10.1177/2055207620963958>
- Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, 13(1), 440–460. <https://doi.org/10.1109/TAFFC.2019.2927337>
- 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.e21–390.e29.
<https://doi.org/10.1016/j.jvoice.2012.12.010>
- Godin, K. W., & Hansen, J. H. L. (2015). Physical task stress and speaker variability in voice quality. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015(1).
<https://doi.org/10.1186/s13636-015-0072-7>
- 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. https://doi.org/10.1007/978-3-540-74200-5_6
- Hemmeter, U.-M., Hemmeter-Sernal, J., & Krieg, J.-C. (2010). Sleep deprivation in depression. *Expert Review of Neurotherapeutics*, 10(7), 1101–1115.
<https://doi.org/10.1586/ern.10.83>

- Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience & Biobehavioral Reviews*, 26(2), 235–258. [https://doi.org/10.1016/S0149-7634\(01\)00068-9](https://doi.org/10.1016/S0149-7634(01)00068-9)
- Juster, R.-P., McEwen, B. S., & Lupien, S. J. (2010). Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, 35(1), 2–16. <https://doi.org/10.1016/j.neubiorev.2009.10.002>
- Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54(2), 910–921. <https://doi.org/10.3758/s13428-021-01670-x>
- Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van Hoecke, S., & Vanderhasselt, M.-A. (2022). Acoustic speech features in social comparison: How stress impacts the way you sound. *Scientific Reports*, 12(1), Article 1. <https://doi.org/10.1038/s41598-022-26375-9>
- Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. (2023). Speech as a promising biosignal in precision psychiatry. *Neuroscience & Biobehavioral Reviews*, 148, 105121. <https://doi.org/10.1016/j.neubiorev.2023.105121>
- Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier social stress test'—A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1–2), 76–81. <https://doi.org/10.1159/000119004>
- König, A., Riviere, K., Linz, N., Lindsay, H., Elbaum, J., Fabre, R., Derreumaux, A., & Robert, P. (2021). Measuring Stress in Health Professionals Over the Phone Using Automatic Speech Analysis During the COVID-19 Pandemic: Observational Pilot Study. *Journal of Medical Internet Research*, 23(4), e24191. <https://doi.org/10.2196/24191>
- Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2021).

Speech as a Biomarker for Depression. *CNS & Neurological Disorders - Drug Targets*,

20. <https://doi.org/10.2174/1871527320666211213125847>

Kurniawan, H., Maslov, A. V., & Pechenizkiy, M. (2013). Stress Detection from Speech and

Galvanic Skin Response Signals. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 209–214.

Langer, M., König, C. J., Siegel, R., Fredenhagen, T., Schunck, A. G., Hähne, V., & Baur, T.

(2022). Vocal-stress diary: A longitudinal investigation of the association of everyday work stressors and human voice features. *Psychological Science*, 33(7), 1027–1039.

<https://doi.org/10.1177/09567976211068110>

Lara, O. D., & Labrador, M. A. (2013). A Survey on Human Activity Recognition using Wearable

Sensors. *IEEE Communications Surveys & Tutorials*, 15(3), 1192–1209.

<https://doi.org/10.1109/SURV.2012.110112.00192>

Li, J., Wang, S., Chao, Y., Liu, X., & Meng, H. (2022). Context-aware Multimodal Fusion for

Emotion Recognition. *Interspeech 2022*, 2013–2017.

<https://doi.org/10.21437/Interspeech.2022-10592>

Li, N., Li, N., Guo, M., & Feng, J. (2021). Research of Speech Biomarkers for Stress

Recognition Using Linear and Nonlinear Features. *2021 7th International Conference on*

Computer and Communications (ICCC), 509–513.

<https://doi.org/10.1109/ICCC54389.2021.9674330>

LiKamWa, R., Liu, Y., Lane, N. D., & Zhong, L. (2011). *Can Your Smartphone Infer Your Mood?*

1–5.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric

disorders using speech: A systematic review. *Laryngoscope Investigative*

Otolaryngology, 5(1), 96–116. <https://doi.org/10.1002/lio2.354>

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., & Choudhury, T. (2012). StressSense: Detecting stress in unconstrained acoustic environments using smartphones. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, 351–360. <https://doi.org/10.1145/2370216.2370270>

Manuck, S. B., Cohen, S., Rabin, B. S., Muldoon, M. F., & Bachen, E. A. (1991). Individual Differences in Cellular Immune Response to Stress. *Psychological Science*, 2(2), 111–115.

Marmar, C. R., Siegel, C., Brown, A. D., Laska, E., Richey, C., Amara, D. A., Smith, J., Knoth, B., & Tsartas, A. (2019). *Speech - based markers for posttraumatic stress disorder in US veterans*. December 2018, 607–616. <https://doi.org/10.1002/da.22890>

Martin, V. P., Rouas, J.-L., Boyer, F., & Philip, P. (2021). Automatic Speech Recognition Systems Errors for Objective Sleepiness Detection Through Voice. *Interspeech 2021*, 2476–2480. <https://doi.org/10.21437/Interspeech.2021-291>

Martin, V. P., Rouas, J.-L., Thivel, P., & Krajewski, J. (2019). Sleepiness detection on read speech using simple features. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1–7. <https://doi.org/10.1109/SPED.2019.8906577>

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin*, 129, 674–697. <https://doi.org/10.1037/0033-2909.129.5.674>

Monkhouse, S. (2005). *Cranial Nerves: Functional Anatomy*. Cambridge University Press.

- Moore, R. C., Depp, C. A., Wetherell, J. L., & Lenze, E. J. (2016). Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. *Journal of Psychiatric Research*, 75, 116–123. <https://doi.org/10.1016/j.jpsychires.2016.01.011>
- Mote, J., & Fulford, D. (2020). Ecological momentary assessment of everyday social experiences of people with schizophrenia: A systematic review. *Schizophrenia Research*, 216, 56–68. <https://doi.org/10.1016/j.schres.2019.10.021>
- Parry, J., DeMattos, E., Klementiev, A., Ind, A., Morse-Kopp, D., Clarke, G., & Palaz, D. (2022). Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning. *Interspeech 2022*, 1158–1162. <https://doi.org/10.21437/Interspeech.2022-10581>
- Paulmann, S., Furnes, D., Bøkenes, A. M., & Cozzolino, P. J. (2016). How Psychological Stress Affects Emotional Prosody. *PLOS ONE*, 11(11), e0165022. <https://doi.org/10.1371/journal.pone.0165022>
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1), 547–577. <https://doi.org/10.1146/annurev.psych.54.101601.145041>
- Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The Network Approach to Psychopathology: A Review of the Literature 2008–2018 and an Agenda for Future Research. *Psychological Medicine*, 50(3), 353–366. <https://doi.org/10.1017/S0033291719003404>
- Rodellar-Biarge, V., Palacios-Alonso, D., Nieto-Lluis, V., & Gómez-Vilda, P. (2015). Towards the search of detection in speech-relevant features for stress. *Expert Systems*, 32(6), 710–718. <https://doi.org/10.1111/exsy.12109>

- Rude, S. S., Gortner, E., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), 1121–1133.
<https://doi.org/10.1080/02699930441000030>
- Serre, F., Fatseas, M., Swendsen, J., & Auriacombe, M. (2015). Ecological momentary assessment in the investigation of craving and substance use in daily life: A systematic review. *Drug and Alcohol Dependence*, 148, 1–20.
<https://doi.org/10.1016/j.drugalcdep.2014.12.024>
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 4(1), 1–32.
<https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
- Shuman, V., Scherer, K., Fontaine, J., & Soriano, C. (2015). *The GRID meets the Wheel: Assessing emotional feeling via self-report.* <https://doi.org/10.13140/RG.2.1.2694.6406>
- Slavich, G. M., & Irwin, M. R. (2014). From Stress to Inflammation and Major Depressive Disorder: A Social Signal Transduction Theory of Depression. *Psychological Bulletin*, 140(3), 774–815. <https://doi.org/10.1037/a0035302>
- Slavich, G. M., Roos, L. G., Mengelkoch, S., Webb, C. A., Shattuck, E. C., Moriarity, D. P., & Alley, J. C. (2023). Social Safety Theory: Conceptual foundation, underlying mechanisms, and future directions. *Health Psychology Review*, 1–55.
<https://doi.org/10.1080/17437199.2023.2171900>
- Slavich, G. M., Taylor, S., Picard, R. W., Slavich, G. M., Taylor, S., & Stress, R. W. P. (2019). *Stress measurement using speech: Recent advancements , validation issues , and ethical and privacy considerations.* 3890.
<https://doi.org/10.1080/10253890.2019.1584180>

- Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Torous, J., Friedman, R., & Keshavan, M. (2014). Smartphone Ownership and Interest in Mobile Applications to Monitor Symptoms of Mental Health Conditions. *JMIR MHealth and UHealth*, 2(1), e2994. <https://doi.org/10.2196/mhealth.2994>
- Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, 9, 1994. <https://doi.org/10.3389/fpsyg.2018.01994>
- Voppel, A. E., de Boer, J. N., Brederoo, S. G., Schnack, H. G., & Sommer, I. E. C. (2022). *Semantic and phonetic markers in schizophrenia-spectrum disorders; a combinatorial machine learning approach* [Preprint]. Psychiatry and Clinical Psychology. <https://doi.org/10.1101/2022.07.13.22277577>
- Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12. <https://doi.org/10.1016/j.wocn.2014.11.001>
- Walton, G. M., & Cohen, G. L. (2011). A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students. *Science*, 331(6023), 1447–1451. <https://doi.org/10.1126/science.1198364>
- Wang, M., & Saudino, K. J. (2011). Emotion regulation and stress. *Journal of Adult Development*, 18(2), 95–103. <https://doi.org/10.1007/s10804-010-9114-7>
- Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, 44(7), 1349–1360. <https://doi.org/10.1017/S0033291713001979>

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3), 329–336.

<https://doi.org/10.1016/j.wocn.2010.04.003>

Zhu-Zhou, F., Gil-Pita, R., García-Gómez, J., & Rosa-Zurera, M. (2022). Robust Multi-Scenario

Speech-Based Emotion Recognition System. *Sensors*, 22(6), 2343.

<https://doi.org/10.3390/s22062343>

Personal Contributions

Chapter 2:

Mitchel Kappen: Conceptualization, Methodology, Formal analysis, Data Curation, Writing-Original Draft, Visualization **Kristof Hoorelbeke:** Conceptualization, Methodology, Formal analysis, Data Curation, Writing - Review & Editing, Visualization **Nilesh Madhu:** Conceptualization, Methodology, Resources, Writing - Review & Editing **Kris Demuynck:** Conceptualization, Resources, Writing - Review & Editing **Marie-Anne Vanderhasselt:** Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition

Chapter 3:

Mitchel Kappen: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing- Original Draft, Visualization **Jonas van der Donckt:** Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft **Gert Vanhollebeke:** Conceptualization, Methodology, Investigation, Data Curation, Writing - Review & Editing **Jens Allaert:** Methodology, Writing - Review & Editing **Vic Degraeve:** Software, Writing - Review & Editing **Nilesh Madhu:** Conceptualization, Resources, Writing - Review & Editing **Sofie van Hoecke:** Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition **Marie-Anne Vanderhasselt:** Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition

Chapter 4:

Jonas Van Der Donckt: Conceptualization, Methodology, Software, Validation, Formal Analysis, Data Curation, Writing - Original Draft, Visualization. **Mitchel Kappen:** Conceptualization, Methodology, Validation, Formal Analysis, Data Curation, Writing - Original

Draft, Visualization. **Vic Degraeve:** Software. **Kris Demuynck:** Conceptualization, Writing - Review & Editing. **Marie-Anne Vanderhasselt:** Conceptualization, Methodology, Supervision, Writing - Review & Editing. **Sofie Van Hoecke:** Conceptualization, Methodology, Supervision, Writing - Review & Editing, Funding acquisition.

Chapter 5:

Mitchel Kappen: Conceptualization, Methodology, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **Gert Vanhollebeke:** Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Review & Editing **Jonas van der Donckt:** Conceptualization, Methodology, Formal Analysis, Data Curation, Writing - Review & Editing **Sofie van Hoecke:** Conceptualization, Writing - Review & Editing, Supervision **Marie-Anne Vanderhasselt:** Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Chapter 6:

Mitchel Kappen: Conceptualization, Writing - Original Draft, Funding acquisition **Marie-Anne Vanderhasselt** Conceptualization, Writing - Review & Editing, Funding acquisition **George Slavich:** Conceptualization, Writing- Original Draft, Supervision

Curriculum Vitae

Mitchel Kappen MSc

HOME: Lange Kruisstraat 12-1
9000, Ghent, Belgium

Tel: +316 55 68 44 63
E-Mail: mitchelkappen@gmail.com

UNIVERSITY: Ghent University
Experimental Psychiatry Lab
Department of Head & Skin
+32493 75 64 96
Mitchel.Kappen@UGent.be

RESEARCH INTERESTS:

My current research is centered around novel methods of measuring physiological and psychological activity. More specifically, detecting mental states (predominantly stress) from speech fragments or facial video recordings.

EDUCATION:

2019 – 2023 PhD in Life Sciences
Ghent University – Experimental Psychiatry Lab, Belgium
Supervised by Professor Marie-Anne Vanderhasselt
Due for submission September 2023

Projects:

- Detecting stress from speech fragments. Developing new paradigms to record speech in a naturalistic way in a controlled setting whilst simultaneously collecting high-quality physiological data and querying psychological constructs to develop models using phonetic and prosodic features of speech.
- Relapse prevention after ECT using CCT for treatment-resistant depression. Relapse after a successful electroconvulsive therapy happens at high prevalence. In our RCT we offer a two-week cognitive intervention after ECT is concluded after which patients are biweekly monitored over a 6-month time period.
- Stress, rumination, and (im/ex)plicit emotion regulation during the different phases of the menstrual cycle in women with premenstrual syndrome and healthy controls.

Premenstrual symptoms occur in more than 60% of women, however, its origin is still debated. We conducted an online study with 500+ participants, consisting of both women with PMS and healthy controls, and conducted questionnaires and an experiment consisting of rating emotional stimuli whilst their faces are recorded. This enables us to compare one's capabilities of introspective capabilities with regard to their emotions as well as group differences in women with PMS as compared to healthy controls.

2018 – 2019 MSc in Applied Cognitive Psychology (Magna Cum Laude)

Utrecht University, the Netherlands

Thesis (9/10): Predicting Person-Organization and Person-Job Fit Objectively: Stress, motivation, and Nervousness During a Video-Based Pre-hire screening
Supervisor: Prof. Marnix Naber

2013 – 2018 BSc in Cognitive and Neurobiological Psychology

Utrecht University, the Netherlands

Thesis (8.5/10): Are Visuo-spatial Working Memory and Fragile Memory Qualitatively Different Memory Systems?

Supervisor: Paul Zerr

2006 – 2012 Bilingual grammar school

International Baccalaureate English Higher Level (Near Native)
Elde College, Schijndel, the Netherlands

RESEARCH EXPERIENCE (IN ADDITION TO EDUCATION):

Oct 2022 - Dec 2022 UCLA Laboratory for Stress Assessment and Research, Stanford University, United States

As part of my PhD I visited prof. dr. George M. Slavich at Stanford University. During this stay, I expanded my knowledge of lifetime stressors and adverse childhood experiences (ACEs). We wrote a viewpoint on speech in precision psychiatry and started a multi-year collaboration on a state-wide research project on ACEs, precision psychiatry interventions, and a first endeavor to investigate whether speech contains information on chronic or early life stressors.

Feb 2019 - Sep 2019 Alpha.One / Expoze.io, Rotterdam, the Netherlands

I worked here as a junior researcher in which I was responsible for setting up online experiments enabling us to test implicit behaviors as well as collect training data for our predictive eye tracking models. In addition, collected and analyzed EEG data, all tasks were ordered by the Erasmus school of Management and big commercial companies.

Sep 2018 – Feb 2019 **Neurolytics, Utrecht, the Netherlands**

In this start-up, I helped develop the testing environment and first models in predicting the match between an employer and an applicant. Using numerous metrics such as human resource questionnaires, facial coding, and remote PPG.

Aug 2017 - July 2018 **Krigolson Lab, University of Victoria, Canada**

While I was visiting the University of Victoria for an exchange semester, I volunteered at the Krigolson Lab as a research assistant. I gained experience with conducting and analyzing EEG, eye-tracking, and physiological data. After graduation, I stayed for another semester to commit full-time to set up my own project at this lab.

ADDITIONAL WORK HISTORY:

2012 – 2019 **Jobs next to studies**

I have had numerous jobs in sales, recruitment, administration, and hospitality in the evening and weekend hours varying from 10 to 32 hours a week.

In addition, I worked for 5 years at a smartphone and computer refurbishing company doing customer support, sales, and search engine optimization, as well as assisting in repairs.

2012 – 2022 **Extracurriculars**

Participated in numerous committees organizing extracurricular activities. Took one full-time board year (student faculty association) in which I was responsible for acquisition, PR, and team coordination as well as a part-time board year (Utrecht University fund) in which my job was assessing grants for student initiatives.

Furthermore, I acted as the vice president of the UVic Bitcoin Club and have participated in scientific outreach activities (Let's Talk Science, Canada) teaching elementary school children about the brain and organizing symposia on neuroscience for high school students. Moreover, I was a founding committee member for the multi-university 2-day workshop on Machine Learning in Psychiatry.

SKILLS:

Languages

- Dutch (mother tongue)
- English (near-native)
- German (basic)

Software & Programming (<https://github.com/mitchelkappen>)

- Python
- R
- Matlab
- JavaScript
- Machine Learning
- Brainvision Analyser & EEGLab
- Photoshop
- InDesign

Research

- Remote measures
 - Speech analysis (acoustic + semantic), rPPG, (automated) FACS, portable EEG systems (e.g. MUSE)
- ECG
- EDA
- Respiration
- EEG (Brainvision, BioSemi, Muse, EGI)
- Eyetracking (EyeLink 2, EyeLink 1000, Webcam tracking)

REFERENCES:

Professor Marie-Anne Vanderhasselt, Ghent University, Belgium, marie-anne.vanderhasselt@ugent.be
Professor Olave Krigolson, University of Victoria, Canada, kriegolso@uvic.ca
Professor Marnix Naber, Utrecht University, the Netherlands, m.naber@uu.nl

PUBLICATIONS:

Vanhollebeke, G., **Kappen, M.**, De Raedt, R., Baeken, C., van Mierlo, P., & Vanderhasselt, M. A. (2023). Effects of acute psychosocial stress on source level EEG power and functional connectivity measures. *Scientific Reports*, 13(1), 8807.

Kuipers, M., **Kappen, M.**, & Naber, M. (2023). How nervous am I? How computer vision succeeds and humans fail in interpreting state anxiety from dynamic facial behaviour. *Cognition and Emotion*, 1-11.

Kappen, M., Vanderhasselt, M. A., & Slavich, G. M. (2023). Speech as a Promising Biosignal in Precision Psychiatry. *Neuroscience & Biobehavioral Reviews*, 105121.

De Smet, S., Ottaviani, C., Verkuil, B., **Kappen, M.**, Baeken, C., & Vanderhasselt, M. A. (2023). Effects of non-invasive vagus nerve stimulation on cognitive and autonomic correlates of perseverative cognition. *Psychophysiology*, e14250.

Razza, L. B., Luethi, M. S., Zanão, T., De Smet, S., Buchpigel, C., Busatto, G., Pereira, J., Klein, I., **Kappen, M.**, Morena, M., Baeken, C., Vanderhasselt, M. A., & Brunoni, A. R. (2023). Transcranial direct current stimulation versus intermittent theta-burst stimulation for the improvement of working memory performance. *International Journal of Clinical and Health Psychology*, 23(1), 100334.

Kappen, M., Raeymakers, S., Weyers, S., & Vanderhasselt, M. A. (2022). Stress and Rumination in Premenstrual Syndrome (PMS): identifying stable and menstrual cycle-related differences in PMS symptom severity. *Journal of Affective Disorders*. Preprint available: <https://psyarxiv.com/nhvb2>

Xu, Y., **Kappen, M.**, Peremans, K., De Bundel, D., Van Eeckhaut, A., Van Laeken, N., De Vos, F., Dobbeleir, A., Saunders, J. H., & Baeken, C. (2022). Accelerated HF-rTMS Modifies SERT Availability in the Subgenual Anterior Cingulate Cortex: A Canine [11C] DASB Study on the Serotonergic System. *Journal of clinical medicine*, 11(6), 1531.

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M. A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, 54(2), 910-921.

Kappen, M., & Naber, M. (2021). Objective and bias-free measures of candidate motivation during job applications. *Scientific reports*, 11(1), 1-8.

Zerr, P., Gayet, S., van den Esschert, F., **Kappen, M.**, Olah, Z., & Van der Stigchel, S. (2021). The development of retro-cue benefits with extensive practice: Implications for capacity estimation and attentional states in visual working memory. *Memory & Cognition*, 1-14.

Van de Velde, N., **Kappen, M.**, Koster, E. H., Hoorelbeke, K., Tandt, H., Verslype, P., ... & Vanderhasselt, M. A. (2020). Cognitive remediation following electroconvulsive therapy in patients with treatment resistant depression: randomized controlled trial of an intervention for relapse prevention—study protocol. *BMC psychiatry*, 20(1), 1-12.

Williams, C. C., **Kappen, M.**, Hassall, C. D., Wright, B., & Krigolson, O. E. (2019). Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage*, 189, 574-580.

In review:

Kappen, M., Vanhollebeke, G., Van Der Donckt, J., Van Hoecke, S., & Vanderhasselt, M. A. (2023). Acoustic and Prosodic Speech Features Reflect Physiological Stress but Not Isolated Negative Affect: A Multi-paradigm Study on Psychosocial Stressors.

Van Der Donckt, J.*, **Kappen, M.***, Degraeve, V., Demuynck, K., Vanderhasselt, M. A., & Van Hoecke, S. (2023). Ecologically Valid Speech Collection in Behavioral Research: The Ghent Semi-spontaneous Speech Paradigm (GSSP).

Li, Z., Pulopulos, M., Allaert, J., De Smet, S., **Kappen, M.**, Puttevils, L., ... & Vanderhasselt, M. A. (2022). Resting HRV as a trait marker of rumination in healthy individuals? A large cross-sectional analysis. *Authorea Preprints*.

Oral presentations:

Kappen, M., Van der Donckt, J., Vanhollebeke, G., Van Hoecke, S., Vanderhasselt, M.A. (2022, September). How your Speech Responds to Stress: the Validation of Acoustic, Prosodic, and Semantic Speech Features in a Multi-Paradigm Stress-Induction Task, Society for Psychophysiological Research, Vancouver, BC, Canada.

Kappen, M., Vanhollebeke, G., Van der Donckt, J., Coquyt, I., Van Hoecke, S., & Vanderhasselt, M.A. (2022, June). The Effects of Stress on the Voice: Acoustic Features from Semi-Spontaneous Speech in a Multi-Paradigm Stress Induction Task, 2022 Annual Meeting of the Belgian Association of Psychological Sciences, Leuven, Belgium.

Kappen, M., Kuipers, M., & Naber, M.M. (2022, April). Where Computers Outperform Humans: Objective and Bias-Free Measures of Complex Emotions and Mental States using Facial Nonverbal Behavior, 18th NVP Winter Conference on Brain and Cognition, Egmond aan Zee, the Netherlands.

Naber, M. M., Kuipers, M., & **Kappen, M.** (2021, December). Interpreting facial features to determine an observer's attention to a video. In PERCEPTION (Vol. 50, No. 1_SUPPL, pp. 97-97).

Kappen, M., Hassall, C.D., & Krigolson, O.E. (2018). Electroencephalographic Correlates for Risk Taking and Aversion in Financial Decision Making. University of Victoria's Making Waves, Victoria, BC, Canada.

Poster presentations:

Kappen, M., Van der Donckt, J., Vanhollebeke, G., Van Hoecke, S., Vanderhasselt, M.A. (2022, September). How your Speech Responds to Stress: the Validation of Acoustic, Prosodic, and Semantic Speech Features in a Multi-Paradigm Stress-Induction Task, Society for Psychophysiological Research, Vancouver, BC, Canada.

Kappen, M., DeSmet, S., Allaert, J., Schoonjans, E., VanderDonckt, J., Raeymakers, S., & Vanderhasselt, M. A. (2022). The Interaction of Transcranial Direct Current Stimulation (tDCS) and Pace Breathing on Acoustic and Lexical Speech Features in the Context of Stress. *Psychiatria Danubina*, 34(suppl 3), 35-35.

Xu, Y., Kappen, M., Peremans, K., DeBundel, D., VanEeckhaut, A., VanLaeken, N., De Vos, F., Dobbelei, A., Saunders, J. H., & Baeken, C. (2022). Sert Availability Modified by Accelerated HF-rTMS in the Subgenual Anterior Cingulate Cortex: a Canine [11C]-DASB Positron Emission Tomography Study. *Psychiatria Danubina*, 34(suppl 3), 44-44.

Naber, M., & Kappen, M. (2021). How motivated do I look? How humans fail and computer vision succeeds in interpreting facial behavior. *Journal of Vision*, 21(9), 1978-1978.

Williams, C.C., Kappen, M., Hassall, C.D., Wright, B., & Krigolson, O.E. (2018). Cognitive Control and Attention: Neurocognitive Mechanisms of System 1 and System 2 Thinking. Society for Psychophysiological Research Meeting, Quebec City, QC.

Kappen, M., Hassall, C.D., & Krigolson, O.E. (2018). Electroencephalographic Correlates for Risk and Ambiguity in Financial Decision Making. Canadian Neuroscience Annual Meeting, Vancouver, BC, Canada.

Kappen, M., Hassall, C.D., & Krigolson, O.E. (2018). Neurophysiological Representations of Risk Taking and Risk Aversion. Northwest Cognition and Memory 2018, Richmond, BC, Canada.

Powell, G., Kappen, M., Berman, T., Colino, F.L., & Krigolson, O.E. (2018). The Effect of Feedback Frequency on the P300 for Motor Learning. Northwest Cognition and Memory 2018, Richmond, BC, Canada.

MEDIA:

Libelle (2022, September 14). Wat vertelt je menstruele cyclus over je gezondheid? De expert legt uit. <https://www.libelle.be/gezond/cyclus-en-gezondheid/>

Goed gevoel, DPG media (2022, July 20). Had ik dat maar eerder geweten! Physical print.

Knack (2021, January 6). Het premenstrueel syndroom is nog steeds een onbekende problematiek. <https://www.knack.be/nieuws/gezondheid/het-premenstrueel-syndroom-is-nog-steeds-een-onbekende-problematiek/>

EOS Wetenschap (2020, December 10). 'Ik krijg vaak te horen dat ik overdrijf'. <https://www.eoswetenschap.eu/psyche-brein/ik-krijg-vaak-te-horen-dat-ik-overdrijf>

VRT Media, Radio 2 (2020, October 30). Wat doet een menstruatiecyclus met het hoofd van de vrouw? UGent onderzoekt het met gezichtsanalyse. <https://www.vrt.be/vrtnws/nl/2020/10/30/wat-doet-een-menstruatiecyclus-met-het-hoofd-van-de-vrouw-ugent/>

FOLLOW ME:

Website: mitchelkappen.github.io

Google Scholar: <https://scholar.google.nl/citations?user=CWCQD9UAAAJ&hl=en&oi=ao>

Research Gate: <https://www.researchgate.net/profile/Mitchel-Kappen>

Twitter: <https://twitter.com/KappenMitchel>

Github: <https://github.com/mitchelkappen>

OSF: <https://osf.io/4xet9>

Bio: <https://www.gheplab.ugent.be/labmembers/mitchel-kappen/>

Projects:

<https://www.gheplab.ugent.be/projects/stress-speech>

<https://www.gheplab.ugent.be/projects/menstrual-cycle-info/>