Universität
Augsburg
University

Assistive Augmentation of Cognitive Processes Using Mobile Signal Processing

Michael Dietz

Dissertation zur Erlangung des Doktorgrades

Doctor rerum naturalium

(Dr. rer. nat.)

Fakultät für Angewandte Informatik

Institut für Informatik

Universität Augsburg

2025

Universität
Augsburg
University

Erstgutachterin:	Prof. Dr. Elisabeth André
Zweitgutachter:	Prof. Dr. Björn Schuller
Tag der mündlichen Prüfung:	10.03.2025

Abstract

Cognitive processes are the foundation for all our abilities, thoughts, and actions. They enable us to perceive our surroundings, remember details about past events, retain large amounts of knowledge, and work out solutions for complex problems. While these processes are typically performed almost effortlessly and without much conscious awareness, we only realize their importance when they no longer function as expected. In these cases, impairments and disorders of cognitive processes can severely impact the lives of affected individuals and their families. Potential consequences can range from difficulties performing everyday tasks to the complete loss of personal independence. However, supporting these conditions can be rather challenging since they often manifest in different ways and require specialized solutions tailored to the specific circumstances of individuals, which can be expensive and time-consuming.

Fortunately, recent technological advancements of mobile and wearable devices present a promising opportunity to build assistive applications that can help mitigate the impact of cognitive impairments and disorders. With their ubiquitous presence, integrated sensors, and increasing computational power, these devices offer an ideal platform to provide timely and context-aware assistance. Consequently, this thesis investigates how their capabilities can be utilized to address the challenges associated with supporting cognitive processes. To this end, we follow the assistive augmentation paradigm, which advocates for developing technology that considers the needs and circumstances of individuals. Based on a thorough conceptual analysis of previous works to identify shared properties and strategies, common design dimensions are derived to classify existing approaches and guide the development of future applications.

Informed by these conceptual findings, a flexible and easily usable software framework is proposed that combines the growing capabilities of mobile devices with advanced signal processing techniques. It enables the rapid prototyping and implementation of assistive augmentation systems to support individuals affected by impairments and disorders of cognitive processes. Additionally, this thesis demonstrates the feasibility and evaluates the effectiveness of the proposed conceptual and technical solutions with empirical research probes in real-world settings. Each of them illustrates the complete workflow of designing, implementing, prototyping, and testing assistive augmentation systems with the proposed framework and showcases how to support various mental conditions, including visual impairment, memory decline, and cognitive disorders.

Zusammenfassung

Kognitive Prozesse bilden die Grundlage für alle unsere Gedanken, Fähigkeiten und Handlungen. Sie ermöglichen uns die Wahrnehmung der Umgebung, die Erinnerung an Details vergangener Ereignisse, die Speicherung großer Mengen von Wissen und die Lösung komplexer Probleme. Während diese Prozesse in der Regel fast mühelos und größtenteils unterbewusst ablaufen, wird uns ihre Bedeutung erst bewusst, sobald sie nicht mehr wie erwartet funktionieren. In diesen Fällen können kognitive Störungen und Erkrankungen das Leben von Betroffenen und ihrer Familien sehr stark beeinträchtigen. Mögliche Folgen reichen von Schwierigkeiten bei der Bewältigung alltäglicher Aufgaben bis hin zum vollständigen Verlust der persönlichen Unabhängigkeit. Die Unterstützung von betroffenen Personen kann sich jedoch als relativ schwierig erweisen, da die Erkrankungen häufig mit unterschiedlichen Symptomen auftreten und spezialisierte Lösungen erfordern, deren Entwicklung zeitaufwendig und kostspielig sein kann.

Allerdings stellen die technologischen Fortschritte der vergangenen Jahre im Bereich mobiler Endgeräte und tragbarer Sensoren eine vielversprechende Grundlage für die Entwicklung unterstützender Anwendungen dar, die den negativen Auswirkungen kognitiver Störungen und Beeinträchtigungen entgegenwirken können. Dank ihrer kontinuierlich steigenden Rechenleistung, allgegenwärtigen Verfügbarkeit und integrierten Sensorik bilden sie eine ideale Plattform für die zeitnahe Bereitstellung kontextbezogener Assistenz. Im Rahmen dieser Arbeit wird daher untersucht, wie die Funktionen und Ressourcen mobiler Geräte genutzt werden können, um die mit der Unterstützung von kognitiven Prozessen verbundenen Herausforderungen zu bewältigen. Zu diesem Zweck wird das Prinzip der assistiven Augmentierung angewendet, welches sich für die Berücksichtigung der Anforderungen und Bedürfnisse von Individuen bei der Entwicklung neuer Technologien einsetzt. Basierend auf einer umfangreichen konzeptionellen Analyse vorangegangener Arbeiten zur Identifizierung gemeinsamer Eigenschaften und Strategien werden allgemeine Dimensionen abgeleitet, um bestehende Ansätze zu klassifizieren und die Entwicklung künftiger Anwendungen zu erleichtern.

Aufbauend auf diesen konzeptionellen Erkenntnissen wird ein flexibles und einfach zu verwendendes Software-Framework vorgestellt, welches die zunehmenden Fähigkeiten mobiler Endgeräte mit fortschrittlichen Signalverarbeitungsverfahren kombiniert. Neben der zeitnahen Erstellung von Prototypen ermöglicht es die Implementierung vollwertiger Augmentierungssysteme zur Unterstützung von Personen mit kognitiven Ein-

schränkungen und Erkrankungen. Darüber hinaus wird im Rahmen dieser Arbeit die praktische Anwendbarkeit und Effektivität der vorgestellten konzeptionellen und technischen Lösungen anhand von empirischen Studien mit betroffenen Personen in natürlichen Umgebungen evaluiert. Jede dieser Forschungsstichproben veranschaulicht dabei den vollständigen Ablauf von der Konzeption über die Prototypenerstellung und Implementierung bis hin zur Validierung der entwickelten assistiven Augmentierungssysteme und demonstriert, wie das vorgestellte Framework zur Unterstützung verschiedener kognitiver Bedingungen, einschließlich Sehbehinderungen, Gedächtnisstörungen und psychischer Erkrankungen verwendet werden kann.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Dr. Elisabeth André for her continued guidance and support throughout my time at her lab. She enabled me to freely pursue my research interests and provided valuable advice that shaped this dissertation and improved my work as a researcher. Special thanks also go to Prof. Dr. Björn Schuller and Prof. Dr. Wolfgang Minker for taking the time out of their busy schedules and agreeing to review this thesis.

Furthermore, I am grateful to all my current and former colleagues for the friendly and respectful atmosphere that made working at the lab such a pleasant experience. I am especially thankful to Ionut Damian for his mentoring during my early years in academia and his work on the SSJ framework. Similarly, I would like to thank Dominik Schiller, Florian Lingenfelser, Tobias Baur, Björn Petrak, Hannes Ritschel, Simon Flutura, Andreas Seiderer, Alexander Heimerl, Tobias Huber, Silvan Mertes, Ruben Schlagowski, Daniel Schork, Chi Tai Dang, and Ilhan Aslan for the collaboration on various projects, fruitful discussions, and enjoyable time we shared inside and outside the lab.

Last but not least, I would like to express my deepest gratitude to my parents for their unconditional love and support throughout my entire life and for always believing in me on every step of the way. Thank you all!

Contents

1 Introduction **1**

1.1 Research Objectives 4

1.2 Thesis Outline 7

I Foundations

2 Cognitive Processes **11**

2.1 Perception 13

2.1.1 Sensation 14

2.1.2 Attention 15

2.1.3 Disorders 18

2.2 Memory Storage 19

2.2.1 Sensory Memory 21

2.2.2 Short-term Memory 22

2.2.3 Long-term Memory 24

2.2.4 Disorders 26

2.3 Higher-Order Cognition 27

2.3.1 Language 29

2.3.2 Problem-Solving 33

2.3.3 Emotions 35

2.3.4 Disorders 38

2.4 Summary 39

3 Non-Verbal Signals **41**

3.1 Behavioral Cues 42

3.1.1 Facial Expressions 44

3.1.2 Gaze Behavior . . . 46
3.1.3 Vocal Cues . . . 47
3.1.4 Gestures and Posture . . . 48
3.2 Physiological Cues . . . 50
3.2.1 Heart Activity . . . 51
3.2.2 Electrodermal Activity . . . 52
3.2.3 Brain Activity . . . 53
3.2.4 Muscle Activity . . . 54
3.3 Summary . . . 55

II Concept & Implementation

4 Assistive Augmentation 59
4.1 Theoretical Context . . . 61
4.2 Literature Analysis . . . 67
4.2.1 Methodology . . . 67
4.2.2 Findings . . . 69
4.3 Strategies . . . 72
4.3.1 Sensory Augmentation . . . 72
4.3.2 Memory Augmentation . . . 85
4.3.3 Cognitive Augmentation . . . 90
4.4 Design Dimensions . . . 94
4.4.1 Target . . . 95
4.4.2 Initiative . . . 97
4.4.3 Presence . . . 99
4.4.4 Direction . . . 101
4.4.5 Adaptation . . . 102
4.5 Summary . . . 104

5 Mobile Signal Processing 107
5.1 Challenges . . . 108
5.1.1 Privacy Concerns . . . 108
5.1.2 Technical Limitations . . . 110
5.1.3 Ethical Considerations . . . 112
5.2 Hardware Selection . . . 113
5.3 Data Collection . . . 116

5.3.1 Annotation . . . 117
5.3.2 Existing Datasets . . . 119
5.4 Model Training . . . 121
5.4.1 Feature-based Machine Learning . . . 121
5.4.2 End-to-End Deep Learning . . . 122
5.5 Summary . . . 125

6 The SSJ Framework 127
6.1 Origins and Basic Concepts . . . 129
6.1.1 Modular Design . . . 129
6.1.2 Sampling . . . 130
6.1.3 Generic Data Handling . . . 131
6.1.4 Synchronization . . . 133
6.2 Existing Solutions . . . 133
6.3 Architecture . . . 137
6.3.1 Providers . . . 140
6.3.2 Transformers . . . 142
6.3.3 Consumers . . . 144
6.3.4 Events . . . 145
6.4 Graphical Interface . . . 147
6.4.1 Technical Foundation . . . 147
6.4.2 Feature Overview . . . 148
6.5 Example Application . . . 150
6.6 Summary . . . 153

III **Research Probes**

7 Assisting Visual Impairment 157
7.1 Augmentation Design . . . 159
7.2 System Overview . . . 161
7.2.1 Color Sonification . . . 163
7.2.2 Text Sonification . . . 165
7.3 Evaluation . . . 167
7.3.1 Experiments . . . 168
7.3.2 Results . . . 169
7.4 Discussion . . . 171

7.5 Summary . . . 172

8 Assisting Memory Decline 173
8.1 Augmentation Design . . . 175
8.1.1 Object Localization Visualizations . . . 176
8.1.2 Design Probe and Evaluation . . . 178
8.2 Condition Analysis . . . 181
8.2.1 Signal Selection . . . 182
8.2.2 Tracking Device . . . 184
8.2.3 Data Collection . . . 185
8.2.4 Data Analysis . . . 188
8.2.5 Model Training . . . 190
8.3 System Overview . . . 196
8.3.1 Visual Search Detection . . . 199
8.3.2 Visual Search Support . . . 200
8.4 Evaluation . . . 202
8.4.1 Procedure . . . 202
8.4.2 Results . . . 203
8.5 Discussion . . . 204
8.6 Summary . . . 206

9 Assisting Emotional Disorder 209
9.1 Augmentation Design . . . 211
9.2 Condition Analysis . . . 214
9.3 System Overview . . . 217
9.4 Evaluation . . . 222
9.4.1 Procedure . . . 223
9.4.2 Results . . . 223
9.5 Discussion . . . 224
9.6 Summary . . . 225

IV Conclusion

10 Contributions 229
10.1 Conceptual Analysis . . . 229
10.2 Technical Implementation . . . 230
10.3 Empirical Validation . . . 232

11 Future Work **235**
11.1 Long-Term Studies 235
11.2 Multi-User Applications 236
11.3 Additional Scenarios 236
11.4 Ethical Considerations 237

Bibliography **239**

Appendix **289**
A Component Options 289
B Smartphone Pipeline for Visual Search Support 291
C Smartphone Pipeline for Video and Audio Analysis 294

List of Figures

1.1 Common challenges of augmentation approaches for cognitive processes. . 3
1.2 Structure and contents of this thesis. 6

2.1 General stages of cognitive processing. 12
2.2 Top-down processing (P B D vs. 12 13 14). 12
2.3 Basic components of perception. Adapted from Groome [2014, p. 37]. . . . 13
2.4 Comparison of selective attention models. 16
2.5 Memory model by Atkinson and Shiffrin [1968]. 20
2.6 Elements of Sperling's [1960] partial reporting experiment. 21
2.7 Revised version of the working memory model by Baddeley [2000]. 23
2.8 Components of long-term memory according to Squire [1992]. 25
2.9 Speech production model by Garrett [1975]. 32
2.10 Two-string problem used by Maier [1931]. 33
2.11 Circumplex model of affect by Russell [1980]. 36

3.1 Social signal composed of multiple behavioral cues. 42
3.2 Examples of activated facial action units. 45
3.3 ECG signal components and example RR-intervals. 52
3.4 EDA signal with tonic and phasic components. 53

4.1 Assistive augmentation continuum. Adapted from Huber et al. [2018, p. 2]. 59
4.2 Relation of Human augmentation and its associated terms. Based on de Boeck and Vaes [2021]. 62
4.3 Assistive augmentation classification. 65
4.4 Flow diagram of the literature analysis based on the PRISMA 2020 stages. . 68
4.5 Accumulative timeline of publications identified in our literature analysis across augmentation areas. 70
4.6 Common system structure of analyzed augmentation approaches. 71

4.7 Overview of different sensory augmentation strategies within the context of assistive augmentation. . . . 73
4.8 Generic template to indicate selected design dimensions. . . . 94
4.9 Spectrum of the *target* dimension. . . . 96
4.10 Spectrum of the *initiative* dimension. . . . 98
4.11 Spectrum of the *presence* dimension. . . . 99
4.12 Spectrum of the *direction* dimension. . . . 101
4.13 Spectrum of the *adaptation* dimension. . . . 103

5.1 Example of obtrusive monitoring (based on cartoon by Wim Boost). . . . 114
5.2 Examples showing trade-off between bulkiness and processing capabilities. 115

6.1 General components of the SSJ framework. . . . 128
6.2 Sampling process of continuous to discrete signals. . . . 131
6.3 Mapping of signals to stream packages. . . . 132
6.4 Architecture of the SSJ framework. . . . 138
6.5 Example of Java reflection usage in the SSJ Creator application. . . . 148
6.6 Primary functions of the SSJ Creator application. . . . 149
6.7 Additional views within the SSJ Creator application. . . . 150
6.8 Building an audio recording pipeline with the SSJ Creator. . . . 151

7.1 Visually impaired users participating in our user study. . . . 157
7.2 Selected design dimensions for augmenting visual impairment. . . . 160
7.3 Architecture of the sensory augmentation system. . . . 161
7.4 Assignment of colors to MIDI instruments. . . . 164
7.5 Conversion of text position to sound. . . . 166
7.6 Training examples (left) and experiment set-up (right). . . . 169
7.7 Questionnaire results. . . . 170

8.1 Typical examples of memory lapses: visual search for misplaced objects. . . 173
8.2 Conceptual pipeline to support visual search for objects. . . . 175
8.3 Selected design dimensions for augmenting memory decline. . . . 176
8.4 Overview of visualization candidates. . . . 177
8.5 Participant (left) and experimenter (right) during the design probe. . . . 179
8.6 Object selection screen displayed on Google Glass. . . . 179
8.7 Comparison of sensor data during different activities. . . . 183
8.8 Google Glass-based eye- and head-tracking device. . . . 184
8.9 Overview of the study location. . . . 186

8.10 Participant wearing the sensor setup. 188
8.11 Visual search task durations. 189
8.12 Relation between window length and accuracy (50% window overlap). . . . 193
8.13 Feature composition for early fusion. 194
8.14 Additional results for different search scenarios. 195
8.15 Architecture of the memory augmentation system. 196
8.16 Impressions from the user study. 202
8.17 Results of the SUS questionnaire. 204
8.18 Results of the NASA-TLX questionnaire. 205

9.1 Individual using the proposed cognitive augmentation system. 210
9.2 Selected design dimensions for augmenting cognitive disorder. 212
9.3 Structure of the ABCZ-model. 213
9.4 Average facial expressions per emotion class. 215
9.5 Multi-task MobileNetV2 architecture. 216
9.6 Architecture of the cognitive augmentation system. 218

List of Tables

2.1 Attributes associated with different types of higher cognitive processes according to the dual-process theory. Adapted from Evans and Stanovich [2013]. 28
2.2 Components of language. Based on Ling et al. [2011] and Groome [2014]. . 30

4.1 Mapping of terms between different research fields. 64
4.2 Sensory amplification approaches. 74
4.3 Sensory hearing substitution approaches. 76
4.4 Sensory vision substitution approaches. 77
4.5 Sensory vision substitution approaches for navigation. 80
4.6 Sensory extension approaches. 82
4.7 Sensory enhancement approaches. 84
4.8 Lifelogging memory augmentation approaches. 86
4.9 Real-time memory augmentation approaches. 88
4.10 Cognitive augmentation approaches. 91

5.1 Examples of existing datasets recorded with wearable sensors. 120
5.2 Example CNN architectures for mobile applications. 123

6.1 Existing frameworks for mobile signal processing. 135

7.1 List of participants. 167

8.1 Design probe questionnaire results. 181
8.2 Classification results after feature selection for baseline vs. search. 194

9.1 Performance comparison between AffectNet baseline and our model. 217
9.2 Results of the BDI-II, PSS, and OLBI-S questionnaires. 223

Chapter 1

Introduction

Perceiving, storing, and processing information are some of the core functions of the human mind. They enable us to experience the world, let us expand our knowledge, and form the basis for solving problems through thinking and reasoning. However, impairments and disorders of these processes can severely impact the daily functioning and quality of life of affected individuals. Since the conditions can be caused by various factors such as accidents, brain injuries, illnesses, or even natural aging, a significant proportion (approximately 25%) of the population in Europe and the United States find themselves confronted with such a situation at least once in their lives [Gravenhorst et al., 2015]. Depending on the severity of the condition, consequences can range from being unable to perform specific actions, such as exercising a profession or participating in social activities, to the complete loss of personal independence [Scherer et al., 2005]. While some effects are only temporary and have relatively mild symptoms, others can lead to long-term deficits in cognitive areas that affect people for the rest of their lives. In addition to the severe impact on individuals, impairments and disorders of cognitive processes can also become a heavy burden for relatives due to the associated caregiving and treatment requirements. In some cases, family members have to take on additional responsibilities and might feel the need to reduce their work hours or give up work entirely to care for their loved ones, which can result in severe financial consequences. These not only affect the families but also society as a whole due to the potential loss of productivity and the high costs of therapy and rehabilitation within the healthcare system [Pérez Fornos et al., 2019]. For these reasons, providing assistance to affected people through automated tools and specialized systems has a high potential to ease their burden and reduce the impact of their conditions. However, supporting impairments and disorders of cognitive processes also comes with its own set of challenges.

Since the conditions are often different and specific to individuals, potential solutions must be tailored to their specific circumstances. This process can be expensive and time-consuming and might be the reason why there are no widely available tools and systems to build such specialized solutions.

One technological advancement that could counteract these problems are wearables and mobile devices. In recent years, they have become available to a wide range of people and have taken over the role as the primary device for everyday computational tasks. According to the latest Ericcson and GSM Association (GSMA) reports, the number of smartphone subscriptions reached 6.93 billion in 2023 [Ericcson, 2024], and more than 4.6 billion people (57% of the global population) used their mobile devices to access the internet [GSMA, 2023]. Due to the relatively low cost and significant improvement of their processing capabilities, they have become a ubiquitous companion that can serve as the foundation for potential applications aiming to assist cognitive processes at any place and time. In addition to the raw computational power, integrated sensors such as accelerometers, cameras, and microphones enable signal-based approaches for recognizing people's current cognitive states. The results of this analysis include essential information to determine when assistance is needed and in which form it should be provided. Furthermore, with physiological sensors becoming small enough to fit into wearables like wristbands and smartwatches, the sensing capabilities can even be extended [Schneegass, 2016]. Beyond analyzing users and their environments, wearable and integrated sensors also enable comprehensive interaction techniques such as gesture, touch, voice, and gaze-based inputs. Additionally, various output modalities, including visual, acoustic, and tactile feedback, are supported through built-in displays, speakers, and vibration motors. This wide variety of input and output possibilities facilitates interaction experiences that can be tailored to the specific needs and circumstances of individuals.

However, despite all the features of mobile devices, more than hardware and potential capabilities are needed to solve the lack of tools and systems that enable the rapid prototyping and development of personalized applications to support impairments and disorders of cognitive processes. While there are some approaches that focus on very specific conditions, they usually can not be adapted or repurposed for different circumstances. Although these solutions might benefit a small number of targeted individuals, developing specialized applications from scratch without the ability to adapt and reuse implemented sensing or processing components quickly becomes expensive and is not sustainable in the long run. Moreover, existing libraries and toolkits often lack particular properties and functionalities required to build comprehensive approaches for assisting cognitive processes. These shortcomings include missing input and output capabili-

ties, insufficient implementations of mobile processing techniques, limited support for sensors and modalities, long iteration times, and overly complex requirements for extensions. Additionally, potential solutions need to overcome various challenges related to hardware constraints and ubiquitous application scenarios. Since current mobile devices are still limited in terms of computational power, memory, storage, transmission bandwidth, and battery life, efficient algorithms are required to achieve suitable augmentation approaches. In this regard, most solutions rely on machine learning techniques, which necessitate the acquisition of training data. Unfortunately, only very few publicly available corpora match the specific circumstances of assistive augmentation scenarios and contain appropriate sensor data from mobile and wearable devices. Consequently, potential approaches must support recording relevant signals in addition to the processing capabilities mentioned above.

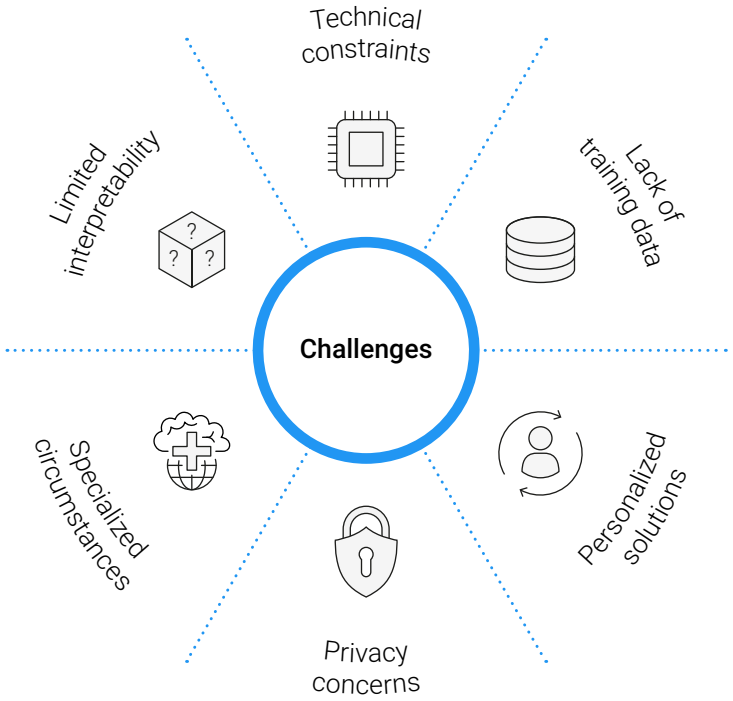


Figure 1.1: Common challenges of augmentation approaches for cognitive processes.

Besides these technical restrictions, various privacy and ethics-related aspects also need to be considered. Since the collected corpora can contain highly sensitive and personal information about users and their daily lives, adequate security mechanisms must be implemented to protect the privacy of individuals and keep them in control of their data. Furthermore, potential reservations about the constant surveillance with invasive sensors, as well as uncertainties regarding how signals are stored, shared, processed, and interpreted, must be addressed to improve people's trust and acceptance of augmenta-

tion solutions. For these reasons, the present thesis focuses on developing a flexible and easily usable framework that utilizes the capabilities of mobile devices and enables the rapid prototyping and implementation of assistive augmentation approaches to support individuals affected by impairments and disorders of cognitive processes. Since another possible cause for the lack of similar tools and systems might be the limited understanding of the specific needs and requirements in targeted situations, which makes it challenging to design effective solutions, this thesis also investigates theoretical concepts and guidelines that can be applied to streamline this procedure. Finally, we illustrate how the proposed conceptual and technical solutions can be employed to support different cognitive processes through multiple research probes of prototypical applications and demonstrate their feasibility in empirical evaluations with affected users.

1.1 Research Objectives

Based on the overarching goal of assisting people with impairments and disorders of cognitive processes, we derived the following conceptual, technical, and empirical research objectives:

- The initial objective of this thesis is to investigate common conceptual properties, patterns, and strategies from previous approaches targeting the assistance of cognitive processes. Based on these findings, a set of design dimensions should be identified that capture the full spectrum of characteristics found in the analyzed systems. Apart from classifying existing approaches, the resulting conceptual framework could also be used to support the design and development process of future systems by illustrating the different variations and their implications within each dimension. This would facilitate the selection of appropriate design choices and serve as a common foundation for comparing different assistive augmentation approaches. In addition to identifying guidelines for supporting cognitive processes, another part of this objective is to provide an overview of the underlying theories from the field of cognitive psychology to better understand their inner workings and consequently create more effective solutions.

- The second research objective is to introduce a technical solution that can be utilized to counteract the absence of tools and systems facilitating the rapid prototyping and development of assistive augmentation approaches for cognitive processes. One core requirement concerns supporting real-time signal processing on mobile devices to detect people's current mental states and provide appropriate

feedback based on analyzed sensor data. In this regard, it should be possible to create, adapt, reuse, replace, repurpose, and rearrange all processing components during the prototyping and development phase with minimal effort to enable a high degree of flexibility and promote experimenting with different approaches. Moreover, the framework should support various sensing devices, data types, and comprehensive mechanisms to synchronize the resulting signals. Based on these capabilities, it should also accommodate a wide variety of input and output modalities, allowing developers to implement personalized interactions and provide tailored assistance. Additionally, the proposed technical solution should address common challenges of assistive augmentation approaches to ensure its feasibility, effectiveness, and acceptance among users. As shown in Figure 1.1, examples include the limited availability of suitable training data, which can restrict the development of robust recognition models, as well as privacy concerns related to sensitive and personal information captured by ubiquitous sensors. In this context, transparency and comprehensibility are equally important aspects that should be considered to increase people's trust towards mobile and assistive technologies. Potential advancements in these areas could reduce the time and resources needed by researchers and developers to implement personalized augmentation systems for cognitive processes, which would lead to a greater diversity of approaches, accelerated research progress, and ultimately benefit affected individuals.

- Finally, the last objective is to evaluate the effectiveness of the proposed conceptual and technical solutions through empirical user studies. To this end, it is essential that the evaluations are conducted under realistic circumstances and with people affected by the targeted conditions to ensure the validity of the results. More precisely, following this procedure facilitates a proper performance and robustness analysis of the proposed systems in their intended environments, which would not be possible in controlled laboratory studies or with unaffected individuals. Additionally, it enables the collection of feedback, insights, and experiences directly from end-users, which can contain valuable information for improving the preliminary solutions. Besides clarifying whether individuals actually benefit from the proposed techniques, such empirical findings would also provide a verified foundation for future system iterations and innovative new approaches.

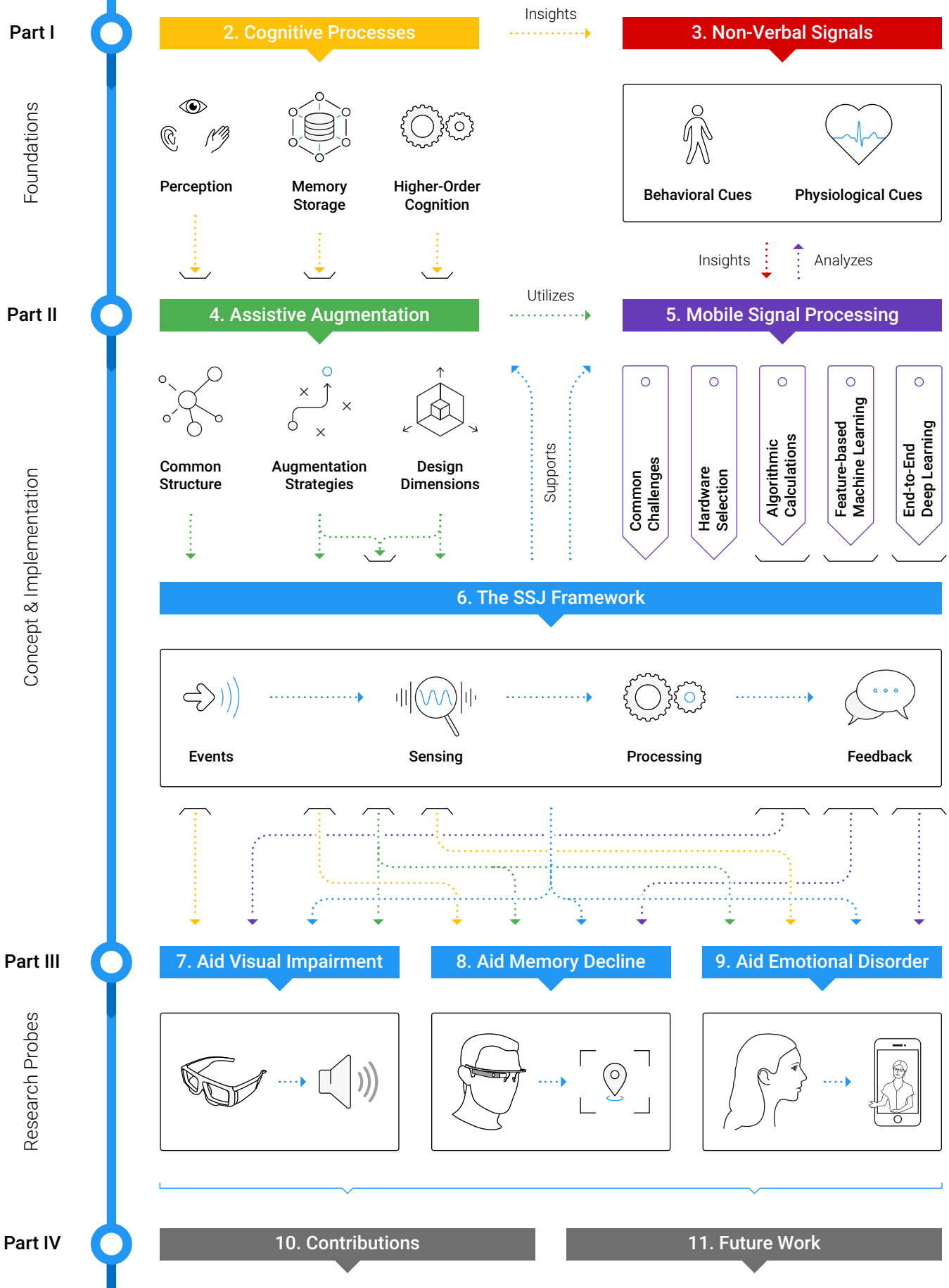


Figure 1.2: Structure and contents of this thesis.

1.2 Thesis Outline

As illustrated in Figure 1.2, this thesis consists of four distinct parts and eleven chapters. The first part is concerned with the theoretical background. It introduces the fundamental concepts and theories regarding mental processes from the field of cognitive psychology to improve the understanding of how they work and which aspects of them can be augmented (Chapter 2). Chapter 3 then provides an overview of available non-verbal signals that can be captured and analyzed to gain insights about the respective processes. To this end, behavioral cues like facial expressions, gaze, gestures, posture, and paralinguistic properties, as well as physiological cues, including brain, heart, electrodermal, and muscle activity are described in more detail.

Following that, Part II focuses on the conceptual aspects of assistive augmentation and the technical implementation of a universal solution that can be utilized to support impairments and disorders of cognitive processes. More precisely, Chapter 4 establishes the theoretical context of assistive augmentation and includes a literature analysis of related works to derive shared properties among them and identify suitable augmentation strategies for each group of cognitive processes. Based on the resulting findings, common design dimensions are proposed to classify existing approaches and guide the development of future applications. Chapter 5 then provides an overview of the typical methods, challenges, and solutions involved in processing signals on mobile devices. While it focuses on conceptual details and available procedures, the practical application of each step is demonstrated through multiple research probes in Part III. Informed by the identified challenges, requirements, and general structure of previous works, Chapter 6 introduces a novel open-source software framework for building and prototyping assistive augmentation systems using mobile signal processing techniques. In addition to addressing the limitations of existing toolkits and highlighting the benefits of the implemented approach, its fundamental design principles, overall architecture, core components, and graphical user interface are also explained.

Part III demonstrates the capabilities and evaluates the effectiveness of the proposed conceptual and technical solutions with research probes for each primary group of cognitive processes. In this regard, the complete workflow of designing, implementing, prototyping, and testing the respective augmentation systems is described. The first research probe in Chapter 7 focuses on augmenting the perception of visual information for blind and visually impaired people. To this end, an application is proposed that enables affected users to explore the environment with their remaining senses by converting the inaccessible signals into acoustic representations. In Chapter 8, a mem-

ory augmentation system is presented that helps older adults remember the location of misplaced objects. To reduce their frustration and mental demand, the application automatically recognizes critical situations and proactively offers appropriate assistance. The third research probe (Chapter 9) demonstrates how the framework can be used to build a cognitive augmentation system that supports the outpatient treatment of individuals with depression and related cognitive disorders. It provides a ubiquitous companion that adapts its behavior to people's current condition and bridges the gap in treatment options between hospitalization and therapy sessions.

Finally, Part IV summarizes the conceptual, technical, and empirical contributions of this thesis in Chapter 10 and outlines potential opportunities for future work to build upon the current findings in Chapter 11. Additionally, the bibliography with all references and the appendices containing supplementary material are also included.

I Foundations

2 Cognitive Processes 11

2.1 Perception
2.2 Memory Storage
2.3 Higher-Order Cognition
2.4 Summary

3 Non-Verbal Signals 41

3.1 Behavioral Cues
3.2 Physiological Cues
3.3 Summary

Chapter 2

Cognitive Processes

One of the main objectives of this thesis is to explore conceptual and technological approaches to aid the development of mobile assistive systems for the augmentation of the human intellect. Consequently, this chapter introduces fundamental concepts and theories from the field of cognitive psychology regarding involved mental processes to provide a better understanding of how they work and which aspects of them can be augmented. Additionally, potential disorders of each process are discussed to serve as starting points for future assistive augmentation approaches.

In general, cognitive psychology can be defined as the scientific study of all mental abilities and processes that are the basis for human behavior. It concerns how people perceive, learn, remember and think about information [Neisser, 1967; Sternberg and Sternberg, 2012]. While the field has evolved through several phases and different theories since the end of the 19th century, the current consensus is based on the *information processing* model, which uses the processing workflow of digital computers as an analogy for human cognition [Eysenck and Keane, 2020]. This paradigm was initially brought to prominence by Broadbent [1958], who argued that the majority of mental processes consist of a sequential series of processing stages, as shown in Figure 2.1.

Typically, each process begins with a stimulus, which is acquired through sensing organs and analyzed in the initial *perception* stage. At this point, the brain already tries to interpret the input in an effort to make sense of its contents [Groome, 2014]. Parts of the resulting information are then transferred to the *memory* storage, where a record of it is retained for later use. Combined with existing knowledge, this provides the foundation for higher-level mental activities in the *cognition* stage, such as language, thinking, or problem-solving. While this basic concept serves well to illustrate the general stages

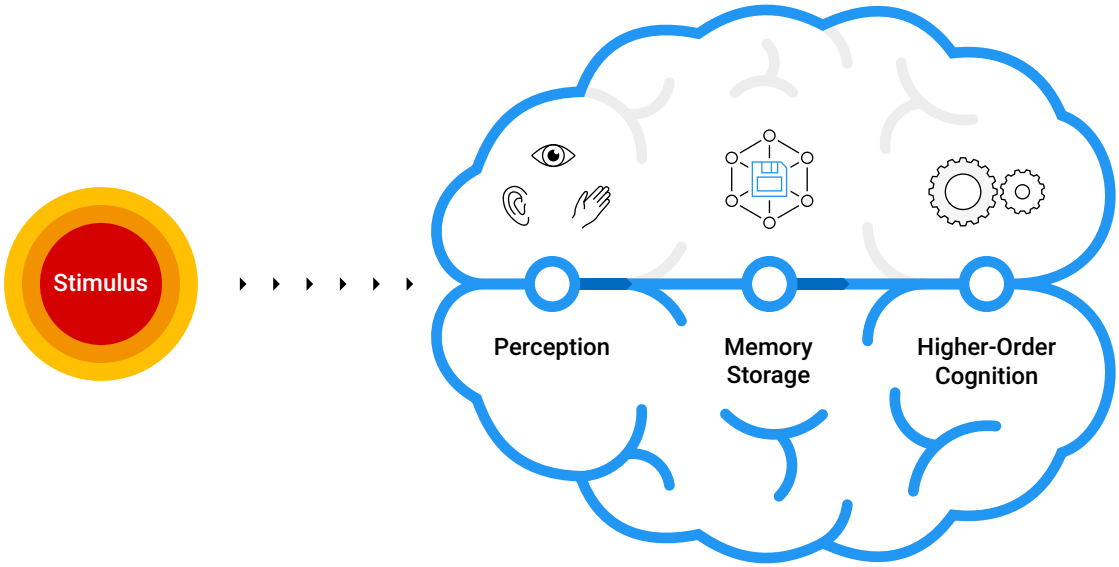


Figure 2.1: General stages of cognitive processing.

of cognitive processing, reality is much more complex than the model implies. For instance, Figure 2.1 suggests that the processing stages are completely distinct but in reality they overlap and interact with each other. This especially becomes apparent when considering different types of input perception such as *bottom-up* and *top-down* processing [Neisser, 1967]. Bottom-up processing refers to the perception of sensory information, which progresses up towards higher cognitive stages based on the nature of the stimulus. In contrast, top-down processing is driven by higher levels of cognition where an individual's existing knowledge, experiences, and expectations are sent down to the lower stages and affect incoming sensory information [Groome, 2014]. An example of this is shown in Figure 2.2, where the same sensory information is perceived differently depending on the expectations set by the surrounding context.

Figure 2.2: Top-down processing (P B D vs. 12 13 14).

As outlined, the actual cognitive processes can be much more complex than illustrated in the fundamental processing model above, which should only be regarded as a simplified representation of the general processing stages. Since the inclusion of all interactions

and relations of every process in a single unified model is still an ongoing challenge that has not been accomplished, the concepts and theories of each individual cognitive process are examined in the following sections of this chapter.

2.1 Perception

Experiencing the world through our senses is one of the primary cognitive processes. It allows us to interact with our surroundings and gain essential information required for our survival. Within the field of cognitive psychology, perception is defined as the acquisition, organization, and interpretation of incoming sensory information about the environment and the internal state of our bodies. While this process subjectively seems effortless, it involves several complex functions of the nervous system which occur outside our conscious awareness and do not require any active thought [Goldstein, 2010]. The complexity of these functions becomes apparent when trying to build systems and applications which attempt to artificially replicate human perception (e.g., for self-driving cars, autonomous robots, or augmentation purposes). Although the process in its entirety is still unmatched by any computer system, specific parts have been successfully imitated due to recent technological advancements especially in the field of computer vision [Eysenck and Keane, 2020]. An overview of all components involved in the perceptual process is illustrated in Figure 2.3.

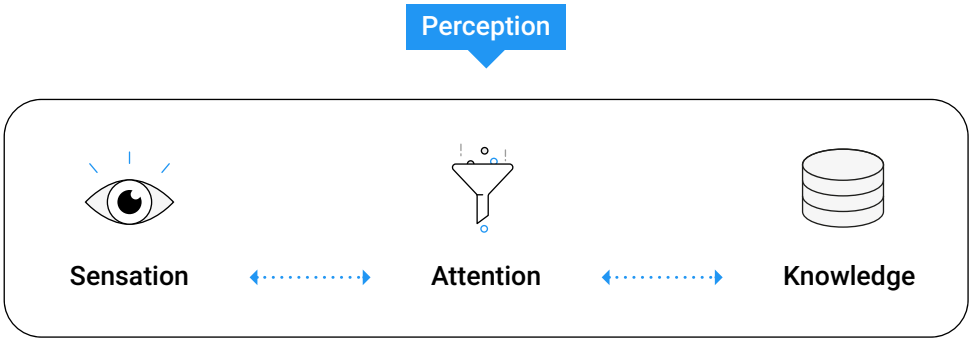


Figure 2.3: Basic components of perception. Adapted from Groome [2014, p. 37].

Initially, perception starts with a stimulus which is captured by our sensory organs. This process is called sensation and refers to the raw, unaltered sensory input. It mainly involves the transformation of incoming stimuli into signals that can be processed by our nervous system. After the conversion, the resulting sensory information gets filtered through attentional processes. This step is necessary since the amount of continuously incoming stimuli would otherwise be too overwhelming. Once the relevant informa-

tion has been selected, it gets combined with existing knowledge and previous experiences to form the basis for interpretation. This includes recognizing patterns, identifying matching mental models, and even altering the perceived information to fit present expectations (top-down processing). At this point, the resulting output of the perceptual process might be a highly modified version of what was initially captured by our senses [Groome, 2014]. Consequently, the perception of the same stimulus might vary from one person to another because it can be interpreted differently based on the individual attitude, knowledge, goals, needs, values, experiences, expectations, and physical condition [Kenyon and Sen, 2015]. It is also the reason why some information does not reach the perception stage at all and gets discarded after sensation.

2.1.1 Sensation

The process of sensation involves the acquisition of sensory information through our organs and is the first step in the perceptual pipeline. It begins with an environmental stimulus that transmits energy such as light and sound waves, mechanical pressure, or chemical reactions. These forms of stimuli are captured by our receptor cells which transform the environmental energy into electrical signals for further processing. The transformation from one form of energy into another one is also called *transduction* and occurs in every sensory organ [Goldstein, 2010]. For example, when light hits the retina of our eyes, it gets converted into electrical signals that represent the sensed visual information. After conversion, the signals are transmitted through the central nervous system to the corresponding areas of the brain, where further steps, such as filtering and interpretation, are performed [McBride and Cutting, 2019].

In order to perceive the environment, humans generally possess five senses: sight, hearing, touch, taste, and smell. Our eyes detect light reflected from our surroundings, our ears pick up sounds in the vicinity, our nose recognizes scents in the air, our tongue reacts to different flavors, and our skin perceives pressure and temperature. In addition to these external senses, humans are also able to perceive information about the internal state of our bodies [Macpherson, 2011]. For instance, we can feel hunger, thirst, suffocation, and tiredness, which are essential properties for our survival. We can sense our posture and the position of our limbs relative to our torso. We can even become aware of physiological phenomena occurring inside our bodies, such as the beating of our heart, the inflation of our lungs, and the stretch of our bladder.

While each of these sensory systems is specialized in perceiving a different type and range of phenomena, all of them have in common that in order to detect a specific

stimulus, a minimum amount of stimulation is required. This perception barrier is also called *absolute threshold* and refers to the smallest quantity of energy necessary for a stimulus to be sensed [Goldstein, 2010]. An example of this is the brightness threshold at which the lowest amount of light energy can still be seen as a flash of light. Similarly, the *difference threshold* refers to the smallest difference between two stimuli of the same type that can still be perceived. For instance, when comparing the heaviness of multiple objects, it is the smallest noticeable weight difference between them. Since both thresholds can vary based on a person's current environment and internal state, it is essential to consider the possible value ranges when building sensory augmentation systems so that users can perceive the targeted stimuli and distinguish their different intensities under any circumstances.

2.1.2 Attention

Within the context of cognitive psychology, attention refers to selecting and prioritizing information for conscious processing. It acts as a filter that blocks irrelevant stimuli and allows individuals to focus on specific details. Due to its close relation to other cognitive processes, such as sensation, memory, language, and problem-solving, it plays a central role in our daily lives [McBride and Cutting, 2019]. While attention can be directed intentionally (e.g., by targeting an object with our eyes), it can also be captured unintentionally by our surroundings due to sudden sounds or movements [Groome, 2014]. However, since attention is a limited cognitive resource, such unwanted focus shifts can negatively impact performance in previously attended tasks. One of the reasons is that attended information is stored in short-term memory for further processing, which only has a limited capacity. As described by William James in one of the first definitions of attention, this constraint necessitates *"[...] the withdrawal from some things in order to deal effectively with others"* [James, 1890]. Without such filtering mechanisms, the amount of incoming stimuli would otherwise overwhelm our cognitive capabilities. Similar to spotlights that highlight certain regions of the environment while hiding others in darkness, this selective process enables individuals to effectively perceive relevant information by allocating cognitive resources to the attended stimuli.

Overall, there are two distinct types of attention: (1) *selective* (or *focused*) *attention*, which involves the ability to concentrate on one specific stimulus while ignoring others; and (2) *divided attention*, which refers to the process of allocating cognitive resources to multiple stimuli simultaneously [Brown, 2006]. One of the first models that provided a complete concept for the selective attention process was the *filter theory* by

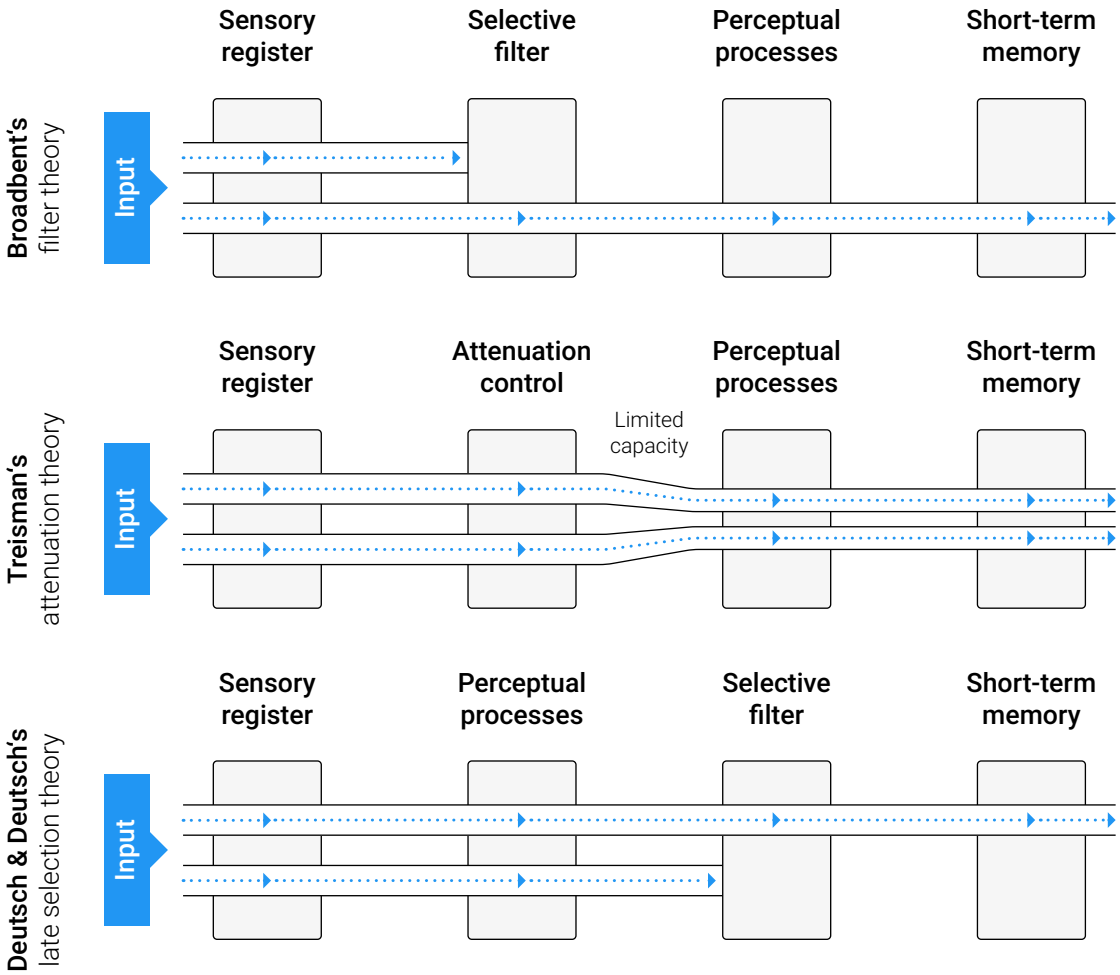


Figure 2.4: Comparison of selective attention models.

Broadbent [1958]. According to his model, incoming stimuli arrive in parallel at a fast-decaying sensory register [Groome, 2014]. From there, only one input at a time can pass through the selective filter into the short-term memory, which can cause processing bottlenecks but prevents information overload. While the model initially matched available research results, later studies by Moray [1959] and Treisman [1960] revealed that unattended messages can break through the selective filter and still be consciously perceived. Deutsch and Deutsch [1963] accounted for these findings in their *late selection theory* and argued that the semantic properties of all inputs are analyzed in parallel before the most relevant stimulus get selected. As shown in Figure 2.4, this concept places the bottleneck much closer to the end of the attentional process than Broadbent's theory. In an attempt to provide a compromise between both models, Treisman [1964] proposed the *attenuation theory*, which suggests a more flexible position for the processing bottleneck. According to her model, the selective filter reduces the strength of

unattended stimuli based on their relevance rather than completely blocking them from perception. Depending on the activation level of the input event and the availability of processing capacity, this enables the conscious awareness of information outside the current focus [Eysenck and Keane, 2020].

While several revisions and alternatives to these models have been proposed throughout the past, they significantly shaped our understanding of the selective attention process and still serve as a foundation for current research. They even inspired models for divided attention, such as Kahneman's [1973] *central capacity theory*, which expands upon Broadbent's initial concept. The model suggests we only have a limited pool of cognitive resources that can be divided between multiple processing tasks. As soon as the combined demands of all activities exceed the available capacity, interference between tasks and performance degradation might occur [Ling et al., 2011]. In this regard, the required amount of resources to process a certain activity primarily depends on its difficulty level. However, the total available capacity can also vary based on a person's current arousal, which in turn is influenced by their goals, effort, and motivation.

One aspect that this model does not fully account for is *structural interference*, which occurs when multiple activities compete for the same perceptual mechanisms. As studies have shown, it is generally more difficult to monitor concurrent tasks within the same modality than within different ones [Treisman and Davies, 1973]. For instance, listening to an audiobook while cooking requires little to no effort, but doing so during a conversation is rather challenging. To address this phenomenon, Navon and Gopher [1979] expanded the central capacity theory and proposed the existence of *multiple resource pools*. According to them, each task requires a combination of resources from different pools and as long as they are available, multiple activities can be performed simultaneously without interference. This distinction also explains the degraded performance of parallel tasks that compete for the same resources even though the total processing capacity is not exhausted. Inspired by these findings, Wickens [1980, 1984, 2002] proposed a refined *multiple resource model* that uses different sides of a cube to represent each of the following dimensions: (1) stages of processing (perception, cognition, and responding), (2) codes of processing (spatial and verbal), (3) modalities (visual and auditory), and (4) visual channels (focal and ambient vision). The general idea is that as long as two tasks require different characteristics along each dimension, their concurrent execution should not lead to interference [Wickens, 2008].

2.1.3 Disorders

Impairments and disorders related to sensing, perception, and attention are conditions that negatively impact a person's ability to perceive, process, and interpret sensory information. They can directly affect any of our senses, including sight, hearing, touch, taste, and smell. While covering all existing conditions exceeds the scope of the present thesis, we still provide some examples to illustrate the range of potential symptoms. For instance, a *cataract* is a cloudy area forming within the eye's naturally clear lens that leads to blurred vision, faded colors, bright light sensitivity, and decreased ability to see at night [Lam et al., 2015]. It typically develops at an older age and can cause vision loss over time. Similarly, *macular degeneration* is an age-related condition that affects an area of the retina (the macula) responsible for sharp and focused vision. Although peripheral perception remains intact, the loss of central vision makes it difficult to read, drive, or recognize faces. Another group of visual impairments are *glaucoma*. They gradually damage the optic nerve, often due to abnormally high pressure within the eye, and can lead to blindness if left untreated.

In contrast, one of the most common hearing disorders is *presbycusis*, which refers to the irreversible hearing loss that progressively occurs in most older adults as part of the natural aging process. It usually affects both ears equally and is characterized by difficulties hearing high-pitched sounds or understanding speech in noisy environments. Similar symptoms also apply to *auditory neuropathy*, which is a condition where the hearing organ successfully perceives sounds but can not properly transmit the captured signals to the brain [de Siati et al., 2020]. The dysfunction might be caused by damage to the inner hair cells or the auditory neurons, leading to impaired recognition of spoken language. *Otosclerosis* on the other hand, is an abnormal deformation of the bone structure surrounding the inner ear that also disrupts the transmission of sounds and can result in tinnitus, vertigo, and hearing loss [Uppal et al., 2009]. While it typically occurs bilaterally, the severity can vary between both ears.

Apart from sensing disorders, there are various conditions that affect the processing of acquired signals. For example, *agnosia* is the inability to recognize or interpret specific stimuli despite intact sensory functioning [Groome, 2014, p. 119]. Since its symptoms usually only impact a single modality, several distinct variations exist [Burns, 2004]. More precisely, visual agnosia includes the failure to recognize familiar faces (*prosopagnosia*), shapes (*apperceptive agnosia*), and object functions (*associative agnosia*), while auditory agnosia encompasses the inability to differentiate speech from other sounds (*verbal agnosia*) and between familiar voices (*phonagnosia*). Another

condition with similar symptoms is *spatial neglect*. It typically occurs after brain damage to one of the hemispheres and results in an attention and awareness deficit on the opposing spatial side (e.g., people with damage to the right hemisphere fail to respond to stimuli on the left side of their field of view) [Groome, 2014, p. 113].

2.2 Memory Storage

The ability to remember and recall information is essential to our daily lives. It enables us to learn from experiences, navigate complex situations, and establish our own personality. Without it, we would be unable to form meaningful relationships, make sophisticated decisions, or perform more than basic actions. Fundamentally, our knowledge and abilities would remain at the same level as that of newborn children [Eysenck and Keane, 2020]. Within the context of cognitive psychology, memory is considered a complex and multifaceted process that involves the following stages: *encoding*, *storage*, and *retrieval* [Baddeley et al., 2015]. During the initial stage, external information from the environment gets transformed into a suitable representation that can be stored and further processed. For that, the perceptual processes described in Section 2.1 are used to capture and provide the sensory stimuli as input for conversion. This step involves interpreting the acquired signals, extracting relevant features, and associating them with existing knowledge. In this regard, the quality and persistence of the encoded information depends on various factors, such as the current emotional state, task complexity, and number of repetitions [Sridhar et al., 2023].

Following that, the converted information enters the storage phase, where it is structured, reorganized, and maintained over time. Based on the quality of the encoding process, the resulting representations are stored in different types of memory structures. These range from transient sensory registers with limited capacity to large persistent data repositories. Although these stores do not correspond to distinct physiological structures in our brains, they serve as conceptual constructs that help us understand the mechanisms and phenomena behind the cognitive process [Sternberg and Sternberg, 2012]. The final stage is concerned with bringing memories back into conscious awareness. This involves accessing, decoding, and reconstructing stored information for further use. Similar to the encoding process, retrieval also requires mental effort and attention to effectively restore the original stimuli. Apart from those, another essential factor that can influence the outcome of this stage are memory cues. Depending on their distinctiveness and association with the desired memories, they can facilitate the process of identifying and recalling the correct information. However, despite all of

these properties and mechanisms, retrieval does not always produce a perfect copy of the original events. Instead, it can result in distorted or inaccurate reconstructions due to the influence of various factors such as prior knowledge, beliefs, and expectations. For these reasons, it is essential to understand the details of the memory process in order to support it effectively with mobile and wearable technologies.

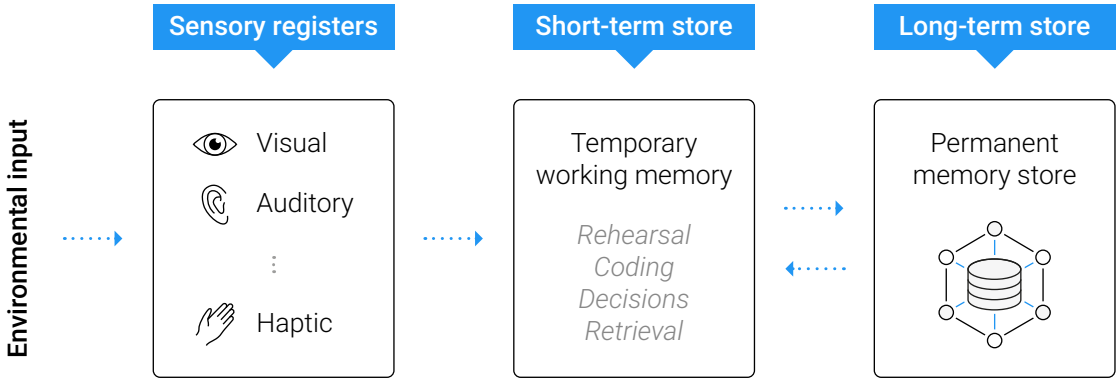
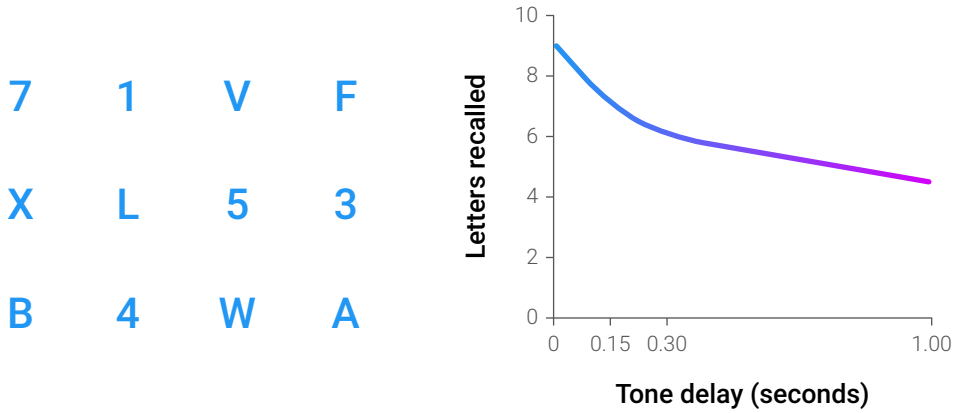


Figure 2.5: Memory model by Atkinson and Shiffrin [1968].

One of the first models that provided a theory for the basic structure and operation of memory was proposed by Atkinson and Shiffrin [1968]. According to their concept, information flows through and is held in one of three separate stores: the *sensory register*, *short-term memory* (STM), and *long-term memory* (LTM). An overview of the processes and relations between these structures is illustrated in Figure 2.5. While the sensory register only retains unprocessed stimuli for a very brief period, short-term memory temporarily maintains information that currently receives our conscious attention [Sternberg and Sternberg, 2012; McBride and Cutting, 2019]. At this stage, it is possible to prolong the retention of processed information through continuous rehearsal, but without it, the encoded representations decay and are forgotten almost immediately [Groome, 2014]. In contrast, long-term memory permanently stores information outside our conscious awareness for future retrieval. With its almost unlimited capacity, it can preserve a lifetime's worth of memories and keep them available for recall at any point in time [Eysenck and Keane, 2020]. Despite its age, Atkinson and Shiffrin's model still gets used to explain the basic concepts of memory and serves as a foundation for current approaches. However, evidence suggests that certain parts, such as the flow of information, might be more complex than indicated by the model [Baddeley et al., 2015]. For this reason, we outline the details and further concepts regarding different types of memory in the following sections.

2.2.1 Sensory Memory

When observing the movement of a bright object in a dark environment, it appears to leave behind a rapidly fading trail along its path. This phenomenon occurs because the perceptual system briefly retains the raw stimuli for processing even after the physical signals cease to exist [McBride and Cutting, 2019]. Thereby, each modality has its own sensory memory. For instance, the visual register is referred to as *iconic memory* while its auditory counterpart is termed *echoic memory* [Neisser, 1967]. Since these transient stores have a relatively short retention time, measuring their exact duration and capacity is rather difficult. In an attempt to overcome this challenge, Sperling [1960] initially presented an array of twelve letters to participants for 50 milliseconds and asked them to recall the displayed information (see Figure 2.6a). On average, they were able to remember around four items, which is in line with prior findings by Brigden [1933]. However, participants also mentioned that they had seen more items but forgot them while responding. To avoid this problem, Sperling repeated the experiment and reduced the number of items to be reported [Baddeley et al., 2015]. For that, he instructed them to only recall the letters from one of the three lines based on the selection indicated by an acoustic signal (i.e., high pitch = first line, low pitch = last line). Since participants did not know beforehand which line would be requested, their partial recall performance of roughly three out of four items per row only represented one third of the stored information [Sternberg and Sternberg, 2012]. Consequently, Sperling extrapolated the results and estimated the total capacity of the iconic memory at around nine symbols.



(a) Example array of letters **(b)** Study results

Figure 2.6: Elements of Sperling's [1960] partial reporting experiment.

In subsequent experiments, he systematically varied the time between displaying the array and prompting the recall with the acoustic signal to measure the duration of the

iconic store [McBride and Cutting, 2019]. As shown in Figure 2.6b, the average recall of nine items was only achieved when immediately prompted following the visual display and rapidly decreased to four or five symbols after just one second. These results indicate that visual stimuli are kept in iconic memory for around 200 to 400 milliseconds before they disappear [Malim, 1994]. Following Sperling's experiments, other researchers focused on determining the durations of the remaining modalities. For instance, Darwin et al. [1972] applied the partial reporting procedure to analyze echoic memory and found that auditory information can be retained for around two to four seconds [McBride and Cutting, 2019]. While the likelihood of errors substantially increased towards the end of visual sequences, this was not the case for auditory representations. Instead, the last one or two items were more likely to be correct than previous entries in the list [Baddeley et al., 2015]. Regarding further modalities, there is very little information available apart from general properties and concepts since most research has focused on visual and auditory senses in the past [McBride and Cutting, 2019].

2.2.2 Short-term Memory

According to the original model by Atkinson and Shiffrin [1968], short-term memory refers to a temporary storage that can maintain small amounts of material for a limited duration. The primary purpose of this store is to hold information that receives our conscious attention and to control the transition towards long-term memory [Sternberg and Sternberg, 2012]. While its capacity of around seven plus or minus two items is relatively similar to that of sensory memory, it can retain the material for a significantly longer period [Miller, 1956]. Typically, information remains in short-term memory for about 30 seconds, but through repeated rehearsal, this duration can be extended up to several minutes [Sternberg and Sternberg, 2012]. Apart from rehearsal, its capacity can also be increased by grouping related pieces of information into larger chunks [Miller, 1956]. For instance, it is relatively challenging to remember this sequence of 22 digits "1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0", but when grouped into larger units such as "100, 1000, 100, 10, 100, 1000, 100", it becomes much more manageable. In such cases, the typical capacity limit then applies to the number of chunks instead of the individual elements they consist of [Miller, 1956]. However, the capacity and duration of short-term memory can also be negatively influenced by interference. This phenomenon primarily occurs when new information replaces similar existing memories (retroactive interference) or when older material keeps new stimuli from being stored (proactive interference) [McBride and Cutting, 2019].

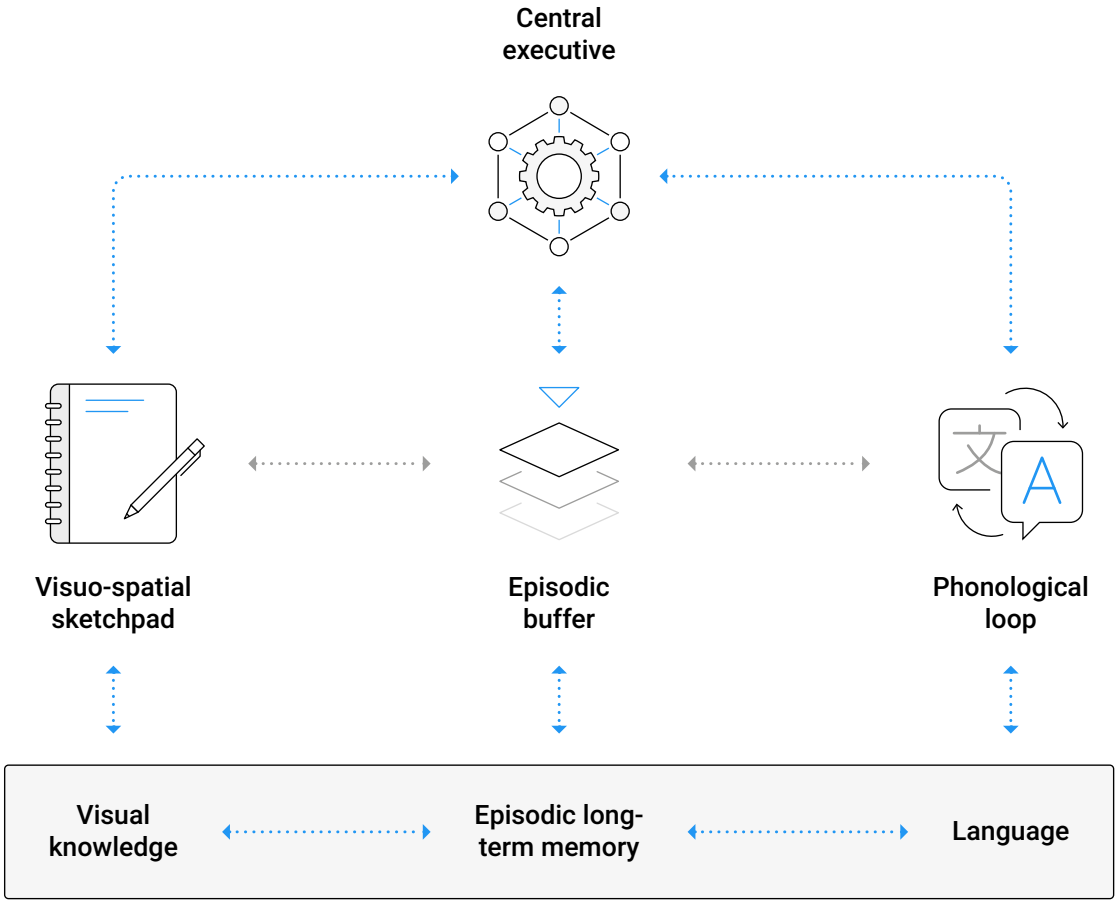


Figure 2.7: Revised version of the working memory model by Baddeley [2000].

Although Atkinson and Shiffrin's concept laid the foundation for understanding short-term memory, most cognitive psychologists today believe it is more than just a passive temporary store with limited capacity and duration [Hills, 2016]. One of the most widely adopted theories that replaces the traditional view with a more active concept is the *working memory model* by Baddeley and Hitch [1974]. In their work, they propose the existence of a memory system that serves as a mental workspace where new and existing information gets manipulated and temporarily maintained [Baddeley et al., 2015]. Similar to the screen of a computer, which is used to perform various tasks on current data, working memory also resembles a space where the analysis and processing of information takes place [Groome, 2014]. As shown in Figure 2.7, the model consists of four major components: *central executive*, *phonological loop*, *visuo-spatial sketchpad*, and *episodic buffer*. The most important subsystem is the central executive, which assigns attentional resources to the other subsystems and controls the flow of information between them [McBride and Cutting, 2019]. While it has no capacity to store material on its own, it decides what information will be processed by which component and

whether the results should ultimately reside in long-term memory [Ling et al., 2011]. However, due to the limited nature of attention (see Section 2.1.2), the central executive can only devote resources to the subsystems in case they are available. Otherwise, processing performance will be degraded, which is also the case when multiple tasks simultaneously require the same component [Eysenck and Keane, 2020].

The first subsystem that receives instructions from the central executive is the phonological loop, which is responsible for processing and storing sequences of verbal information. To achieve that, it uses two subcomponents: the *phonological store*, which holds the sounds for a brief period, and the *articulatory control process*, which enables the silent rehearsal of stored items through verbal repetition with a person's inner voice [Brown, 2006]. The second subsystem is the visuo-spatial sketchpad, which is the visual equivalent of the phonological loop. It is responsible for holding and manipulating visual and spatial information, similar to a whiteboard that can be erased and rewritten. According to Logie [1995], it can be further divided into the *visual cache*, which stores material related to colors and shapes, and the *inner scribe*, which processes orientation, location, as well as movement information and is involved in its rehearsal [Groome, 2014]. The final subsystem is the episodic buffer, which was retroactively introduced by Baddeley [2000] to address certain phenomena that could not be explained with the original model. It serves as an interface between the other subsystems and long-term memory and provides a temporary store where information from these sources can be integrated into an episodic representation. This mechanism enables us to re-evaluate existing knowledge and memories with more recent experiences and allows us to combine material from different modalities [Sternberg and Sternberg, 2012]. Additionally, it can briefly hold information initially intended for other subsystems while they are otherwise engaged [McBride and Cutting, 2019].

2.2.3 Long-term Memory

The final stage of Atkinson and Shiffrin's model is concerned with the long-term preservation of memories. While the general capacity and duration of sensory and short-term memory have been narrowed down through extensive research, determining these limits for long-term memory still remains an open challenge. In this regard, some psychologists have even suggested that its capacity might be infinite, although there is currently no evidence to support this assumption [Sternberg and Sternberg, 2012]. Independent of these properties, several concepts and theories have been developed to explain the internal structure and processes of long-term memory. One of the most comprehensive

overviews that encompasses a classification of different persistent storage systems was proposed by Squire [1992]. As illustrated in Figure 2.8, his concept broadly distinguishes between *explicit (or declarative)* and *implicit (or nondeclarative) memory*.

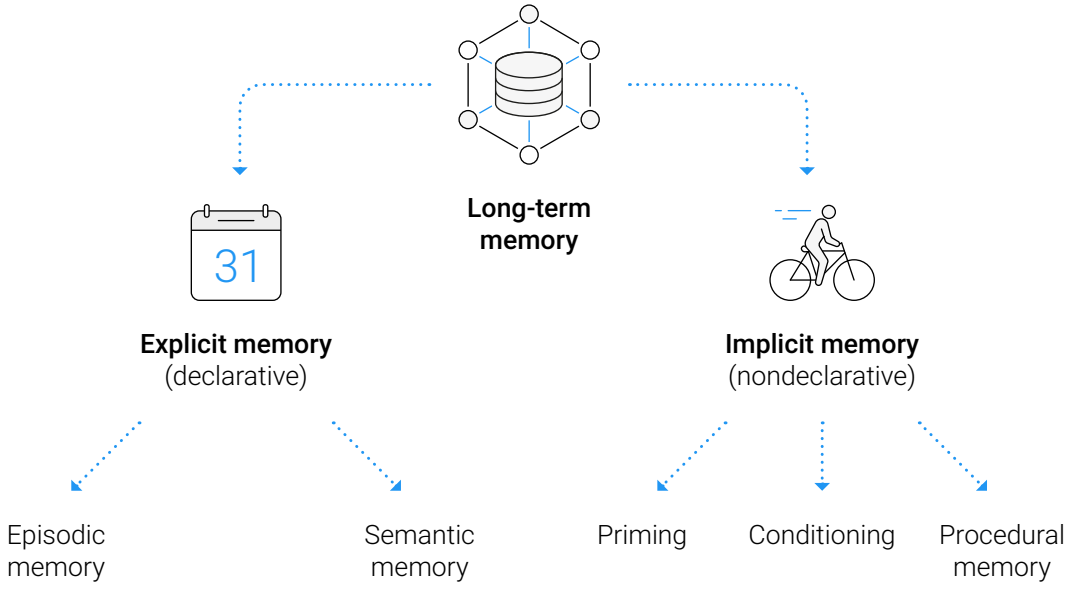


Figure 2.8: Components of long-term memory according to Squire [1992].

On the one hand, explicit memory involves the conscious recall of specific facts and events, such as a person's date of birth and the experience of celebrating it. According to the model by Tulving [1972], this type of stored information can be further separated into *semantic* and *episodic* memory. Thereby, semantic memory refers to static knowledge about facts, concepts, objects, people, processes, and the world. Beyond general information, this also extends to sensory attributes, such as the taste of various foods or the texture of different materials [Baddeley et al., 2015]. In contrast, episodic memory is concerned with storing and retrieving specific events and experiences, such as remembering activities performed during the last vacation or recollecting past instances of family reunions. Additionally, it maintains the temporal and geographical relationships between them, which serve as contextual references during recall. This mechanism allows us to relive certain aspects of the past by remembering the times and places where we originally experienced them [Baddeley et al., 2015].

On the other hand, implicit memory refers to the unconscious recollection of skills, habits, and motor sequences, such as playing an instrument or riding a bike. It allows us to perform these actions automatically without requiring mental effort or conscious thought [Ling et al., 2011]. The acquisition of such memories is usually achieved through various forms of repetition, which can be grouped into the following major cate-

gories: *classical conditioning*, *priming*, and *procedural memory*. Classical conditioning refers to the repeated pairing of a neutral stimulus with a naturally occurring stimulus reflex response. After a few presentations, this association is learned and can be evoked even if the original stimulus is not present [Baddeley et al., 2015]. A popular example of this is Pavlov's [1927] experiment with dogs, where a bell was rung immediately before they were fed. Through repeated association of these stimuli, the sound of the bell alone was sufficient to elicit a salivation response, which would otherwise only occur in the presence of food [Pavlov, 1927].

Priming is another phenomenon of implicit memory that occurs when the appearance of a stimulus subconsciously influences the perception and processing of subsequent stimuli [Baddeley et al., 2015]. In this regard, the initial exposure to the material activates related concepts and associations in memory that affect our thoughts and responses by providing a frame of reference for further inputs. For instance, seeing the terms "towel", "shower", and "shampoo" before being asked to complete the word fragment "s o _ p" would likely lead to "soap" as a response, while the terms "juice", "bread", and "broth" might elicit the word "soup" as an answer. Additionally, priming improves the processing speed for successive presentations of the same or related stimuli [Eysenck and Keane, 2020]. This enables us to identify them more rapidly and efficiently after a previous encounter. Finally, procedural memory refers to the retention of skills, habits, and workflows. They are acquired through continued practice and repetition, which progressively improves our capabilities until they can be performed automatically without conscious effort. Once an activity is learned, procedural memory takes over and handles its execution without the need for active thought. Examples include riding a bike, typing on a keyboard, swimming, or reading a book. Since amnesic patients are also able to perform such activities, one possible explanation is that their impairments only affect explicit memory while their procedural memory remains intact.

2.2.4 Disorders

Memory-related disorders affect a person's ability to encode, store, retain, and recall information. They can range from mild memory lapses to severe impairments interfering with daily functioning. For instance, one of the most impactful conditions is *amnesia*. It involves a significant memory loss that exceeds ordinary forgetfulness and can be extremely disruptive. In general, there are two types of amnesia: (1) *retrograde amnesia*, which refers to the loss of access to previous memories acquired before its onset, and (2) *anterograde amnesia*, which is the inability to store new information in long-term

memory [Baddeley et al., 2015, p. 438]. Amnesia can either be caused by psychological factors involving the temporary suppression of disturbing memories (*psychogenic amnesia*) or physical damage to specific parts of the brain (*organic amnesia*) [Groome, 2014, p. 205]. One of the most common causes of organic amnesia is *Alzheimer's disease*. It is a degenerative disorder that initially appears as a progressive loss of episodic memory but later leads to the broader cognitive decline of various functions with symptoms such as disorientation, language problems, and behavioral issues.

A related condition that is often a precursor to Alzheimer's disease and other forms of dementia is *mild cognitive impairment (MCI)*. It is characterized by a noticeable cognitive decline affecting memory, thinking, decision-making, and language capabilities. Although its symptoms are more severe than regular age-related changes, MCI usually does not significantly impact people's daily lives on its own [Petersen et al., 1999]. However, it increases the risk of developing more serious conditions and should be monitored for potential changes. One disorder that primarily affects short- and long-term memory is the *Korsakoff syndrome*. It results from a thiamine (vitamin B_1) deficiency and is commonly associated with chronic alcoholism but can also originate from malnutrition [Arts et al., 2017]. Over time, the deficiency causes irreversible damage to the brain cells responsible for memory-related functions and leads to symptoms like anterograde and retrograde amnesia. Additionally, affected individuals may also exhibit *confabulation*, where memory gaps are unconsciously filled with reasonable but inaccurate information [Svanberg and Evans, 2013].

2.3 Higher-Order Cognition

While the term cognition encompasses all cognitive processes, including perception and memory, higher-order cognition only refers to higher-level activities such as thinking, reasoning, language understanding, problem-solving, and decision-making [Braisby and Gellatly, 2005, pp. 344ff.]. From the perspective of the information processing model, it involves the processes at the end of the workflow, which generally rely on the results of previous stages (i.e., memory and perception) [Broadbent, 1958]. However, the boundaries of higher-order cognition are relatively fluid because even basic functions, such as perceiving and interpreting visual signals, can require complex cognitive processes [Ragni and Stolzenburg, 2015]. Additionally, emotional states and reactions can influence the outcome of these processes and lead to different results despite similar external conditions, which can not be explained with the computer analogy of the information processing theory. This is another instance where Broadbent's model reaches its

limits and mainly serves to illustrate the existence of this stage. An alternative concept that shares a similar definition is *complex cognition*. According to Knauff and Wolf [2010], it can be described as follows:

> *"As 'complex cognition' we define all mental processes that are used by individuals for deriving new information out of given information, with the intention to solve problems, make decision, and plan actions. The crucial characteristic of 'complex cognition' is that it takes place under complex conditions in which a multitude of cognitive processes interact with one another or with other noncognitive processes".*
>
> —Knauff and Wolf [2010]

In addition to targeting the same mental activities, this definition highlights the goal-oriented nature of complex cognition and emphasizes its ability to coordinate the processes involved in response to changing demands and difficult situations [Funke, 2010]. Due to the similarities between higher-order cognition and complex cognition, we will use both terms throughout the remaining thesis to address higher-level mental activities that rely on the combination and interaction of lower-level processes [Sternberg, 2019].

	Type 1 process (intuitive)	**Type 2 process (reflective)**
Features	*Does not require working memory* *Autonomous*	*Requires working memory* *Cognitive decoupling* *Mental simulation*
Attributes	Fast	Slow
	High capacity	Capacity limited
	Parallel	Serial
	Nonconscious	Conscious
	Biased responses	Normative responses
	Contextualized	Abstract
	Automatic	Controlled
	Associative	Rule-based
	Experience-based decision-making	Consequential decision-making
	Independent of cognitive ability	Correlated with cognitive ability

Table 2.1: Attributes associated with different types of higher cognitive processes according to the dual-process theory. Adapted from Evans and Stanovich [2013].

Independent of the chosen terminology, several theories have suggested a distinction between two general types of processes in higher cognition: the first one is fast, intu-

itive, and automatic, while the second one is slow, reflective, and deliberate [Evans and Stanovich, 2013]. The foundation for these so-called *dual-process theories* was initially proposed by Wason and Evans [1974]. Since then, this concept has gained increasing popularity and has become the focus of current research [Evans and Stanovich, 2013]. An overview of commonly associated attributes for each type of process is provided in Table 2.1. Usually, higher-order cognition involves the parallel activation and coordination of both types. For instance, when making decisions, individuals may rely on intuitive judgments (Type 1) while also engaging in reflective reasoning (Type 2) to consider options and anticipate consequences. Some researchers even went one step further and suggested the existence of two evolutionary distinct brain systems that are responsible for each respective type [Epstein, 1994; Stanovich, 1999]. In addition to the inherited processing attributes, these systems have the following characteristics: while System 1 evolved early and shares similarities with animal cognition, System 2 developed more recently and only contains uniquely human features [Evans and Stanovich, 2013]. Since there is still a debate regarding the validity of these two systems, we will primarily focus on the underlying functions. However, due to the large amount of processes involved, we only examine a limited selection in more detail.

2.3.1 Language

One of the primary aspects that separate us from animals is language. It is an essential part of human life and plays a fundamental role in our culture, technology, and society [Ling et al., 2011]. While it can be used to perform various functions, such as expressing emotions, formulating thoughts, or recording information, its main purpose is communication with others [Crystal, 2008]. In general, language can be defined as a system of symbols (words) and rules (syntax) that determine how they should be arranged to form meaningful sentences [Harley, 2014]. Since each symbol represents a specific meaning, this system enables us to encode and decode information. Apart from these two components, there are several others that contribute to the hierarchical structure of language. As shown in Table 2.2, the smallest units are phonemes, which represent the building blocks and address the acoustics of spoken language [Groome, 2014]. Combining them in the right order results in morphemes, which are sub-structures of words (i.e., stems and affixes) that can alter their meaning. While some words only consist of one morpheme (e.g., cat), others are formed through an arrangement of multiple units (e.g., un-break-able). According to the grammatical rules that define the syntax of a language, these words can be organized into sentences with specific meanings (semantics).

Component	Definition
Phoneme	The smallest unit of speech which contributes to its linguistic meaning: changing a phoneme will change the meaning of a word (e.g., /p/ and /b/ are similar phonemes, but "pit" and "bit" are two different words with distinct meanings).
Morpheme	Units of meaning within words. A word like "descendant" contains a number of morphemes which contribute to its meaning (e.g., "de-" = from, "-scend-" = climb, "-ant" = person with the property of).
Word	Lexical unit which can stand alone in terms of its use in a language and its meaning. Words have meanings which map onto things and ideas: words are the level at which languages convey meaning.
Syntax	Grammatical rules of a language. These rules govern the ways that words can be combined (and declined). Syntax can be independent of meaning: a sentence can be syntactically correct but meaningless (e.g., "colorless green dreams sleep furiously").
Semantics	The meanings of words and the ways that this knowledge is structured and interpreted. Sentences can be ungrammatical but fully semantically comprehensible (e.g., "to sleep I no want").

Table 2.2: Components of language. Based on Ling et al. [2011] and Groome [2014].

Based on these structural insights from the field of linguistics, researchers began to study the underlying mental processes in a sub-branch of cognitive psychology called psycholinguistics. It initially emerged in the second half of the 20th century and primarily focuses on the mechanisms responsible for language acquisition (learning), comprehension (understanding), and production (speaking) [Ling et al., 2011]. Since it is a relatively young discipline compared to other fields like math or physics that have been studied for hundreds of years, there is only little consensus regarding common concepts and most aspects are still subject to debates between opposing views [Harley, 2014]. One example concerns the acquisition of language. During the first half of the 20th century, the dominant approach for studies in psychology was behaviorism, which assumed that all behavior can be acquired through conditioning and reinforcement, as it is simply the result of experience and repetition [Skinner, 1938]. According to Skinner [1957], this also applies to language, which must be learned like any other behavior. However, in a critical review of Skinner's book, Chomsky [1959] proposed an opposing theory. He argued that language could not be acquired through learning alone since children are able to understand and construct utterances they have not heard before [Groome, 2014]. Therefore, Chomsky concluded that humans must possess some form of innate knowledge about language, which further develops under suitable conditions, similar to the ability to walk [McBride and Cutting, 2019].

Another area of uncertainty regards the perception of speech. On the one hand, some researchers suggest it requires specialized processes that are distinct from the perception of other stimuli [Ling et al., 2011]. For instance, Liberman et al. [1967] demonstrated that speech signals are typically assigned to separate phoneme categories (e.g., either /b/ or /d/) without any variations in between (categorical perception). At the same time, regular sounds are usually associated with a continuous spectrum (e.g., pitch or loudness), which is why Liberman et al. [1967] hypothesized that phonemes must be decoded through different mechanisms compared to non-speech sounds. On the other hand, opposing models suggest that speech perception involves the same principles and processes used to acquire regular stimuli [Carbonell and Lotto, 2014]. In this regard, multiple attempts have been made to disprove competing hypotheses by illustrating the occurrence of categorical perception in non-speech sounds [Harnad, 1990; Mirman et al., 2004]. Since there is evidence for both sides of the argument, more universal concepts and solutions might be found somewhere in between.

Regardless of the mechanisms with which speech signals are perceived, several theories have been proposed that address the further processing steps. One example is the *TRACE model* by McClelland and Elman [1986]. It assumes that once the first phoneme of a word is heard, various related words with the same starting sound become activated, similar to interconnected neurons in a neural network. With each subsequent phoneme, this set of potential word candidates gets reduced until only one remains [Ling et al., 2011]. Thereby, contextual information can influence the activation of suitable candidates, which might lead to different recognition results of the same stimuli (top-down and bottom-up interaction). Once words have been identified from the perceived signals, the sentences they form are analyzed to extract the represented meaning. For that, several models have been proposed to explain processes involved in comprehension. One of them is the *constraint-based theory* by MacDonald et al. [1994]. It assumes that the initial interpretation of a sentence is based on multiple sources of information, such as syntactic and semantic knowledge. In this regard, each source has its own constraints, which limit the number of potential results. During analysis, suitable interpretations are activated and ranked according to their compliance with the associated constraints, similar to the words in the TRACE model [Eysenck and Keane, 2020]. The final meaning of the sentence is then selected based on the structure with the highest activation.

Besides word recognition and sentence comprehension, language production is another area of interest in psycholinguistic research. While theorists agree that it generally involves the stages of *conceptualization* (determining what to say), *formulation* (translating concepts into linguistic form), and *articulation* (phonetic planning and execution),

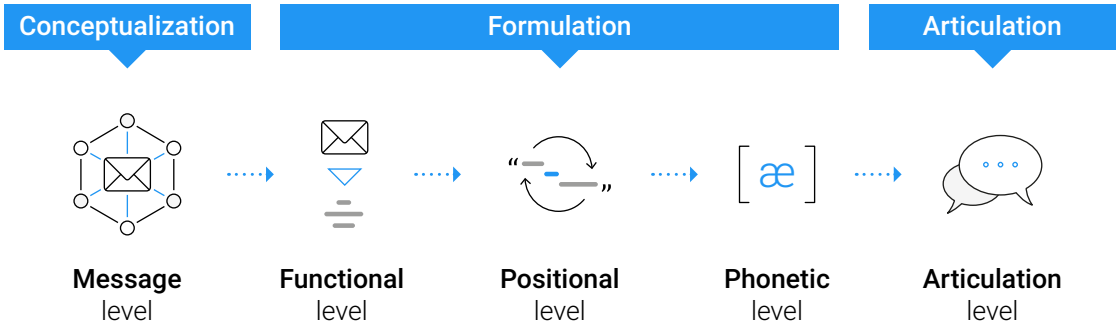


Figure 2.9: Speech production model by Garrett [1975].

there is still a debate regarding the nature and interactions of the processes involved [Eysenck and Keane, 2020]. On the one hand, some theories argue that speech production is serial and occurs in an orderly fashion without any interactions between stages. An example of this is the model by Garrett [1975]. As shown in Figure 2.9, he divided the general stages into five independent levels: (1) the *message level*, where concepts and thoughts are gathered; (2) the *functional level*, at which these concepts are expressed through semantic representations (words) and assigned to syntactic roles (e.g., verb, subject, and object); (3) the *positional level*, where words are ordered to form sentences; (4) the *phonetic level*, at which these sentences are transformed into phonetic sequences (including speed, prosody, and intonation); and (5) the *articulation level*, where the phonetic sequences are realized through motor instructions of the vocal apparatus [Groome, 2014]. According to Garrett's model, the intermediate results of each level are processed sequentially and in complete isolation by the subsequent layers without any influences or interactions between them.

On the other hand, alternative models suggest that the processes involved in language production are highly interactive and occur in parallel. A popular example is the spreading activation theory by Dell [1986], which consists of four levels: (1) the *semantic level*, which handles the meaning of what is to be communicated; (2) the *syntactic level*, which is responsible for the grammatical structure of the words; (3) the *morphological level*, which organizes the morphemes that are part of the planned sentence; and (4) the *phonological level*, which produces the phonemes for the indented utterances [Ling et al., 2011]. While this structure is relatively similar to the model by Garrett [1975], a major difference is that the processing occurs simultaneously across all of these levels. Additionally, activations within one level can spread and influence others, similar to the structures in MacDonald et al.'s constraint-based theory and McClelland and Elman's TRACE model [Eysenck and Keane, 2020].

2.3.2 Problem-Solving

Solving problems is a common and essential part of our daily lives that generally involves three main components: (1) the initial state, (2) the goal state that should be achieved, and (3) the actions or operations that need to be performed to get from the current to the target state [Ling et al., 2011]. Based on these components, there are two primary types of problems. The first are *well-defined problems*, where all states, operations, and conditions are fully specified. In contrast, the second type are *ill-defined problems*, where certain aspects are unclear or not defined at all. While the majority of challenges we encounter on a daily basis are ill-defined, most research, especially in the early days, has focused on well-defined problems, since there is usually an optimal strategy to solve them [Eysenck and Keane, 2020]. One of the first psychologists to study the processes involved in problem-solving was Thorndike [1898]. After observing the almost random behavior of cats under experimental conditions[1], he argued that their approach to finding solutions was through trial and error.

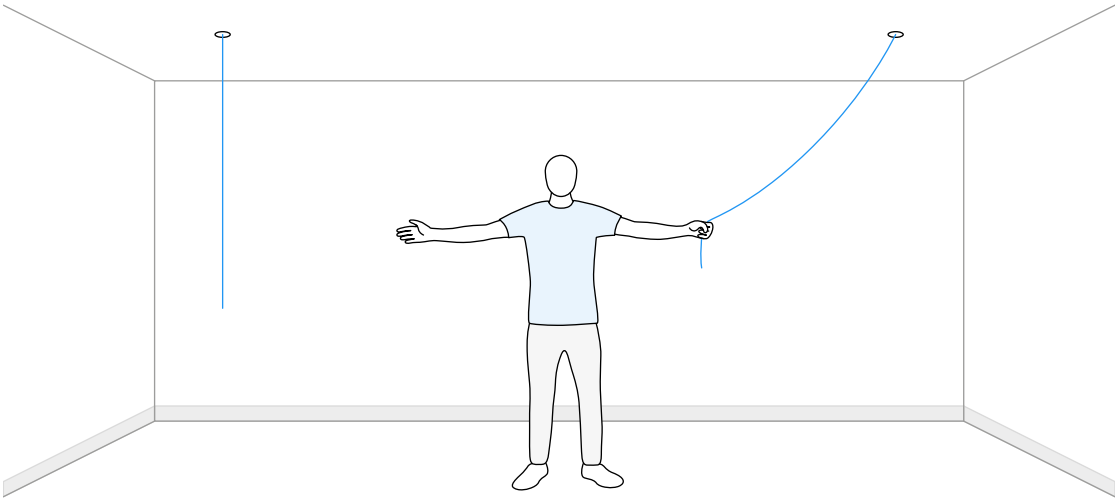


Figure 2.10: Two-string problem used by Maier [1931].

While it is true that even humans sometimes apply this method to solve problems, researchers from the field of Gestalt psychology advocated for alternative explanations.

[1] Thorndike [1898] placed cats into cages with special mechanisms that allowed them to open the doors from inside. Initially, they performed various kinds of behaviors until they randomly found the appropriate solution. In subsequent trials, they gradually learned the necessary steps until they were able to escape almost immediately.

Based on a series of experiments with apes[2], Köhler [1925] argued that they were able to find solutions through a sudden restructuring of the problem, which he called "insight". This idea inspired the work of Maier [1931], who used the "two-string problem" illustrated in Figure 2.10 to study insight among humans. In his experiment, participants were brought into a room with various objects on the floor (e.g., poles, pliers, and cables) and two strings hanging from the ceiling. The task was to tie both strings together, but they were placed in such a way that made it impossible to reach one of them while holding the other. A solution to this problem was to tie one of the objects to the end of a string and make it swing like a pendulum. While around 39% of participants solved the task on their own, others made little progress even after intense periods of thinking. In these cases, Maier inconspicuously brushed against one of the strings to make it swing, which triggered an "ah-ha" experience in most participants and led them to the correct approach. Based on these observations, he concluded that insight and problem solutions could be facilitated through external cues [Eysenck and Keane, 2020].

Along with the emergence of the information processing view in the 1960s, Newell and Simon [1961, 1972] worked on a more systematical approach, which resulted in a computer program called the General Problem Solver (GPS). They argued that it could be used to simulate the processes involved in human problem-solving and demonstrated that the solutions to most well-defined problems could be found by breaking them down into a series of stages [Eysenck and Keane, 2020]. In the first stage, a *problem space* is constructed, which represents both the initial and the target state, as well as instructions, constraints, and other information from long-term memory that might be relevant. The idea behind this step is to create a space with all possible states of a problem that can be searched for appropriate solutions. In the second stage, suitable operators (actions) are selected, which transform the initial state and achieve a specific (sub-)goal. The implementation of the selected actions in the third stage then leads to a new state within the problem space. Finally, this new state gets evaluated in the fourth stage, and if it corresponds with the target state, a potential solution has been found. Otherwise, the previous stages are repeated until this is the case [Groome, 2014]. To validate their theory, Newell and Simon asked participants to describe their thoughts while solving various problems. After that, they compared the verbal protocols with the steps performed by the GPS program and found a high degree of similarity.

[2] Köhler [1925] provided an ape with two sticks, both of which were too short to reach the bananas placed outside the cage. While seeming lost at first, the ape eventually combined both sticks and was able to obtain the bananas. A similar behavior was observed when provided with several crates and bananas hanging from the ceiling.

While evaluating all possible sequences of operations and their resulting states is suitable for simple problems, this method can become very time-consuming and almost impossible for complex challenges. Instead, a process called *problem reduction* can be used, which divides the problem into smaller sub-problems that are easier to solve. One example of this approach is the *means-ends analysis*. By working backward from the goal, it identifies a sub-goal that reduces the difference between the initial and the target state. Following that, suitable mental operators are selected to achieve the sub-goal [Groome, 2014]. Another essential strategy for solving problems is the application of analogies. It involves using knowledge from related tasks and domains to find solutions for the current problem. This process can be broken down into the following three phases: (1) recognizing that the problem shows similarities to a previously solved task, (2) retrieving the source analogy from long-term memory and abstracting the general properties that were used to solve it, and (3) mapping these elements from the source analogy to the target problem [Mayer, 2013]. Besides providing transferable solutions, information gained from analogies can also be used to find alternative perspectives, establish relationships between entities, or draw new conclusions.

2.3.3 Emotions

Emotions play a fundamental role in our daily lives. They are deeply connected with most cognitive processes and can mutually influence each other's outcomes. Although there is no universally accepted definition, emotions are generally associated with the following characteristics: (1) subjective experiences of internal cognitive states (feelings); (2) physiological reactions to these experiences (e.g., raised heart rate or faster breathing); and (3) behavioral responses, including facial and vocal expressions (e.g., smiling or screaming) [Braisby and Gellatly, 2005]. In this context, several related terms, such as *mood* and *affect*, are often used interchangeably, despite having slightly different meanings. While emotions typically refer to intense short-term events and experiences, moods usually have a lower intensity and last for longer periods [Hills, 2016]. In contrast, affect is primarily used as an overarching term that encompasses all other notions. Besides that, there are two general approaches to characterize different types of emotions. The first one is the categorical view, which assumes that a small set of discrete emotions (e.g., anger, fear, sadness, disgust, and happiness) serves as the foundation for all affective experiences. Similar to the mixture of colors, it is argued that different combinations within this set can produce all other states. One of the most prominent proponents of this view is Ekman [1984], who initially proposed the *basic emotion theory*. According to his findings, these basic emotions can be recog-

nized universally across cultures and are typically associated with specific expressions and physiological responses [Ekman, 1999a].

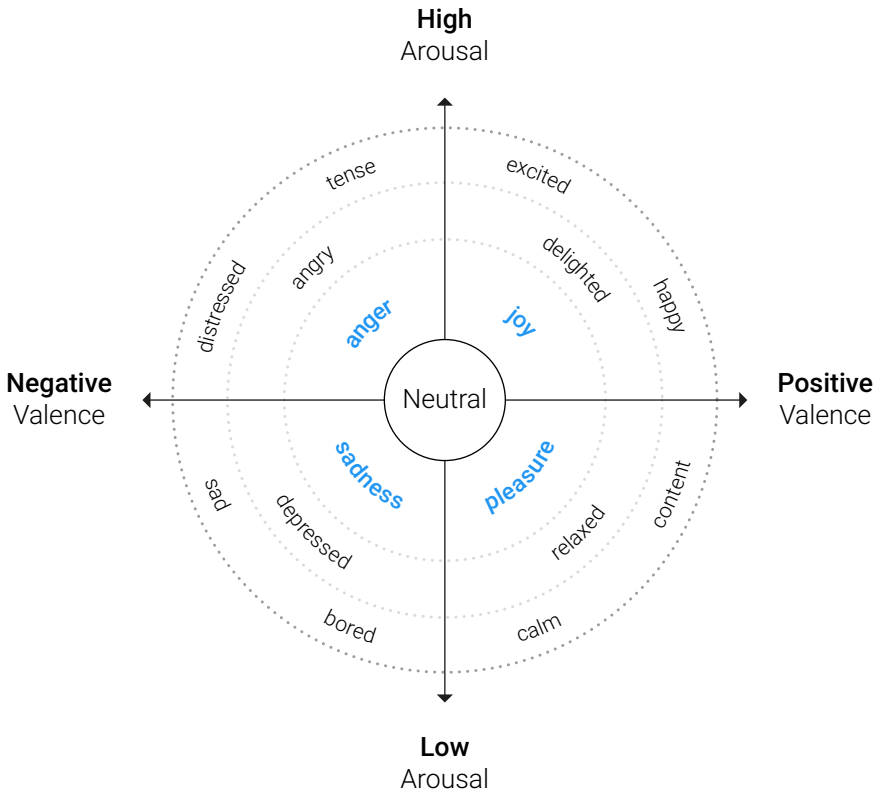


Figure 2.11: Circumplex model of affect by Russell [1980].

Another method for specifying emotions is through dimensional models, which represent each emotion as a point along continuous scales in a multidimensional space. Typically, these models use two axes to characterize emotions in terms of *valence* and *arousal*. One example of this is the *circumplex model of affect* by Russell [1980]. As shown in Figure 2.11, it distributes various emotional states around a circular pattern in a two-dimensional space. While the valence axis in this model defines the pleasantness of an emotion (positive vs. negative), the arousal dimension is concerned with the level of agitation or activation (calm vs. aroused). Apart from these two variables, other models propose the inclusion of further dimensions, such as *dominance* [Mehrabian, 1995, 1996]. It reflects how much a person feels in control and can help to distinguish emotions, such as anger and fear, which would be relatively similar in a two-dimensional model (negative valence and high arousal). One advantage of this dimensional approach is that categorical emotions can still be represented as points or areas in the multidimensional space. For instance, happiness is located in the top-right quadrant, while depression and boredom fall into the bottom-left corner [Eysenck and Keane, 2020].

Since both categorical and dimensional models are only sufficient to distinguish different types of emotions, several approaches have been proposed to explain the processes involved in their occurrence. In this regard, one of the most influential concepts is the *appraisal theory* by Lazarus [1966, 1991]. It argues that emotions are an elicited response to mental evaluations (appraisals) of perceived, remembered, or imagined situations and events [Roseman and Smith, 2001]. For instance, feeling sadness when a job application gets declined might be a reaction to the appraisal that something desired has not been achieved. According to Lazarus [1966], there are three types of evaluations: *primary appraisal*, *secondary appraisal*, and *reappraisal*. Initially, primary appraisal referred to assessing the impact of a situation or event on a person's well-being. However, in a later revision of his theory, Lazarus [1991] expanded it to include the following three components: (1) *goal relevance*, which determines whether the current circumstances are related to an individual's goals; (2) *goal congruence or incongruence*, which asses whether a transaction prevents or facilitates personal goals; and (3) *type of ego-involvement*, which analyzes the implications on various identity-related aspects, such as self-esteem, moral values, ideals, life-goals, and the well-being of others.

Similarly, secondary appraisal was initially defined as the subsequent evaluation of options and resources available for coping and was later expanded to include the following components: (1) *blame or credit*, which determines who is responsible or accountable and can be directed internally (self) or externally (other people or groups); (2) *coping potential*, which evaluates the prospects of different coping strategies based on situational demands; and (3) *future expectancy*, which estimates whether the current circumstances are likely to change in the future [Power and Dalgleish, 2016]. Finally, reappraisal refers to the continuous monitoring and evaluation of events and appraisal results. This includes modifying primary and secondary appraisals in response to changed conditions. While these forms of appraisal imply deliberate and conscious processing, they can also occur automatically and unconsciously in certain situations [Eysenck and Keane, 2020]. Based on the central idea that emotions are elicited in response to subjective evaluations of events, current research suggests a direct correlation between distinct appraisal patterns and emotional states [Roseman and Smith, 2001]. More specifically, this means that any situation with the same evaluation outcome will evoke the same emotion. However, it is also possible that an identical situation is appraised in different ways, which results in changed perspectives and altered emotions (e.g., frustration after a certain period because a desired condition is still not met). Further information regarding current challenges and approaches associated with the recognition of these different emotional states can be found in the literature review by Can et al. [2023].

2.3.4 Disorders

Impairments and disorders of higher-order cognition involve disruptions to complex mental processes, which are crucial for language, problem-solving, decision-making, reasoning, and emotion regulation. They can result from a variety of causes, including neurological damage, psychological conditions, neurodevelopmental disorders, and age-related changes. While covering all existing impairments exceeds the scope of the present thesis, we still provide selected examples to illustrate the range of potential symptoms. One of the most common conditions among older adults is *dementia*. It is characterized by a general decline of cognitive functions and can result from various disorders, like the previously mentioned Alzheimer's disease and mild cognitive impairment (see Section 2.2.4). Aside from memory deficits, typical symptoms include confusion, disorientation, emotional problems, inappropriate behavior, difficulties with language, personality changes, and a reduced ability to solve problems and make decisions. While Alzheimer's disease accounts for 60-70%[3] of cases, there are other forms of dementia, such as *vascular dementia* (damage to blood vessels or reduced blood flow), *Lewy body dementia* (abnormal deposits of alpha-synuclein protein), and *frontotemporal dementia* (damage to the frontal and temporal lobes of the brain).

Another common cognitive condition whose prevalence has further increased due to the COVID-19 pandemic is *depression*. It is characterized by a prolonged mood disorder that negatively impacts people's thoughts, feelings, and behaviors. Potential symptoms include changed appetite, hopelessness about the future, loss of interest, low energy, reduced pleasure, sadness, tiredness, and suicidal thoughts. The condition can be caused by a variety of biological, environmental, medical, psychological, and social factors, such as a genetic predisposition running in the family, chemical imbalances within the brain, prolonged exposure to stress, traumatic events, social isolation, significant life changes, or side effects from medication. Depending on the circumstances leading to its development, there are different types of depression with distinct properties. For example, *major depressive disorder* is characterized by the pervasive impact of symptoms on a person's daily life over a period of more than two weeks. In contrast, *persistent depressive disorder* (also known as *dysthymia*) is a chronic form of depression with less severe but longer-lasting symptoms (usually more than two years). Another variation is *bipolar disorder*. It involves alternating episodes of depression and highly elevated mood (*mania*), each lasting between a few days and multiple weeks [Anderson et al., 2012; American Psychiatric Association, 2013].

[3] `https://who.int/news-room/fact-sheets/detail/dementia`

Apart from conditions with relatively broad symptoms, some disorders only affect individual functions and processes. For instance, *aphasia* is a communication-related impairment that impacts people's ability to comprehend language (*receptive aphasia*), produce sentences (*expressive aphasia*), or both (*global aphasia*) [Groome, 2014, p. 338]. It is characterized by the dysfunctional conversion between mental representations and structured language elements. The disorder is caused by damage to specific brain regions, usually resulting from a stroke or head injury, but can also develop over time in conjunction with cerebral tumors. It can affect all types of language, including spoken words, written symbols, and even visual sign gestures used by deaf people. Additionally, it can compromise various aspects of language, such as sentence structure (syntax), formation of words (morphemes), and correct pronunciation (phonemes) [Damasio, 1992]. The prevalence and severity of these potential symptoms depend on the specific manifestation of the condition and can vary between individuals. While one person might only show signs of non-fluent speech, another might have problems comprehending the meaning of certain words and producing coherent sentences.

2.4 Summary

This chapter provided an overview of the general stages involved in human cognition to establish a fundamental understanding of their mechanisms and reveal potential areas for enhancement. To this end, it introduced general concepts and theories from the field of cognitive psychology and discussed common disorders of each process that can serve as foundations for assistive augmentation approaches. Initially, the first stage of human cognition starts with the perception of stimuli through sensory organs. It involves various mechanisms, including knowledge, sensation, and attention, that are responsible for acquiring, organizing, and interpreting incoming signals about the environment and the internal states of our bodies. Parts of the resulting details are then transferred to the memory storage, which performs complex operations to encode, retain, and retrieve appropriate representations. Depending on the quality of these processes, the encoded information is stored in different types of memory structures, ranging from transient sensory registers with limited capacity to large persistent repositories for long-term archival. Combined with existing knowledge and memories, these records provide the foundation for higher-level mental activities in the cognition stage. Examples include thinking, reasoning, language understanding, problem-solving, decision-making, and emotional reactions, which can influence the outcome of other processes and lead to different results despite similar conditions.

Chapter 3

Non-Verbal Signals

Verbal communication is our primary method of conveying information to other people. It involves encoding the intended meaning into words, sentences, and spoken language. On the receiving end, the message is perceived with our auditory senses and decoded using the same rules as during encoding process (see Section 2.3.1). Apart from speech, there are several non-verbal signals that convey information about an individual's personality, feelings, mental state, and other properties [Richmond and MacCroskey, 1995]. Since capturing and analyzing these types of information can produce valuable insights for assistive augmentation systems, this chapter provides an overview of available signals. According to Poggi and Francesca [2010], a *signal* can be generally described as a stimulus, such as a behavior, a morphological trait, a chemical trace, an electrical pattern, or a series of events, that is produced by an emitter (i.e., an individual or a group of people) and can be interpreted by a receiver (i.e., other humans or sensing devices) to extract its meaning. While some signals are emitted consciously and deliberately, others simply occur as a byproduct of events or an automatic physiological reaction that can not be controlled.

In this regard, the authors distinguish between informative and communicative signals. Typically, a signal is *informative* if the emitter produced it without the goal, intention, or biological function of conveying the information to a receiver. This is often the case for physiological signals and accidental stimuli that occur due to a random combination of events. In contrast, *communicative* signals are emitted with the goal of communicating specific information to the receiver. Examples include conscious intentions, such as gestures to symbolize what was said (e.g., winking after a joke), as well as behavior with a lower level of awareness, like facial expressions that indicate a person's emotion

(e.g., raising an eyebrow when confused). Although these signals are universal across all humans, their meaning and interpretation can differ depending on the cultural, regional, and situational context [Poggi and Francesca, 2010]. For instance, slurping noodles can communicate a sign of appreciation in Japanese culture, but the same behavior is often considered rude or inappropriate in Western regions. Apart from distinguishing between informative and communicative signals, they can also be categorized as behavioral and physiological cues based on their origin. While *behavioral cues* are often performed consciously to communicate a specific meaning, *physiological cues* usually have an informative nature and are emitted involuntarily or subconsciously. The details of these two types are described in the following sections.

3.1 Behavioral Cues

The term *behavioral cue* typically refers to an externally observable stimulus that is consciously or intentionally emitted by a person to convey a non-verbal message to the receiver. It usually complements verbal communication and only lasts for a short period of time (milliseconds to minutes) [Vinciarelli et al., 2009]. As shown in Figure 3.1, one or more behavioral cues can occur simultaneously to indicate a shared social signal. According to Poggi and Francesca [2010], *"a social signal is a communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions, social emotions, social attitudes, and social relationships"*.

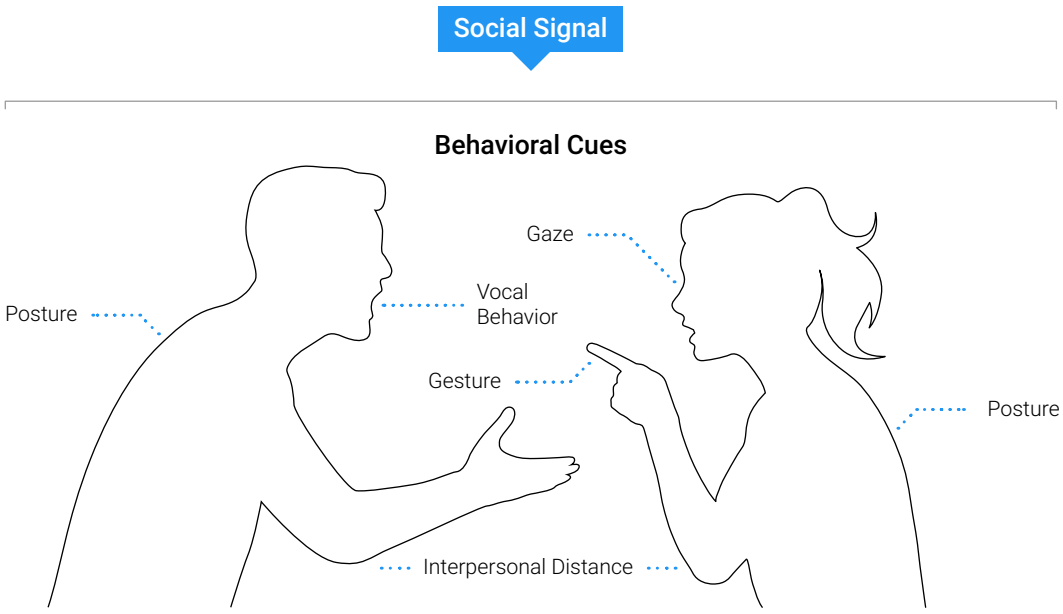


Figure 3.1: Social signal composed of multiple behavioral cues.

Since social signals can express a variety of meta-information about a person's current state, all involved behavioral cues should be considered to correctly interpret their intended meaning (i.e., analyzing a behavioral cue in isolation could lead to false conclusions). To further clarify potential differences between social signals, Ekman and Friesen [1969b] classified non-verbal behavior into the following five categories:

Emblems are non-verbal acts that directly represent spoken words and phrases. They possess a well-known definition and translation within a particular group, class, or culture, allowing them to replace verbal communication entirely in specific contexts. For instance, a thumbs-up gesture might be used to signify approval instead of saying *"Good job!"* in a loud or noisy environment. In most cases, emblems are performed consciously and intentionally, but there are also certain situations where they occur without people's awareness (e.g., forming a fist during anger).

Illustrators are body movements that relate to and visually emphasize or clarify spoken content. They act as a complementary communication channel, enhancing the understanding and impact of verbal messages. Examples include pointing to an item while talking about it or spreading both arms to indicate the dimensions of a large object. In terms of conscious usage, they are relatively similar to emblems, although people typically perform them with slightly less awareness and intentionality.

Affect Displays are behaviors and expressions that communicate a person's emotional state. They are typically shown with a culture-independent set of facial movements that exist for each primary type of affect, such as happiness, surprise, fear, sadness, anger, and disgust. Although some body movements (e.g., trembling or being startled) can also indicate an affective state, they mainly occur as behavioral consequences in response to the underlying emotion rather than displaying affect on their own.

Regulators are behavioral cues that manage the flow and pacing of conversations between two or more people. They provide feedback to speakers and listeners regarding the direction and control of interactions. For instance, nodding can encourage a speaker to elaborate on their current topic, raising an eyebrow can signal a listener's request for clarification, and looking away can indicate the intention to disengage from a conversation. Other examples include eye contact, posture shifts, and hand movements, which are usually performed with less awareness than emblems or illustrators.

Adaptors are movements and behaviors that individuals use to control their emotional state, satisfy bodily needs, or manage physical activities. They are usually performed in social situations to cope with boredom, discomfort, or anxiety. Examples include nail-biting, leg shaking, or fidgeting with objects. The meaning and interpretation of adaptors can vary depending on the context and cultural background. Similar to habits, they occur with low awareness and are not intended to communicate a message.

While even Ekman and Friesen [1969b] emphasize that these categories are not complete or final, they provide a solid foundation to identify the different properties of non-verbal behaviors. This information can be used in assistive augmentation systems to draw conclusions about a person's state and intentions based on behavioral observations. In this regard, certain acts, movements, and expressions can be associated with multiple categories. For instance, emblems can include affect displays or adaptors with culture-specific meanings. To better understand the behavioral cues related to individual modalities, we take a closer look at each of them in the following sections.

3.1.1 Facial Expressions

Facial expressions are one of the primary channels for communicating non-verbal information, such as attitudes, moods, and intentions. They are produced by contracting and relaxing different groups of facial muscles and provide a dynamic window into people's emotional states [Knapp et al., 2013, p. 258]. While they can be consciously controlled to pretend or suppress certain conditions, there are many involuntary expressions that reveal our true feelings and intentions to others. Ekman and Friesen [1969a] call them *micro facial expressions*, which are performed unconsciously and only last for a very short duration (less than half a second). When detected (e.g., with a slow-motion camera), they can be used to spot deceptive behavior and even identify lies. However, in accordance with the findings of Haggard and Isaacs [1966], Ekman [2009] discovered that these microexpressions look the same for both deliberate concealment and emotional repression, which is why he concluded that they can only be differentiated by considering the context of their occurrence.

Overall, facial expressions have been analyzed for more than 150 years [Ekman, 1999b]. One of the first and most influential researchers in this field was Darwin [1872], who argued that these adaptive facial responses formed as part of human evolution to overcome survival-related challenges and facilitate non-verbal communication. Based on his evolutionary theory, he suggested that certain expressions became innate and are now universally present across all humans. These assumptions were supported by Ekman

and Friesen [1971] in studies with people from an isolated population of New Guinea. In their experiments, they told participants a story and asked them to select the most appropriate photo from a collection of pictures showing different facial expressions of Western people. Despite having minimal contact with other cultures and populations, they associated the respective stories with the same expressions as people from Western civilizations. Likewise, the facial behavior of New Guinean individuals was correctly recognized by citizens from the United States, which indicates a universal presence and understanding of basic emotions, such as anger, happiness, fear, surprise, sadness, and disgust [Ekman and Friesen, 1971; Ekman, 1984, 1992].

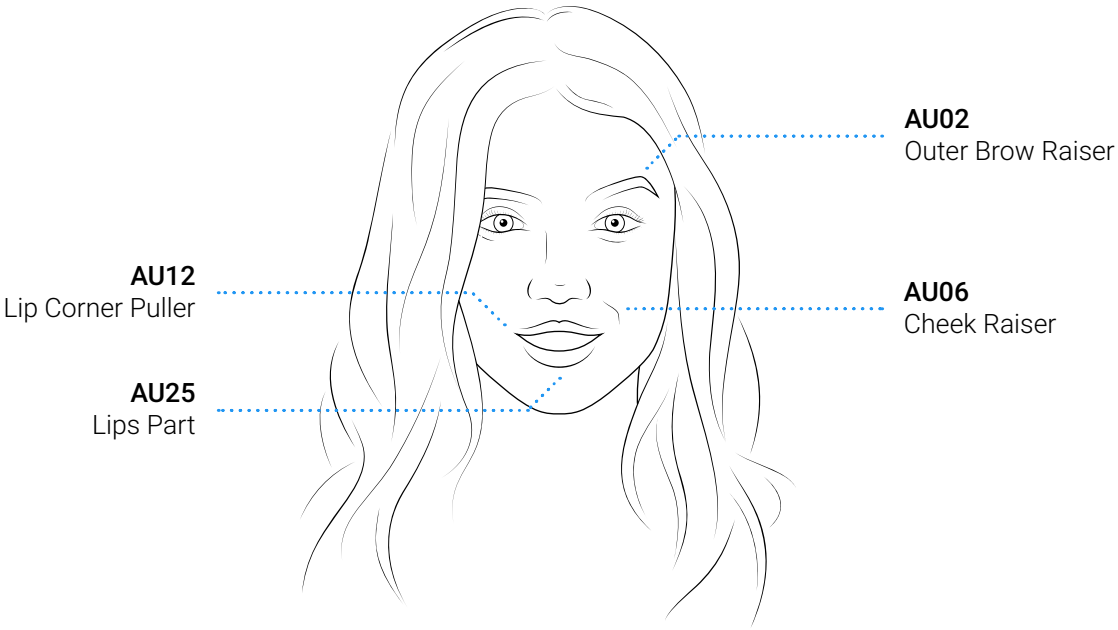


Figure 3.2: Examples of activated facial action units.

To further improve the objective analysis of facial expressions, Ekman and Friesen [1978] introduced the Facial Action Coding System (FACS). It enables observers to describe nearly any possible facial configuration based on the position and occurrence of 64 so-called *Action Units (AU)*. Each AU is assigned to a specific muscle group and defines its anatomic state (i.e., contracted or relaxed). For instance, raised lip corners (e.g., when smiling) are denoted with AU 12 (see Figure 3.2), while lowered lip corners (e.g., during sadness) are labeled with AU 15. The precise distinction between different states allows trained coding experts to describe almost any facial expression by referencing the present action units. Due to the advantages of this method, action units have also been used as feature values in combination with machine learning approaches to infer information about analyzed individuals. For that, several software solutions, such as

SHORE [Ruf et al., 2011] and OpenFace [Baltrusaitis et al., 2016], can provide access to automatic face tracking, landmark estimation, and action unit recognition. However, with the widespread adoption of neural networks, newer approaches have transitioned to detect the desired information based on raw facial images (e.g., Toisoul et al. [2021]).

3.1.2 Gaze Behavior

In addition to enabling the perception of visual information as their primary function, the human eyes can provide valuable insights into people's cognitive processes and mental states. Depending on where we direct our gaze and how long we look at something or someone, we not only reveal information about ourselves but also emit non-verbal signals that establish social connections, regulate the flow of conversations, and communicate unspoken messages [Cañigueral and Hamilton, 2019]. During social interactions, eye contact with other people can be a fundamental mechanism to coordinate the timing of speaking turns and prevent overlaps or interruptions. In this regard, Kendon [1990] found that speakers often look away at the beginning of a turn and resume eye contact at the end to indicate the possibility of a role change. In contrast, direct eye contact by listeners can express attention, interest, and engagement with the current topic. However, too much eye contact can make us feel uncomfortable, which is why maintaining the correct amount requires a delicate balance between mutual gaze and looking away [Argyle and Cook, 1976]. Apart from direct eye contact, the gaze direction can also be used to cue other people's attention or shift the focus of social interactions to a different target [Frischen et al., 2007].

Outside social settings, eye movement patterns often reveal information about cognitive activities [Van der Stigchel et al., 2006]. For instance, shifting the gaze point between various objects can indicate their consideration during problem-solving tasks, while longer fixations on specific items can signal increased cognitive load or deeper processing of thoughts [Just and Carpenter, 1976]. One of the first researchers to analyze these gaze patterns was Buswell [1935]. He showed that eye movements differ distinctively during a visual search task on an image compared to a free viewing task with no instructions. Several years later, Yarbus [1967] confirmed that the visual task indeed plays an important role in the observed scan paths and patterns. Since then, extensive research has been conducted regarding the inference of cognitive processes based on the analysis of eye movement behavior. Example applications include the automatic recognition of fatigue [Eriksson and Papanikotopoulos, 1997], mind wandering [Bixler and D'Mello, 2016; Drummond and Litman, 2010], and decision-making [Gidlöf et al., 2013].

Unlike the previous types of ocular behavior, which can be consciously controlled, pupil dilation is a completely involuntary physiological response. This characteristic makes the information derived from its analysis more trustworthy and reliable compared to other gaze-related signals. As indicated by studies from Kahneman et al. [1969], pupil size also correlates with cognitive load and enables conclusions about people's current mental state. Additionally, it can reflect affective processing [Partala and Surakka, 2003] and memory-related functions, such as encoding and retrieval of information [Goldinger and Papesh, 2012]. While capturing the different types of gaze behavior initially required complex and intrusive hardware setups (e.g., static cameras and fixed head positions), technology has now advanced to a point where eye-tracking devices can be worn like regular glasses [Jacob and Karn, 2003]. This enables the ubiquitous collection and analysis of gaze-related data and facilitates its usage in assistive augmentation systems.

3.1.3 Vocal Cues

Beyond the literal content of speech, paralinguistic or vocal behavior typically refers to *how* something is said [Knapp et al., 2013]. It can complement a spoken message and influence its intended meaning through various vocal cues. Examples include acoustic variations of pitch, loudness, rhythm, intonation, and speech rate. These prosodic features (also known as *voice quality* [Richmond and MacCroskey, 1995]) are essential for interpersonal communication and provide additional layers of meaning to verbal content. For instance, statements can be turned into questions by changing the pitch, and the meaning of utterances can even be inverted through sarcastic intonations. Additionally, these characteristics can provide information about the speaker's mental and emotional state. While feelings like anger and fear are frequently accompanied by vocal bursts (e.g., shouting) [Vinciarelli et al., 2009], boredom typically involves a lower speech rate and monotone rhythm [Scherer, 2003]. Another type of vocal behavior is *linguistic vocalization*, which includes sounds such as *"ehm"*, *"uhm"*, or *"uh-huh"* to fill pauses when the right words or answers do not come to mind and require more time to think. These expressions can also be used for so-called *back-channeling* behavior and indicate agreement or engagement with the speaker [Shrout and Fiske, 1981].

Furthermore, *non-linguistic vocalizations* like laughing, crying, groaning, or whispering can provide insights about a person's state and attitude towards the current situation [Vinciarelli et al., 2009]. While others can easily recognize the occurrence of these vocal cues, interpreting their meaning is not always as straightforward and depends on the situational context [Anikin et al., 2018]. For instance, crying is typically related

to sadness but can also occur in moments of overwhelming happiness. Consequently, considering other modalities can be an important measure to identify the true meaning of these expressions. Apart from that, silence is another non-verbal cue that can indicate hesitation, the need to think about a proper response, or difficulties in dealing with a conversation [Richmond and MacCroskey, 1995]. It is also used as a sign of respect, can emphasize subsequent statements, and influences the behavior of others (e.g., "silent treatment" of children). Since all of these vocal cues are related to specific sounds (or their absence in case of silence), capturing them with microphones is a relatively straightforward method. However, recognizing their occurrence among speech and other sounds from the recorded audio streams can be a rather challenging task. Fortunately, several publicly available tools like openSMILE [Eyben et al., 2010], PRAAT [Boersma, 2001], and EmoVoice [Vogt et al., 2008] can be utilized to facilitate this process. Additionally, established neural network models such as Wav2vec 2.0 [Baevski et al., 2020], Audio Spectrogram Transformer (AST) [Gong et al., 2021], and HuBERT [Hsu et al., 2021] can be fine-tuned to recognize vocal cues and their meaning based on raw audio data (e.g., Wagner et al. [2023]).

3.1.4 Gestures and Posture

The term *gesture* refers to conscious or unconscious movements of hands, arms, and other body parts [Poyatos, 1984]. They can emphasize verbal messages, illustrate concepts, reveal intentions, and convey emotions [Vinciarelli et al., 2009]. For instance, covering the facial region with your hands can indicate embarrassment [Costa et al., 2001], and showing a thumbs-up gesture can signify approval [Pease and Pease, 2008]. According to McNeill [1992], more than 90% of gestures are performed during speech and relate to at least one of Ekman and Friesen's [1969b] non-verbal behavior categories described in Section 3.1 (i.e., emblems, illustrators, etc.). To further classify the different types of movements, he refined their categorization and introduced the following dimensions: iconics, metaphorics, deictics, and beats [McNeill, 1992].

Iconics are gestures that directly represent specific objects, actions, or events. They complement spoken words and demonstrate what was said by mirroring the physical properties or movements associated with the referenced constructs to facilitate comprehension. Examples include forming a circular shape with your hands to illustrate the surface of a ball or making a wave-like motion to describe an ocean current.

Metaphorics also convey meaning through hand and body movements but represent more abstract concepts and ideas than iconics. They create a visual analogy to aid the expression and understanding of notions that are not physically tangible. For example, balancing your hands like a scale might indicate the consideration of different options, or making an upward motion can symbolize an increase in value.

Deictics are pointing gestures directed at another person, object, or location. They help establish the context of conversations and guide the listener's attention to a referenced point. Typically, deictics are performed with the index finger, but other body parts, such as the head, nose, or eyes, can also be used.

Beats are simple, often rhythmic movements that emphasize specific words or phrases. They usually occur in conjunction with important elements in spoken language but do not represent a concrete meaning or concept by themselves. Examples include pounding your fist on the table to emphasize a statement or using a chopping motion with your hand to mark the end of a point.

Although these dimensions apply to all types of human gestures, their usage and interpretation can vary across different cultures. For instance, forming a circle with your thumb and index finger might symbolize "ok" in Europe and North America but is considered an insult in Russia, Brazil, and Turkey [Pease and Pease, 2008]. Apart from intentionally performed movements, some gestures can also occur unconsciously in specific situations. This especially applies to gestures from Ekman and Friesen's [1969b] *adaptors* category (e.g., nail-biting, leg shaking, or fidgeting with objects), which can reveal insights about a person's true attitude and feelings. Since these signals are produced unconsciously, their analysis can provide more trustworthy and reliable conclusions than other movements [Pentland, 2008].

In contrast to gestures, *postures* refer to the overall alignment and orientation of the human body. They have a more static nature and are less frequently used for intentional communication. However, their analysis can still provide valuable insights regarding people's current affective state and social status [Poyatos, 1984]. For example, open postures are often associated with confidence and the willingness to cooperate, whereas closed postures usually imply the opposite [Pease and Pease, 2008]. To further classify their characteristics and social implications, Scheflen [1964] proposed three easily observable dimensions. The first one distinguishes between *inclusive* and *non-inclusive* behaviors and considers to which extent postures involve or exclude others. For instance, body positions with open arms typically indicate inclusiveness, while crossed

arms or turned backs can express defensiveness and the desire to create social distance. The second dimension identifies whether activities are performed *face-to-face* or with a *parallel body orientation*. The rationale behind this category is that individuals tend to be more active and engaged in interactions when facing each other than in parallel orientations. Finally, the third dimension is concerned with *congruence vs. incongruence*. It determines whether people adopt the same posture as their communication partners during interactions. This imitation behavior is also called *mirroring* and often indicates a deeper connection between individuals [Chartrand and Bargh, 1999].

Overall, the categories and dimensions proposed by McNeill [1992] and Scheflen [1964] provide valuable insights regarding the interpretation of gestures and postures. By automatically recognizing and responding to the meaning of these movements and body positions, assistive systems can adapt their experiences and support individuals in real-time. For that, several methods and techniques can be used to achieve the desired outcomes. While markers and sensors placed at a person's limbs typically produce the most accurate tracking results, they are also relatively intrusive and reduce the naturalness of interactions, which makes them only suitable for specific situations [Ibraheem and Khan, 2012]. In contrast, vision-based approaches provide more flexibility and enable the continuous analysis of gesture and posture data. To this end, various approaches, such as the Gesture Recognition Toolkit (GRT) [Gillian and Paradiso, 2014], BlazePose [Bazarevsky et al., 2020], and OpenPose [Cao et al., 2021] can be utilized.

3.2 Physiological Cues

In order to understand the cognitive processes that occur inside human beings, other sources of information beyond observing their actions and behaviors can provide additional insights. This is where physiological cues, also known as biosignals, come into play. They are measurable parameters that originate within the human body and offer a window into the underlying biological processes [Kaniusas, 2012a]. By analyzing the temporal progression of these signals, they enable conclusions about people's health, emotional state, and cognitive abilities. Example parameters include brain activity, heart rate, respiration, and skin conductance. In contrast to behavioral cues, which can be consciously controlled, most physiological reactions are regulated by the autonomic nervous system and can not be intentionally manipulated by untrained individuals [Jerritta et al., 2011]. Consequently, these cues provide a valuable, reliable, and trustworthy source of information that can complement or even contradict externally observable behaviors in case of intentional deceptions.

Since most of these signals are always present in living human beings, they can be continuously captured and analyzed over extended periods of time. However, current sensing technology still requires direct contact with the human body to acquire most physiological signals, which increases the intrusiveness of approaches and creates an additional entry barrier when attempting to use them for assistive systems. In return, the directly captured signals are typically more accurate and reliable than those acquired with less intrusive methods. While our bodies possess numerous physiological parameters [Kaniusas, 2012b], not all of them can be utilized to draw conclusions about people's cognitive processes and mental states (e.g., stomach volume or kidney filtration rate). For this reason, the following sections only provide an overview of the most commonly used physiological cues.

3.2.1 Heart Activity

The heart's primary function is to circulate blood, oxygen, and nutrients within the body's vascular system through rhythmic contractions. This continuous process is essential for maintaining homeostasis and is achieved by dynamically responding to various physiological demands and mental states [Marieb and Hoehn, 2019]. Consequently, analyzing the nature and frequency of heartbeats can provide valuable insights into a person's cardiovascular health, autonomic nervous system function, cognitive processes, and overall physiological condition [Shaffer et al., 2014]. One of the most common and accurate methods to measure heart activity is electrocardiography (ECG). It involves placing multiple electrodes on a person's skin and capturing the electrical signals generated by the cardiac muscle. The resulting signal consists of several distinct components, including the P wave (atrial depolarization), QRS complex (ventricular depolarization), and T wave (ventricular repolarization), as shown in Figure 3.3. Each component represents a specific phase of cardiac activity, originating from electrical stimuli (emitted by the sinoatrial node) that travel through the heart and trigger the contraction or relaxation of the upper (atria) and lower chambers (ventricles).

Based on the collected data, the heart rate (HR) can be calculated by measuring the duration between two consecutive R-spikes, also known as RR-interval or inter-beat interval (IBI), and extrapolating how many of these periods occur within a minute. Combined with heart rate variability (HRV), which refers to the fluctuations of time intervals between sequential heartbeats, these measures can be used to draw conclusions about a person's physical and emotional condition [Shaffer and Ginsberg, 2017]. For instance, higher HRV is typically associated with a healthy autonomic nervous system function,

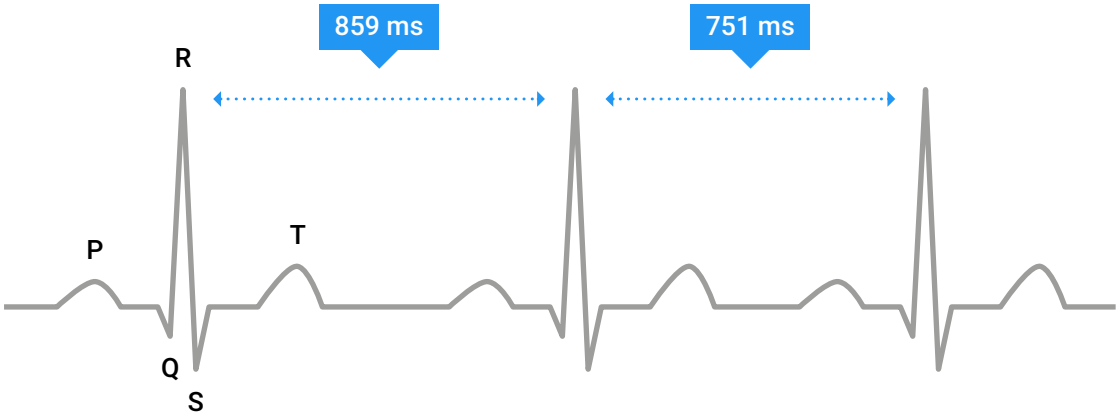


Figure 3.3: ECG signal components and example RR-intervals.

while lower HRV can indicate stress, heightened arousal, or potential cardiovascular anomalies [Quintana et al., 2012]. An alternative method to measure these parameters is photoplethysmography (PPG). It utilizes the light absorption characteristics of blood to detect volumetric changes within the blood vessels after each heartbeat (also known as blood volume pulse) [Sinex, 1999]. Since this process only requires illuminating the skin with an LED and measuring the amount of reflected or absorbed light, PPG sensors are typically less intrusive than ECG. On the other hand, the resulting signal is less accurate and can be subject to movement artifacts, which reduces its reliability and makes the data more challenging to process [Weiler et al., 2017].

3.2.2 Electrodermal Activity

Electrodermal activity (EDA), also known as galvanic skin response (GSR), refers to the constantly changing electrical properties of the human skin. It is based on the state of around three million sweat glands that are distributed in varying densities across the human body and produce sweat in response to signals from the autonomic nervous system [Boucsein, 2012, pp. 2–14]. The secreted fluids temporarily increase the electrical conductivity of the skin, which can be measured by applying a small electrical current to the surface (*exosomatic*) or recording the potential differences originating from the skin itself (*endosomatic*) [Dawson et al., 2007]. In this regard, Vigouroux [1879] and Féré [1888] were among the first who used these sensing techniques to discover a close relation between people's skin resistance level and their psychological state. Several studies confirmed these findings and showed that EDA is linked to both physical and mental arousal (e.g., excitement, stress, or anxiety) [Neumann and Blanton, 1970].

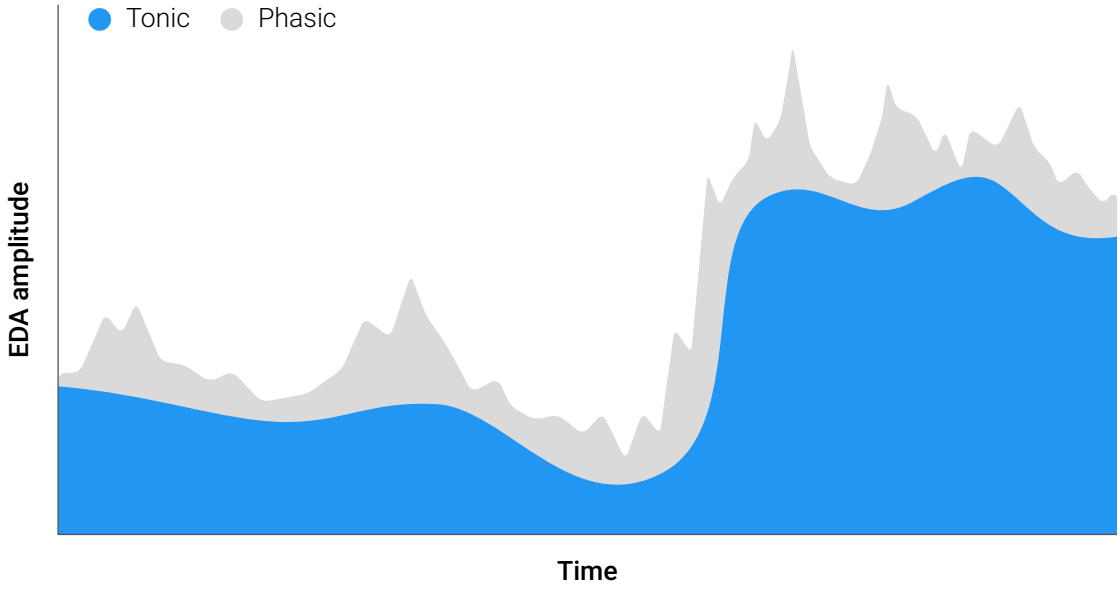


Figure 3.4: EDA signal with tonic and phasic components.

Additionally, early research also implied the existence of two distinct components within the captured electrodermal activity (see Figure 3.4). While the *tonic* level of skin conductance or resistance refers to the slowly changing background characteristics of the signal, the superimposed *phasic* component relates to rapidly changing skin conductance or resistance responses [Dawson et al., 2007]. These reactions typically occur within one to four seconds after perceiving an external stimulus and provide valuable insights regarding short-term emotional responses [Boucsein, 2012, pp. 151ff., 369ff.]. A common method for automatically analyzing and reacting to relevant changes in the phasic component is to remove the tonic baseline and calculate statistical features for machine-learning-based approaches (e.g., Picard et al. [2001] or Wagner et al. [2005]). Based on the classification results, assistive systems can adapt their interactions to a user's current state and provide appropriate support to deal with negative emotions.

3.2.3 Brain Activity

Brain activity refers to the complex interactions between billions of neurons that communicate with each other through electrical signals and chemical substances called neurotransmitters [Frackowiak et al., 2004]. This continuous exchange of cues and information serves as the foundation for all human functions, including cognitive processes, emotional states, and bodily behaviors. Consequently, analyzing the dynamic interplay of signals within our brains is essential for understanding and supporting the underlying mechanisms and procedures. The primary method to measure brain activity is elec-

troencephalography (EEG). It involves placing electrodes directly on a person's scalp to detect voltage fluctuations generated by neuronal activations [Biasiucci et al., 2019]. The captured signals can be categorized into five primary frequency bands called alpha (8–12 Hz), beta (12–35 Hz), gamma (> 35 Hz), delta (0.5–4 Hz), and theta (4–8 Hz) waves [Abhang et al., 2016]. Each of these bands is associated with specific states of consciousness and cognitive processes. For instance, alpha waves are prominent during relaxed and attentive states, while beta waves are linked to problem-solving and active thought. In contrast, observing irregular patterns in these waves can indicate neurological disorders or cognitive dysfunctions. To detect such conditions, clinical devices typically require stationary settings along with intrusive electrode caps for precise measurements. Although recent advancements in mobile EEG technology have enabled greater flexibility and ease of use through gel-free and wireless devices, their signal quality still represents an ongoing challenge that needs to be addressed before they can serve as viable alternative solutions [Radüntz, 2018].

3.2.4 Muscle Activity

Muscle activity is a fundamental physiological process that reflects the coordinated effort of muscle fibers to contract and produce movements [Sherwood, 2015]. It originates from the nervous system, where motor neurons transmit electrical impulses that trigger a biochemical reaction within the respective muscle cells, resulting in their contraction. The primary method to analyze muscle activity is electromyography (EMG). Similar to EEG and ECG, electrodes are placed on the skin above the targeted region to capture the electrical potential of muscle cells generated during activation [Partridge and Partridge, 2003]. The degree of tension depends on the number of stimulated muscle fibers and is directly reflected in the signal's amplitude. Therefore, capturing and analyzing the resulting EMG data can provide valuable insights about a person's neuromuscular function, coordination, fatigue, and overall health. For instance, one area of application involves monitoring gestures and movement patterns to support physiotherapeutic rehabilitation and improve athletic performance [Hogrel, 2005]. Additionally, electromyography can be used to analyze emotional expressions, such as happiness, anger, or fear, by placing electrodes around the facial region [Fridlund et al., 1984]. While this method is more intrusive than camera-based solutions, it is able to detect even the slightest movements and microexpressions.

3.3 Summary

This chapter provided an overview of available non-verbal signals that can be captured and analyzed with sensing devices to gain insights about the cognitive processes introduced in Chapter 2. It initially defined the term "signals" more precisely and distinguished between different types of stimuli. The first category are behavioral cues, which refer to externally observable actions and behaviors performed by a person to communicate non-verbal messages. They often complement spoken language and can be expressed through multiple modalities to indicate specific social signals. Examples include facial expressions, gaze behavior, vocal cues, gestures, and posture. Since these signals can be captured relatively easily from a distance with external cameras and microphones, no intrusive sensors are required that could influence people's behavior. In contrast, physiological cues typically occur without the intention of conveying information to other individuals. They originate within the human body and offer a window into the underlying biological processes (e.g., brain activity, heart rate, or skin conductance). Since most of these parameters are regulated by the autonomic nervous system and can not be intentionally manipulated by untrained individuals, they represent a relatively reliable source of information. However, capturing them typically also requires more intrusive sensors, which could reduce people's acceptance of potential augmentation solutions. Consequently, suitable compromises must be found according to the targeted circumstances and accompanying requirements. Further details and guidelines regarding the selection of appropriate sensing devices are provided in Section 5.2.

II Concept & Implementation

4 Assistive Augmentation 59

4.1 Theoretical Context
4.2 Literature Analysis
4.3 Strategies
4.4 Design Dimensions
4.5 Summary

5 Mobile Signal Processing 107

5.1 Challenges
5.2 Hardware Selection
5.3 Data Collection
5.4 Model Training
5.5 Summary

6 The SSJ Framework 127

6.1 Origins and Basic Concepts
6.2 Existing Solutions
6.3 Architecture
6.4 Graphical Interface
6.5 Example Application
6.6 Summary

Chapter 4

Assistive Augmentation

Suffering from impairments or disorders of cognitive processes can have severe effects on a person's daily life. These consequences include a variety of limitations ranging from not being able to perform specific actions, such as participating in social activities or exercising a profession, to the complete loss of personal independence [Scherer et al., 2005]. In order to reduce the burden of these conditions and to support affected individuals with appropriate solutions, this chapter focuses on conceptual approaches to augment the respective cognitive processes (Chapter 2). For that, we follow the core principles of the *assistive augmentation* paradigm. According to Huber et al. [2018, p. 2] assistive augmentation technology *"should be socially acceptable, work coherently for disabled and non-disabled alike, and support independent and portable interaction"*. While some researchers associate the term *augmentation* only with enhancements beyond natural human capabilities [Kiss, 2020], this distinction does not apply to the assistive augmentation paradigm. Rather than solely focusing on amplifying the abilities of the average population beyond natural limitations, it considers

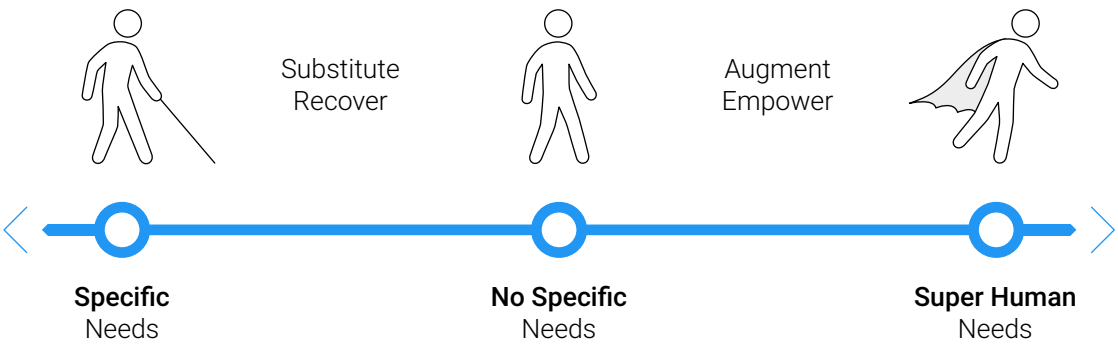


Figure 4.1: Assistive augmentation continuum. Adapted from Huber et al. [2018, p. 2].

the personal needs and circumstances for the development of technology that benefits individuals regardless of their capabilities on the augmentation spectrum displayed in Figure 4.1. As a research field, assistive augmentation encompasses both the recovery and amplification of sensory, memory, and higher cognitive capabilities. Depending on the targeted area of the augmentation spectrum, research is spread across multiple disciplines, including human-computer interaction, assistive as well as accessibility technology, and human augmentation [Huber et al., 2018]. Since the overall incentive of this thesis is to support people affected by impairments and disorders of cognitive processes, we will primarily focus on the left side of the continuum.

Another important aspect of assistive augmentation is the commitment towards independent and portable interaction. This requirement ensures that the conditions and circumstances which can occur during the usage of potential solutions in the wild are considered early on in the design and development process. Otherwise, intended behavior and interactions only validated in controlled environments might not translate to the real world and could cause unwanted or even harmful user experiences. More importantly, focusing on portable approaches enables assistive augmentation technology to be available throughout a person's daily life. This opens up new opportunities to support affected individuals in ways that are not possible with stationary or desktop applications. For example, the ubiquitous presence of augmentation systems allows people to utilize their assistance at any place and time. Additionally, it establishes the foundation for potential solutions which can automatically detect when the user is in a critical situation and proactively offer support at the appropriate time.

The concepts behind all of these approaches will be discussed in the following sections. More precisely, we start by placing the assistive augmentation paradigm within the larger theoretical context of human augmentation to provide a better understanding of related terms and concepts. Following that, we conduct a literature analysis to identify suitable augmentation strategies for each respective cognitive process and review related works regarding their applied methods and procedures. Based on the previous findings, we then identify common design dimensions that can be used to guide the design process of future assistive augmentation technologies and provide examples for each direction to demonstrate potential implications of the respective choices.

4.1 Theoretical Context

The fundamental idea of using computational technology to support and extend human capabilities is not a new phenomenon. Its roots can be traced back to the early 1960s, when Joseph Licklider anticipated the development of future systems with close symbiotic relationships between humans and computers. In his work *"Man-Computer Symbiosis"*, he foresaw a period of flexible collaborations where humans and computers work together to make joint decisions and solve complex problems [Licklider, 1960].

Shortly thereafter, Douglas Engelbart expanded upon this idea and explored how computers could be used to augment the human intellect. His vision was to amplify the natural human capabilities to deal with complex situations through improved comprehension aided by interactive applications [Engelbart, 1962]. To achieve that, he proposed a conceptual framework where humans are part of a larger system (which he named H-LAM/T) that consists of the *language*, *artifacts*, and *methodology* they are *trained* in. He argued that improving the system's overall performance can be accomplished by augmenting any individual part of it. This includes assigning meaningful terms to useful concepts for easier reference, using artifacts such as computers and information displays to enhance comprehension of complex problems, applying more effective procedures to complete specific tasks, and improving training techniques to acquire new skills and abilities [Xia and Maes, 2013].

While technology has advanced significantly since then, the underlying concepts are still valid today. The most notable difference is the availability of computational resources. Instead of having to share a mainframe computer between multiple people like in the past, the majority of the global population nowadays has access to a manifold of computational power at the tip of their fingers through smartphones and other wearable devices [Ericcson, 2024; GSMA, 2023]. The emergence of these technologies plays an essential part in the process of augmenting human capabilities the way it was envisioned by Engelbart and Licklider. It facilitates the development of personalized applications which seamlessly integrate with the physical world and provide immediate access to digital information. Moreover, it enables users to interact with smart objects and benefit from the provided augmentation at any place and time without the need for intrusive extensions or modifications of their bodies.

Although the core concepts behind *human augmentation* were formulated several decades ago, a general definition of the term has only recently been proposed. According to Raisamo et al. [2019], human augmentation is *"an interdisciplinary field that*

addresses methods, technologies and their applications for enhancing sensing, action and/or cognitive abilities of a human. This is achieved through sensing and actuation technologies, fusion and fission of information, and artificial intelligence (AI) methods". While it remains to be seen whether this definition will prevail within the research community, for now, it provides a solid foundation to address the general scope of the field, which in turn enables a clear distinction from related terms and concepts.

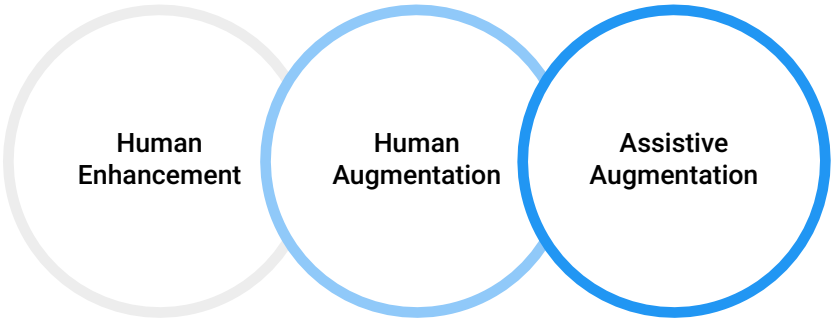


Figure 4.2: Relation of Human augmentation and its associated terms. Based on de Boeck and Vaes [2021].

One of these terms is *human enhancement*, which describes a relatively broad field spanning several disciplines, including mechanical, electrical, chemical, and genetic engineering. It refers to any pharmaceutical (e.g., medication or chemical stimulants [Robbins, 2005]), biomedical (e.g., surgical operations, transplants, or implants [Suthana et al., 2012]), or genetic (e.g., genome editing of embryos [de Araujo, 2017]) modification aimed at improving human abilities, capacities, and performances beyond the scope of restoring and sustaining health [Giubilini and Sanyal, 2015; Juengst and Moseley, 2019]. Due to the rather permanent and invasive nature of these enhancements, several ethical concerns were raised in the past that are still part of an ongoing discussion today [Savulescu and Bostrom, 2009]. Examples include the potential impact on the liberty of future generations who have to accept the genetic choices of their ancestors [Habermas, 2003] and the further exacerbation of inequalities within the population due to the possibly limited availability of enhancement technologies only to the wealthiest people [Mehlman and Botkin, 1998]. In contrast, human augmentation primarily focuses on temporary and device-based technologies to achieve the intended improvements through human-machine interactions [de Boeck and Vaes, 2021]. For example, an exoskeleton is considered human augmentation, while biomechanical implants fall into the category of human enhancement. Consequently, human augmentation can be viewed as a subset of human enhancement, due to its more confined scope (see Figure 4.2). Similarly, assistive augmentation can be regarded as part of human augmenta-

tion, focusing on individual limitations and how to overcome them with augmentation technologies. This includes recovering abilities lost due to impairments and disabilities as well as improving the performance and capacities of sensory, memory, and higher cognitive capabilities.

Overall, assistive augmentation can be divided into three major categories: *physical*, *social*, and *mental augmentation* (see Figure 4.3). Physical augmentation focuses on the interactions with the physical world and its objects surrounding us. This encompasses extended motor functions, amplified strength, speed, dexterity, and endurance, as well as remote presence and teleoperation [Development, Concepts and Doctrine Centre, 2021]. One of the most prominent examples of it are the previously mentioned exoskeletons, which can be worn to support certain physical activities such as lifting heavy objects or performing rehabilitation exercises [Chen et al., 2019].

Meanwhile, social augmentation primarily relates to the enhancement of our interactions with other individuals. This includes non-verbal behavior, communication, and collaboration with other humans and computers [de Boeck and Vaes, 2021]. For example, Damian et al. [2015] proposed a system that analyzes the user's behavior during public speaking and provides feedback on a head-mounted display (HMD) regarding the speaker's openness, body energy and speech rate. The information gained from these indicators enables the users to become aware of their social behavior and lets them learn how to adjust it in order to improve their communication skills.

Finally, mental augmentation is concerned with the amplification of all cognitive processes introduced in Chapter 2. While this involves various (sub-)processes such as attention, long-term memory, decision-making, and problem-solving, they can be grouped into the three major processes *perception*, *memory storage*, and *higher-order cognition*. Since each of them has its own set of challenges and respective augmentation strategies, mental augmentation can be further divided into the three corresponding areas *sensory*, *memory*, and *cognitive augmentation* as shown in Table 4.1 [Schmidt, 2017]. Following conceptual models from the field of cognitive psychology [Eysenck and Keane, 2020], this distinction allows for a more in-depth exploration of the underlying challenges and their potential solutions, which will be the focus throughout the rest of this thesis.

Firstly, sensory augmentation applies methods and technologies to enhance the processes of acquiring information from sensory organs and interpreting the perceived signals [Raisamo et al., 2019]. As sensor technology has surpassed human capabilities, it can be employed to amplify existing senses beyond their natural limits. For instance, microphones have a higher temporal resolution than the human ear, enabling the record-

Cognitive Psychology	Assistive Augmentation	Description
Perception	Sensory Augmentation	Refer to perceptual processes
Memory Storage	Memory Augmentation	Refer to memory-related processes
Higher-Order Cognition	Cognitive Augmentation	Refer to higher cognitive processes
Cognition	Mental Augmentation	Umbrella terms that refer to all cognitive processes
Cognitive Processes	Cognitive Processes	

Table 4.1: Mapping of terms between different research fields.

ing of frequencies outside the audible spectrum (20 Hz – 20 kHz). After conversion into a signal perceivable by humans, it allows us to capture and react to sounds that we would otherwise not be able to hear (e.g., ultrasonic sound waves). In addition to exceeding the limits of natural human perception, sensory augmentation can also be used to supplement impaired senses. For that, the signal is either amplified significantly to overcome the impairment or converted into another sensory modality to bypass the limitations entirely [Huber et al., 2018]. An example of this was proposed by Olwal et al. [2020], who visualized real-time transcriptions of spoken language on a head-mounted display for people who are deaf or hard of hearing. It enabled them to understand and participate in conversations with others by converting acoustic to visual information.

However, perceiving information is usually only the first part of the cognitive processing we perform in our daily lives. Encoding, retaining, and recalling it are equally important functions of the human memory that can lead to forgetting if a failure occurs in any one of them [Dingler et al., 2021]. Previous research indicates that a lack of attention during an event results in significantly reduced recall performance [Craik et al., 1996]. Similarly, retrievability is also influenced by the time spent to process a stimulus after its perception [Craik and Lockhart, 1972]. As a result, many ordinary instances of forgetting can be attributed to the ineffective integration of new information with existing memories [Harvey et al., 2016]. These cases are where memory augmentation technology shows its greatest potential. Due to its automatic nature, it can act as surrogate memory that stores information reliably and is not prone to distractions. It can help users compensate everyday memory failures by providing the desired information at the right time. For that, methods such as lifelogging can be applied, which continuously capture a person's experiences from various data sources (e.g., images, video, audio, GPS position, or physiological sensors). One example that makes use of this approach is SenseCam [Hodges et al., 2006], a small camera device worn around a per-

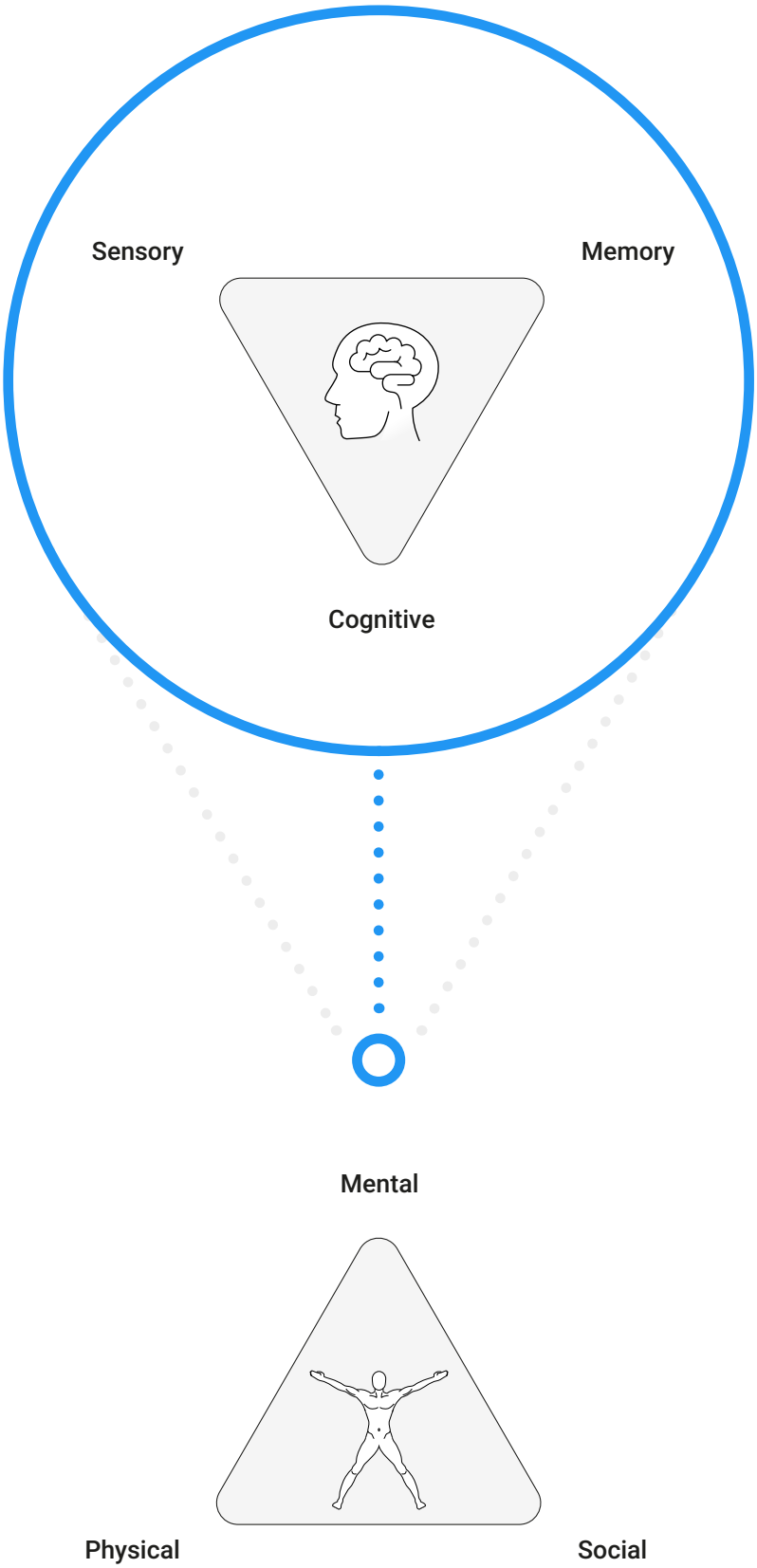


Figure 4.3: Assistive augmentation classification.

son's neck that automatically takes photographs based on changes in certain conditions (e.g., brightness, temperature, or elapsed time). Reviewing the captured information can help users recollect forgotten details of past events and can positively impact long-term retention performance [Roediger and Karpicke, 2006].

In addition to compensating ordinary instances of forgetting, memory augmentation can also be used to support people affected by memory impairments and disorders. Although the consequences of these conditions are much more severe, similar methods can be applied to alleviate their impact. Personalized hints, reminders, manuals, and navigation instructions are only a few of the commonly used approaches to assist affected individuals. For example, Hamilton et al. [2021] developed an augmented reality application for people with dementia. Their caregiver can customize it to display personalized reminders through text, images, videos, 3D models, voice messages, or music. Furthermore, the system can detect the names of objects within the field of view and supports indoor navigation in case people forget the path to their destination. While this is just one example, it shows the potential of using memory augmentation technology to assist individuals and enhance their personal independence.

In order to further support the intellectual abilities of humans, approaches from the field of cognitive augmentation (also known as augmented cognition) can be applied. Since the targeted conditions are usually rather specific to each individual, cognitive augmentation focuses on methods and technologies to determine a person's current cognitive state and uses the acquired information to adapt all involved systems to the identified needs and requirements [Schmorrow and Kruse, 2004]. This symbiotic coupling between humans and computers, as Licklider [1960] envisioned it several decades ago, is achieved through non-invasive physiological and behavioral sensing [Schmorrow et al., 2006]. Based on the continuously measured parameters, corresponding cognitive states can be inferred automatically and in real-time. Once the targeted state has been detected, appropriate adaptation strategies can be applied to mitigate potential information-processing bottlenecks [Stanney et al., 2009b]. These include limitations in language, learning, comprehension, and decision-making among others [Reeves and Schmorrow, 2007].

In addition to enhancing human performance in these processes, cognitive augmentation can also be used to rehabilitate and support cognitive impairments and disorders [Stanney et al., 2009a]. While the general concepts for that are relatively similar, there are some distinct differences between these two directions. Specifically, the thresholds for certain states and the targeted conditions themselves differ from those focused

on performance optimization. Furthermore, the mitigation and adaptation strategies to support impaired cognitive functions rely on different aspects to provide appropriate assistance in corresponding situations. For instance, Stanney et al. [2009a] proposed several strategies to compensate the specific effects of traumatic brain injuries. These approaches include self-monitoring of behavioral impulses (e.g., multimodal feedback), frustration reduction (e.g., task simplification, calming environment), and motivational enhancement techniques (e.g., goal visualization), which are typically not applied to amplify cognitive capabilities. However, in the targeted situations, they allow individuals to perform affected functions despite impairments and support rehabilitation through restorative exercises and activities.

4.2 Literature Analysis

After establishing the theoretical foundation, it is equally important to identify common strategies and concepts of previous works that can inform the design and development of a generalized solution for the assistive augmenting of cognitive processes, which is one of the main objectives of this thesis. To this end, we conducted a systematic literature analysis of related research and applications within the context of human-computer interaction (HCI). The primary goal was to identify similarities and shared properties of previously proposed systems that might be useful to consider when implementing future approaches. The details of this analysis are described in the following sections.

4.2.1 Methodology

For our analysis, we followed the general structure proposed in the updated *Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA 2020)* statement [Page et al., 2021]. It includes guidelines designed to aid the transparent reporting of systematic literature reviews and is based on the following stages: *identification*, *screening*, and *inclusion*. Before conducting each of the three phases, the first step was to define the scope and search criteria of the analysis. We decided to focus primarily on publications describing sensory, memory, and cognitive augmentation approaches that follow the principles of assistive augmentation. This selection excludes all applications that do not target the respective cognitive processes and do not restore or enhance the capabilities of individuals in these areas. Furthermore, we excluded theoretical approaches that have yet to be implemented and evaluated since they might be based on assumptions that do not apply during real-world usage.

Figure 4.4: Flow diagram of the literature analysis based on the PRISMA 2020 stages.

With these criteria in mind, we first performed the identification phase of the PRISMA methodology, during which publication sources such as central databases and records are selected as the basis for all further steps. To this end, we compiled the following list of HCI conferences and journals with a high probability of containing publications focusing on the augmentation of cognitive processes: Augmented Human International Conference (AH), Augmented Humans International Conference (AHs), Conference on Human Factors in Computing Systems (CHI), International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), International Symposium on Wearable Computers (ISWC), and Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT). In order to identify relevant publications from the selected records, a common method is to apply a targeted keyword-based search. However, since the field

of human augmentation and, more specifically, the assistance of cognitive processes is still relatively young, a consistent and widely adopted terminology has not yet been established that could be used for a comprehensive search. Additionally, previous works that meet all inclusion criteria but were not explicitly positioned as augmentation approaches would not be identified with such a method. For these reasons, we decided to include the publications of all proceedings and volumes of each previously listed venue released before conducting this analysis (in early 2023), resulting in an initial set of 15,358 publications (Figure 4.4).

Based on this corpus, we conducted the screening phase, which is used to filter out publications that are not relevant or related to the topic of the analysis. In our case, the screening process was achieved with the following steps: First, we removed all publications with a title clearly referring to a different subject, which resulted in 772 remaining records. We then screened them again and excluded 590 entries based on the contents of their abstracts. In each step, unclear or borderline cases were included for a more informed decision at the next stage.

The remaining 182 publications served as candidates in the inclusion stage, where the final decision, whether to include a specific paper or not, was made. After examining all records, we rated the contents of each publication regarding its relevance on a 3-point Likert scale (*low*, *medium*, and *high*). Papers with a "low" rating were excluded and records with a "medium" rating were reviewed again for consideration. In the end, only papers with a "high" rating were included, resulting in a total of 99 publications. Based on this selection of papers, we conducted a comparative analysis to identify similarities and shared characteristics within the body of related works.

4.2.2 Findings

Although our analysis only covered a subset of all existing assistive augmentation approaches, it yielded a sufficient amount of records to gather insights about commonly used structures and strategies of proposed systems. This might not have been the case with a different methodology (e.g., a keyword-based search) since most identified approaches did not use a consistent and widely adopted terminology, which can be attributed to the relatively young age of the research area. As illustrated in Figure 4.5, the exploration of attempts to augment cognitive processes sporadically emerged in the early 2000s and only started to gain traction around 2011. Since then, the number of publications has significantly increased, indicating the research community's growing interest in this topic. One factor that might have contributed to the sudden rise in pop-

ularity was the technological advancement of mobile and wearable devices during this time, which enabled many new augmentation techniques that were previously impossible. In our analysis, more than 63.6% of approaches used these methods to focus on sensory augmentation. The remaining publications were split almost evenly between memory (17.2%) and cognitive assistance (19.2%).

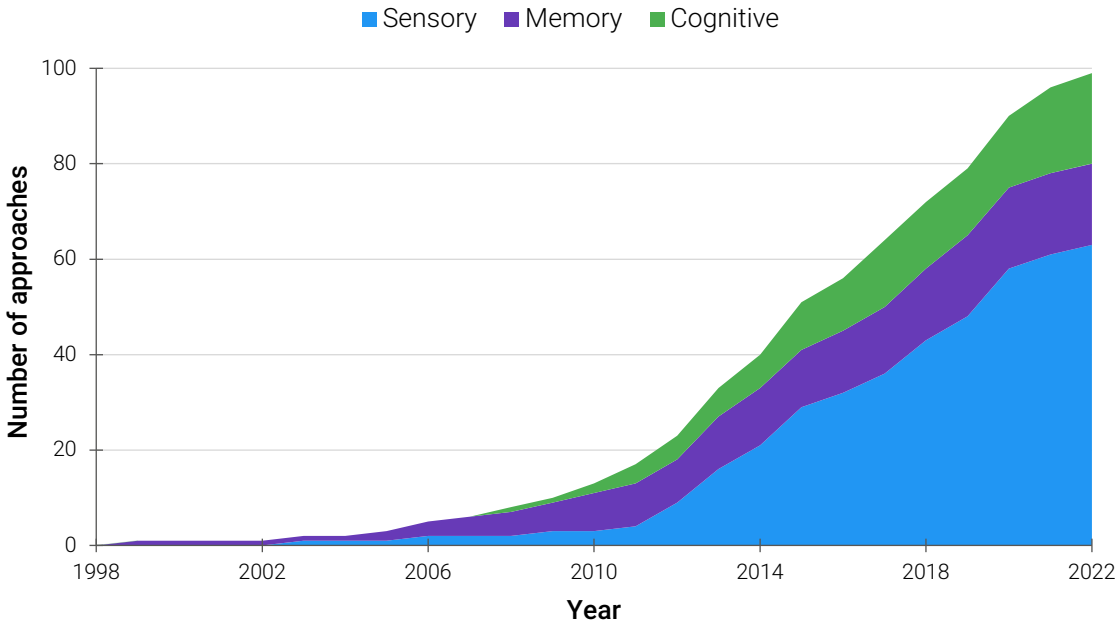


Figure 4.5: Accumulative timeline of publications identified in our literature analysis across augmentation areas.

One of the most prominent similarities across the analyzed augmentation areas and publications was the technical structure of the proposed approaches. Almost every system included all or at least a subset of the components shown in Figure 4.6. As illustrated, the first step usually consisted of a sensing component to analyze the users, their environment, or both depending on the augmentation target. Examples range from head-mounted cameras and microphones used to capture memorable events for retrospective assistance to wearable physiological sensors that provide information about a user's current activity. In this regard, the same types of sensors were often used in many different ways. For instance, Sohn et al. [2005] utilized the GPS position to provide location-based reminders, while Yatani et al. [2012] employed the same data to guide visually impaired users to points of interest. These two examples also illustrate the different interaction flows that can be achieved based on the selected trigger type. While some approaches used the sensor data directly as input for their augmentation systems, others continuously analyzed the signals to identify specific thresholds and suitable conditions to trigger the assistance. Regardless of how a system is initiated, the sensing component

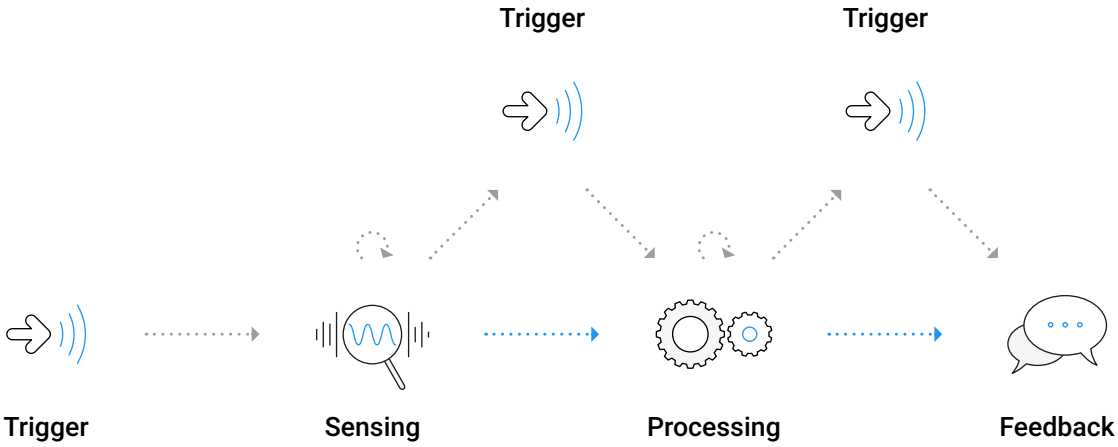


Figure 4.6: Common system structure of analyzed augmentation approaches.

always plays an essential role since it serves as the foundation for providing assistance and understanding a user's current state.

Following the initial sensing phase, the acquired signals were typically processed to extract certain details, identify specific conditions, or enhance the individual perception of sensory information. This step commonly involved the use of algorithms and machine learning techniques to enhance, transform, and interpret the raw data collected by the sensors. For example, Jain et al. [2022] trained a neural network model that used the raw audio data from a microphone as input to recognize the contained sound categories for deaf and hard of hearing users. One important aspect to consider in this phase is the processing time of acquired signals. While short delays might still be acceptable, longer latencies can lead to negative experiences if a system can not adapt to changing user inputs and environmental conditions in real-time. For this reason, most analyzed approaches either optimized their processing steps to run efficiently on wearable and mobile devices or performed resource-intensive operations on more powerful external servers and only returned the results back to the client devices.

The last step commonly involved generating and providing user feedback as part of the augmentation. Based on the processed signals from the previous phase, relevant information was collected and delivered through one or often multiple modalities, including visual, auditory, and tactile representations. To achieve that, the output capabilities of various mobile devices such as head-mounted displays, smartphones, or wearables were used. For instance, Goodman et al. [2020] utilized a smartwatch's display and built-in vibration motors to convey the relative direction and loudness of sounds. Besides multiple modalities, this example also illustrates the usage of continuous feedback. Thereby, the system constantly provides information based on the processed sensor data until it is

turned off or disabled by the user. As an alternative, other approaches employed trigger- or event-based feedback, such as prompts and notifications. This technique is primarily used when the provided information is only relevant in specific situations or conditions that can be automatically detected.

Overall, the shared technical structure across all identified systems can be used as a foundation for future assistive augmentation approaches. It provides a general framework for the requirements and responsibilities of involved components and can aid the design process of more generalized solutions. A first step towards this goal could be a flexible system with various reusable and easily exchangeable modules that can be combined to fulfill the roles of each component within the identified model. However, besides understanding the general structure of past approaches, it is equally important to consider the previously applied strategies when designing new augmentation concepts.

4.3 Strategies

In addition to the general findings presented in the previous section, our analysis of related works also revealed that each assistive augmentation area (see Figure 4.3) has its own set of strategies that should be considered when designing new assistive systems and technologies in the respective field. To support the process of deciding which method to use in a specific circumstance, we provide an overview of the most commonly applied sensory, memory, and cognitive augmentation strategies found in our review.

4.3.1 Sensory Augmentation

Overall, the identified approaches applied five general strategies for sensory augmentation: (1) amplifying certain aspects of a sensory modality that are naturally perceivable by humans, (2) mapping naturally perceivable information from one sensory domain to another, (3) extending the perceivable range of stimuli within a specific sensory domain, (4) enhancing the perception of stimuli beyond natural limits across sensing domains, and (5) adding artificial sensory modalities. Figure 4.7 illustrates the general approach of each method and outlines the differences between each strategy.

Depending on whether the goal is to restore sensing capabilities or to enhance them beyond natural boundaries, different approaches are suitable in each case. While *amplification* and *substitution* are often used for restorative purposes, *extension* and *enhancement* are primarily applied to expand the natural range of perceivable stimuli. Although creating additional sensing modalities is currently only a theoretical strategy,

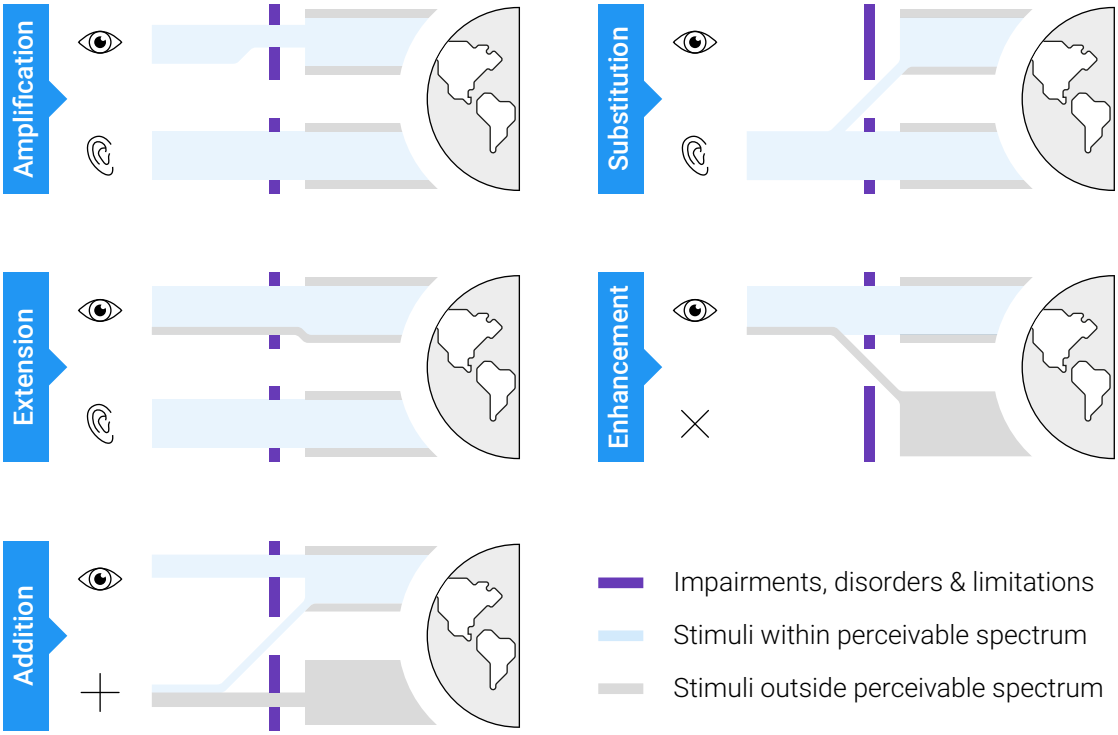


Figure 4.7: Overview of different sensory augmentation strategies within the context of assistive augmentation.

we included it in this overview for the sake of completeness. As a result, the identified strategies are mostly in line with the sensory augmentation types found by Kiss [2020]. The main differences are that we included sensory substitution as it is commonly applied within the context of assistive augmentation and omitted the *annotation* type since it follows the same approach as sensory enhancement. The details of each strategy are explained in the following sections.

Amplification

Human perception allows us to focus our attention on specific aspects of the environment in a way that makes us more aware of them than those outside our attention span (see Section 2.1.2). Similarly, sensory amplification attempts to artificially increase the magnitude (or strength) of certain stimuli within the naturally perceivable spectrum of the respective sensory organ to improve their perception. Examples include increasing the volume of faint sounds through sensitive microphones and speakers or magnifying small visual details (e.g., texts on distant signs) with a high-resolution camera and a head-mounted display (see Table 4.2). The primary intention is to overcome sensory limitations by amplifying naturally perceivable stimuli so that more information can be

captured with the respective sensory organs. In addition to providing more information, this strategy can also be applied to circumvent sensory impairments caused by accidents, disabilities, or age-related circumstances. There the stimulus gets amplified until it is perceivable by the unaffected receptors of the sensory organ.

			Sensors	
Publication	Condition	Amplification		
Aoki et al. [2003]	regular hearing	adjust conversation audio	○	●
Fardoun et al. [2013]	blind spots	transform field of view	●	○
Tanuwidjaja et al. [2014]	color blindness	map colors within the field of view	●	○
Flatla et al. [2015]	color blindness	overlay color information	●	○
Itoh and Klinker [2015]	optical defocus	overlay compensation image	●	○
Zhao et al. [2016]	low vision	provide attention cues during search	●	○
Langlotz et al. [2018]	color blindness	map colors within the field of view	●	○
Amini et al. [2020]	blind spots	transform field of view	●	○
Knierim et al. [2020]	regular vision	enable slow motion vision	●	●
Eghtebas et al. [2021]	regular vision	enlarge areas within the field of view	●	○
Min Htike et al. [2021]	low vision	enhance contrast & contours	●	○

Legend: Camera Microphone

Table 4.2: Sensory amplification approaches.

For instance, one popular application found in our literature analysis was amplifying color information to make it perceivable for people with color vision impairments (Tanuwidjaja et al. [2014]; Flatla et al. [2015]; Langlotz et al. [2018]). In order to achieve this, all identified approaches used a visual overlay to convey the missing information about the underlying objects. While Flatla et al. [2015] focused on patterns and words to indicate the corresponding color, Tanuwidjaja et al. [2014] and Langlotz et al. [2018] applied different processing techniques to directly alter the colors and make them distinguishable on a head-mounted display. In addition to changing colors, overlays have also been used to address other visual impairments, such as optical defocus (Itoh and Klinker [2015]), partial field of vision loss (Fardoun et al. [2013]; Amini et al. [2020]) and low vision (Zhao et al. [2016]; Min Htike et al. [2021]). To counteract optical defocus, Itoh and Klinker [2015] generated a compensation image based on an estimation of the user's vision and the current field of view captured with a camera. This combination of stimuli resulted in the desired optical effect and improved the visual perception of affected users. Similarly, Fardoun et al. [2013] and Amini et al.

[2020] also used an estimation of the visual condition in their approaches for people with partial field of vision loss to distort the images at the position of blind spots and move the hidden stimuli to visible areas. For people with low vision, Zhao et al. [2016] and Min Htike et al. [2021] focused on guiding their attention by highlighting important stimuli within the field of view. Using this technique, they were able to improve everyday activities such as finding groceries and avoiding obstacles during navigation.

While most amplification approaches identified in our analysis focused on restoring impaired perceptual capabilities, we also found some examples that aim to amplify what an average person can typically perceive (Aoki et al. [2003]; Knierim et al. [2020]; Eghtebas et al. [2021]). For instance, Knierim et al. [2020] explored how to overcome the temporal limitations of human visual perception when observing fast-moving objects. To achieve that, they used a head-mounted display and slowed down the video stream of the field of view by repeating each frame multiple times to create the desired effect. Based on the same hardware, Eghtebas et al. [2021] developed a prototype that allows users to zoom in on desired parts of their field of view with a simple tap gesture. The system then recognizes the intended physical object and overlays it with a zoomed in digital counterpart. Apart from amplifying visual perception, Aoki et al. [2003] explored how to improve the acoustic intelligibility of conversations in crowded places. For that, they automatically identified groups of dialog participants and reduced the volume of people outside each respective conversational group. This shows that mitigating unintended signals can also result in an indirect amplification of the targeted stimuli.

Substitution

Sensory substitution describes the process of transforming certain types of stimuli from one sensory domain to another (e.g., vision to touch). The primary goal is to compensate for deficiencies of a sensory modality by converting otherwise not or only partially perceivable information into a type of stimulus for which receptors are intact [Deroy and Auvray, 2012]. This transformation enables individuals to still perceive the original sensory information through other channels despite being affected by an impairment or disorder of the corresponding sensory organ. It can also serve as an alternative strategy to amplification in cases where the initial sense is too severely damaged and can not fully perceive the enhanced stimulus.

One group of people that significantly benefits from sensory substitution are deaf and hard of hearing users. To aid their perception of audible signals, our literature analysis revealed several approaches that strive to convey different acoustic aspects to the

Publication	Hearing Substitution	Sensors: Camera	Sensors: Microphone
Matthews et al. [2006]	transcribe acoustic signals to text	○	●
Nanayakkara et al. [2009]	convey musical experience through vibrations	○	●
Jain et al. [2015]	show direction of sound sources	○	●
Sicong et al. [2017]	recognize sounds and convert to event notifications	○	●
Luzhnica and Veas [2018]	convert speech to vibration patterns	●	●
Peng et al. [2018]	display utterances as speech bubbles	○	●
Petry et al. [2018]	convert music rhythm to vibration patterns	○	●
Goodman et al. [2020]	show direction of sound sources	○	●
Jain et al. [2020]	convert loudness to vibration patterns	○	●
Jain et al. [2022]	recognize sounds and convert to text	○	●

Legend: Camera Microphone

Table 4.3: Sensory hearing substitution approaches.

affected group by converting them into alternative modalities (see Table 4.3). For instance, Matthews et al. [2006] and Peng et al. [2018] focused on transforming speech signals during conversations into textual representations to enable the participation of deaf and hard of hearing users. While Matthews et al. [2006] provided the information through text messages on a smartphone, Peng et al. [2018] utilized a head-mounted display to show the dialog contents in speech bubbles next to the corresponding conversational partner. Instead of using a textual representation, Luzhnica and Veas [2018] converted the speech signals into haptic patterns on a wearable vibrotactile display to make the information accessible for users with both hearing and vision impairments.

In addition to conversational content, acoustic events have also been a primary target for hearing substitution approaches (Jain et al. [2015]; Sicong et al. [2017]; Jain et al. [2020]; Goodman et al. [2020]; Jain et al. [2022]). For instance, both Sicong et al. [2017] and Jain et al. [2022] proposed smartphone applications that listen to the current acoustic landscape and recognize certain sound events, such as car honks, fire alarms, or doorbell rings. In case of successful detections, the corresponding event names were displayed on the screens of mobile devices. Jain et al. [2015] took this concept one step further and indicated the direction of sound events combined with their names on a head-mounted display. This enabled deaf participants to localize the sources of sounds and improved their environmental awareness. Similarly, Goodman et al. [2020] used a

smartwatch to indicate the direction of acoustic events and included their loudness as additional information.

Instead of recognizing specific events, Jain et al. [2020] focused primarily on their general properties and converted the loudness to the vibration intensity of a wrist-worn device. Based on the resulting patterns, participants became aware of the soundscape of their environment and could correlate audible events with visual cues. Another area where vibrations have been used is to substitute musical experiences (Nanayakkara et al. [2009]; Petry et al. [2018]). To this end, Nanayakkara et al. [2009] developed a haptic chair that enables the perception of music through tactile feedback. Using the same concept, Petry et al. [2018] proposed a wearable device that not only conveys rhythm information to deaf users but also allows them to play musical instruments themselves in sync with other musicians.

		Sensors			
Publication	Vision Substitution				
Yoshida et al. [2011]	convert image features to sounds	●	○	●	○
Guilbourd et al. [2012]	convert text to speech	●	○	○	○
Banf and Blanz [2013]	convert colors to sounds	●	○	●	○
Nanayakkara et al. [2013]	convert text to speech	●	○	○	○
Tang and Li [2014]	convert object position to sounds	●	○	○	●
Shilkrot et al. [2015]	convert text to speech	●	○	○	○
Woźniak et al. [2015]	convert colors to sounds	●	○	○	○
Carcedo et al. [2016]	convert colors to vibration patterns	●	○	○	○
Boldu et al. [2018]	recognize objects and provide descriptions	●	○	○	○
Zhao et al. [2018]	recognize faces and provide information	●	○	○	○
Feiz et al. [2019]	recognize form fields and provide guidance	●	○	○	○
Ahmetovic et al. [2020]	recognize objects and provide description	●	○	○	○
Boldu et al. [2020]	recognize objects and provide description	●	○	○	○
Lee et al. [2020]	recognize people and provide description	●	○	○	○
Chen et al. [2022]	convert object position to sounds	●	●	○	○

Legend: Camera Microphone Touch Depth

Table 4.4: Sensory vision substitution approaches.

Aside from audible information, sensory substitution has also been used to make visual aspects accessible for people with vision impairments. Through our literature analysis, we identified several approaches (see Table 4.4) that cover a wide spectrum of visual

properties for substitution ranging from fundamental details such as colors (Banf and Blanz [2013]; Woźniak et al. [2015]; Carcedo et al. [2016]) to higher-level information about objects and people (Boldu et al. [2018]; Zhao et al. [2018]; Ahmetovic et al. [2020]; Boldu et al. [2020]; Lee et al. [2020]). One approach that targets the whole spectrum by itself was proposed by Banf and Blanz [2013], who applied a multi-level sonification method to convert colors, edges, patterns, and objects within images into sounds. More precisely, they enabled blind and visually impaired users to explore images on a touchscreen by transforming the visual details at their fingertips into acoustic signals. For that, they mapped the color space to different MIDI instruments, represented patterns and edges with varying drum rhythms, and notified users about objects at the current finger position through synthesized speech.

Other approaches in this space focus more on individual visual properties such as just the colors or edges. For example, Woźniak et al. [2015] proposed the ChromaGlove, a wearable device that converts color information into haptic feedback. The system mainly consists of a color sensor attached to the palm of the glove and a vibration motor that communicates the differences in hue. Following the same principles, Carcedo et al. [2016] prototyped a wristband that indicates the color based on spatial and temporal patterns of the haptic motors positioned around the device. Both systems enabled blind and visually impaired users to perceive the colors of objects around them and improved their color awareness. In addition to colors, Yoshida et al. [2011] focused on conveying the shapes of objects by transforming the edges and contours within images into sound waves with different frequencies. For their application, they used the same interaction principle as Banf and Blanz [2013] and let users explore the images on a touchscreen.

Instead of tasking users with inferring objects by their shapes, another method is to automatically recognize them with machine learning techniques. Boldu et al. [2018, 2020] followed this approach and prototyped a finger-worn as well as a head-mounted wearable device to support visually impaired people while grocery shopping. Both systems use an integrated camera to capture images of the targeted objects, which serve as input for the detection models. Once an object has been identified, a description is generated and read to the user. Ahmetovic et al. [2020] took this concept one step further and proposed a smartphone application that allows blind and visually impaired people to take pictures of their personal objects to train the detection models further. This enabled the recognition of objects that are otherwise difficult to distinguish with pre-trained detection models. In addition to identifying objects, Tang and Li [2014] and Chen et al. [2022] focused on conveying their position to the users. While Tang and Li [2014] employed a head-mounted depth sensor to convert the object's direction and

distance into stereo sounds appearing to originate from the same position, Chen et al. [2022] prototyped a neck-worn dual camera-based wearable that detects the object of interest's position and generates audio instructions on how to reach it.

Besides recognizing objects and their position, identifying people and faces has also been a target of previous research. For instance, Zhao et al. [2018] proposed a smartphone application that utilizes the photos of social network contacts with computer-vision methods to recognize their faces and provide details about their name and current appearance through speech output. Rather than identifying friends and acquaintances, Lee et al. [2020] focused more on a general solution to describe the looks of other pedestrians. For that, they captured pictures of people within a person's field of view with a head-mounted camera and detected their age, gender, looking direction, and distance. Despite the differences in each application, both approaches contributed to improved interactions of blind and visually impaired people with others.

Apart from detecting people and objects, recognizing texts and conveying their meaning is another area of interest for vision substitution research (Guilbourd et al. [2012]; Nanayakkara et al. [2013]; Shilkrot et al. [2015]; Feiz et al. [2019]). To this end, the typical process involves capturing pictures of objects and surfaces containing texts, applying image processing methods to identify and prepare the relevant regions, converting the texts into machine-readable form with optical character recognition (OCR), and outputting the detection results with synthesized speech. The main differences between approaches mostly relate to the utilized hardware and interaction paradigms. While Guilbourd et al. [2012] used two cameras integrated into the frame of glasses to detect texts within the field of view, Nanayakkara et al. [2013] and Shilkrot et al. [2015] prototyped finger-worn wearable devices that can be used to scan documents by moving your hands along text regions. Besides extracting information, the same general process has also been used to identify input areas in forms and guide users to fill out the necessary fields. For that, Feiz et al. [2019] designed a custom 3D-printed smartphone stand with mirrors to redirect the camera input towards documents on a table surface. The resulting images were then used to recognize the relevant form areas as well as the current user position to provide instructions on where to move and what to write.

In addition to these general vision substitution strategies, several approaches found in our literature analysis focused on converting specific spatial aspects to support the navigational capabilities of blind and visually impaired people. An overview of these approaches is shown in Table 4.5. One of the targeted fundamental spatial properties is the distance towards objects and locations (Yatani and Truong [2012]; Twardon et al.

Publication	Substitution	Sensors							
		Camera	Microphone	GPS	IMU	Depth	Eye Tracking	Ultrasonic	Bluetooth Beacon
Guy and Truong [2012]	describe intersection layout	○	○	●	○	○	○	○	○
Yatani et al. [2012]	indicate distance and direction	○	○	●	○	○	○	○	○
Panëels et al. [2013]	describe nearby points of interest	○	○	●	●	○	○	○	○
Twardon et al. [2013]	convert depth information to sounds	○	○	○	○	●	●	○	○
Manduchi and Coughlan [2014]	describe directions towards objects	●	○	○	○	○	○	○	○
Shangguan et al. [2014]	detect crossroads and provide crossing instructions	●	●	○	○	○	○	○	○
Berning et al. [2015]	convert distance to pressure	○	○	○	○	○	○	●	○
Buchs et al. [2015]	convert distance to sounds and vibrations	○	○	○	○	●	○	○	○
Freeman et al. [2017]	indicate proximity of nearby beacons	○	○	○	○	○	○	○	●
Flores and Manduchi [2018]	describe backtracking path	○	○	○	●	○	○	○	○
Guerreiro et al. [2019]	provide indoor navigation instructions	○	○	○	○	○	○	○	●
Kayukawa et al. [2019]	detect pedestrians and warn about collisions	●	○	○	○	●	○	○	○
Kayukawa et al. [2020]	guide users along path to avoid collisions	●	○	○	●	●	○	○	○
Kuribayashi et al. [2021]	guide users to stand in lines	●	○	○	●	●	○	○	○

Legend: Camera Microphone GPS IMU Depth Eye Tracking Ultrasonic Bluetooth Beacon

Table 4.5: Sensory vision substitution approaches for navigation.

[2013]; Berning et al. [2015]; Buchs et al. [2015]). For that, Buchs et al. [2015] prototyped a hand-held device that captures depth information with an infrared sensor and transforms it into haptic feedback. The intention was to build a system that could be used similarly to a cane to scan the immediate surroundings. Based on the same sensing method, Twardon et al. [2013] proposed a head-mounted wearable device that converts depth information at the current gaze position into acoustic signals. This interaction method enabled users to scan the environment with their eyes and construct mental images from the resulting sounds. Instead of relying on the active exploration of the current surroundings, Berning et al. [2015] used an array of ultrasonic sensors positioned around a hat to simultaneously measure the distance towards obstacles in all cardinal directions. The captured information was then transformed into different pressure levels indicated by actuators positioned next to the corresponding sensors.

Aside from depth sensors, the Global Positioning System (GPS) has also been utilized to detect and convert spatial properties. Like the previously mentioned approaches, Buchs et al. [2015] used it to measure the distance and direction towards target locations. The results were indicated through a grid of vibration motors attached to the back of a smartphone. Instead of tactile feedback, Panëels et al. [2013] employed synthesized speech to inform blind and visually impaired users about nearby points of interest based on their current GPS position. The same concept was applied by Guy and Truong [2012] to provide information about the layout and connected streets of nearby intersections. Due to the dangers and challenges related to pedestrian navigation, Shangguan et al. [2014] expanded upon this idea and provided instructions to safely guide users across streets and crossroads. To this end, they used the integrated cameras of smartphones to capture pictures of the road and locate zebra patterns with computer-vision algorithms.

In addition to supporting the traversal through outdoor environments, several indoor navigation approaches have been proposed (Manduchi and Coughlan [2014]; Flores and Manduchi [2018]; Guerreiro et al. [2019]). For instance, Guerreiro et al. [2019] placed multiple Bluetooth beacons around indoor spaces to locate the position of blind and visually impaired users and provide instructions on how to reach their destination. However, this method requires precise location mappings that need to be updated in case of physical changes. One approach that does not rely on specifically prepared environments was proposed by Manduchi and Coughlan [2014]. They used the camera images from smartphones to identify desired objects and guide users towards them. Flores and Manduchi [2018] applied a different method for environment-independent navigation. Instead of supporting people to reach a destination, they focused on guiding them back along the same path. For that, they recorded the sequence of steps and

turns towards a target location with the integrated inertial measurement unit (IMU) from smartphones and generated verbal instructions for backtracking in reversed order.

One challenge that can occur during navigation independent of the applied method is avoiding collisions with obstacles and other people (Kayukawa et al. [2019, 2020]; Kuribayashi et al. [2021]). To this end, Kayukawa et al. [2019, 2020] equipped a suitcase with multiple sensors (camera, IMU, LiDAR) and used the acquired data to analyze the current surroundings. This included monitoring the behavior of nearby pedestrians, predicting the potential risk of intersecting paths, and alerting users to avoid collisions. Kuribayashi et al. [2021] followed the same principles but used the gathered sensor data (camera, IMU, depth) to guide blind people to stand in line behind others and prevent collisions with the person in front of them.

Extension

Unlike amplification, which primarily focuses on improving the perception of signals that could be naturally sensed by humans, sensory extension aims to enable the perception of stimuli outside the natural scope of a given sense but still within the respective modality. This is achieved by shifting the otherwise undetectable signals into the perceivable spectrum of the corresponding sensory domain. As a result, a more comprehensive range of stimuli can be processed to gain additional information about the current surroundings. Examples of such signals include infrared light for thermal or night vision and ultrasonic sounds for echolocation.

		Sensors				
Publication	Extension					
Fan et al. [2014]	overlay back movement	✓				
Kasahara et al. [2016]	combine multiple views	✓				
Abdelrahman et al. [2017a]	overlay thermal & depth vision	✓		✓	✓	
Abdelrahman et al. [2017b]	overlay thermal vision			✓		
Lilija et al. [2019]	overlay occluded views	✓				
Liang et al. [2020]	enable 360 degree field of view	✓				✓
Watanabe and Terada [2020]	map inaudible frequencies		✓			

Legend: Camera Microphone Thermal Depth Eye Tracking

Table 4.6: Sensory extension approaches.

Table 4.6 provides an overview of the sensory extension approaches identified in our literature analysis. As illustrated, most strategies focused on making otherwise unavailable views of our surroundings accessible to people through external cameras and optical sensors (Fan et al. [2014]; Kasahara et al. [2016]; Lilija et al. [2019]; Liang et al. [2020]). For instance, Kasahara et al. [2016] combined the first-person views of four people using head-mounted displays and cameras to enable different perspectives in group activities at the same time. This allowed participants to gain a better understanding of their own physical embodiment and spatial relationship with others. Instead of relying on multiple cameras to capture simultaneous views of our environment, Liang et al. [2020] proposed an alternative approach using only a single 360-degree camera. Combined with a head-mounted display and an eye-tracker to dynamically enlarge the current area of interest, they were able to extend people's field of view to their entire surroundings. Based on a similar idea, Fan et al. [2014] attached a back-facing camera to a head-mounted display and made users aware of activities happening behind their backs. To reduce information overload, they only overlaid the camera image during periods when movement was detected.

Another approach to convey information outside the natural scope of human vision was proposed by Lilija et al. [2019]. Their research focused on different visualization methods to make occluded objects visible and improve interactions with them (e.g., plugging in cables behind a TV). For their prototype system, several cameras and tracking markers on participants' hands were required to achieve the desired optical see-through effect. As an alternative, Abdelrahman et al. [2017a] explored the use of depth and thermal vision to make occluded stimuli visible. The main application areas of their concepts were environments with challenging visual conditions, such as smoke during a fire or fog and snow in mountain regions. In addition to extending visual perception, Watanabe and Terada [2020] investigated different techniques to make inaudible sound frequencies perceivable with our ears. For that, they applied multiple manipulation and transformation methods to the inaudible signals recorded with microphones and conveyed the resulting sounds through earphones.

Enhancement

In general, sensory enhancement also aims to enable the perception of information beyond the natural capabilities of human sensing. However, instead of extending the perceivable scope within a sensory domain, it focuses on making partially abstract properties and information available that originate from other domains than the senses used to perceive them. Similar to substitution, this goal is achieved by converting stimuli

into signals that can be naturally perceived with the human senses. Nevertheless, one significant difference is the absence of a sensory organ that could otherwise be used to capture the original information. Without the transformation process performed during sensory enhancement, it would not be possible to perceive the respective stimuli at all. An overview of enhancement approaches identified in our literature analysis is shown in Table 4.7. One area where sensory enhancement has been used in the past is to improve our awareness of the inner functions and properties of our bodies (Hasegawa et al. [2012]; Norooz et al. [2015]). To this end, Norooz et al. [2015] created a textile prototype that combines biometric sensing with physical models of inner organs to reveal their current state and function. In contrast, Hasegawa et al. [2012] focused more on general body properties and proposed a system that conveys the center of gravity during activities such as skiing to improve control and posture.

		Sensors					
Publication	Enhancement						
Hasegawa et al. [2012]	convey center of gravity	○	○	○	○	●	○
Yamano et al. [2012]	convey navigation direction	○	○	●	○	○	○
Carton and Dunne [2013]	indicate distance to objects	○	○	○	○	○	●
Mateevitsi et al. [2013]	indicate distance to objects	●	○	○	○	○	○
Norooz et al. [2015]	display inner body functions	○	●	○	○	○	○
Kiss et al. [2019]	display swimming direction	○	○	○	●	○	○

Legend: Camera Heartrate GPS IMU Force Ultrasonic

Table 4.7: Sensory enhancement approaches.

Besides conveying internal details of our bodies, sensory enhancement has also been used to provide external information about our current environment (Yamano et al. [2012]; Carton and Dunne [2013]; Mateevitsi et al. [2013]; Kiss et al. [2019]). For example, Carton and Dunne [2013] and Mateevitsi et al. [2013] proposed systems that enable the perception of distances towards objects within our surroundings using tactile feedback. While Carton and Dunne [2013] provided the stimuli through a modified glove, Mateevitsi et al. [2013] employed multiple wearable devices positioned at different locations on a person's body to indicate the sensory information. Similarly, Yamano et al. [2012] and Kiss et al. [2019] also proposed approaches related to the position of objects. However, instead of conveying the distance, they focused on indicating the direction of objects and locations. For that, Yamano et al. [2012] modulated the phase of musical sounds to make it seem like they originated from the same direction as the

targeted object. In place of acoustic feedback, Kiss et al. [2019] relied on visual indicators to convey the intended direction. This included an absolute positioning mode where the color spectrum was mapped to cardinal directions and a relative mode where the stimulus would only indicate if the location was towards a person's right or left side.

Addition

While the creation of new sensory modalities in addition to natural human senses is currently only a theoretical augmentation approach, its foundations are based on the same principles that enable sensory neuroprostheses. For that, artificial sensing devices are used to replace impaired or missing sensory organs. The stimuli gathered this way are then transformed into electrical signals and transmitted through the central nervous system to the corresponding areas of the brain [Pérez Fornos et al., 2019]. However, instead of providing artificial signals that mimic those of a natural sensory organ, the same process could be applied to feed new information to the human brain [Kiss, 2020]. In combination with neuroplasticity, which is the ability of the brain to rearrange neural pathways and structurally adapt itself to different circumstances, new sensing modalities could be created [Costandi, 2016].

4.3.2 Memory Augmentation

Based on our analysis of approaches for human memory augmentation, we found two general strategies that have been commonly applied: (1) capturing a person's daily life with various sensors for retrospective reflection and (2) providing real-time assistance through reminders and access to potentially forgotten information. While there are other approaches that focus on improving the human abilities to encode, retain, and recall information (e.g., through training, repetition, or cues), these methods might not be suitable for people with certain memory impairments or disorders. Consequently, we only included applications that serve as external memory prostheses and can be used by every individual regardless of their memory capabilities or impairments.

Lifelogging

Although there is no universally accepted definition of lifelogging, it can be described as *"a form of pervasive computing consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive"* [Dodge and Kitchin, 2007]. More precisely, it involves capturing and archiving information about daily activities, experi-

ences, thoughts, and even physiological data using various digital tools such as smartphones, wearable devices, cameras, and other data-capturing technologies. The primary objective is to create a detailed and accurate record of a person's life, with the intention of enhancing memory and preserving a comprehensive digital archive for personal reflection and future retrieval. By utilizing technology to capture and store information, lifelogging allows individuals to extend their memory beyond the natural biological limitations of the human mind (see Table 4.8).

		Sensors						
Publication	Strategy							
Hoisko [1999]	capture data at fixed interval	●	●	○	○	○	○	○
Hodges et al. [2006]	recognize and capture events	●	●	○	●	●	●	●
Sellen et al. [2007]	recognize and capture events	●	○	○	○	○	●	○
Lee and Dey [2008]	capture data at fixed interval	●	●	○	●	○	○	○
Chen and Jones [2010]	recognize and capture events	●	○	○	○	○	○	○
Kalnikaite et al. [2010]	recognize and capture events	●	○	○	●	●	●	●
Gouveia and Karapanos [2013]	capture data at fixed interval	●	○	○	●	○	○	○
Niforatos et al. [2017]	capture images manually	●	○	○	○	○	○	○
Niforatos et al. [2018]	capture images at fixed interval	●	○	●	○	○	○	○

Legend: Camera Microphone Heartrate GPS IMU Light Temperature

Table 4.8: Lifelogging memory augmentation approaches.

In general, there are two primary techniques to automatically capture relevant lifelog information: (1) record data in fixed time intervals, and (2) capture information upon recognizing important events and conditions. While it is also possible to let users manually decide which moments are relevant, this method is more intrusive, requires more effort, and can lead to incomplete representations of events [Niforatos et al., 2017]. For these reasons, lifelogging research primarily focuses on automated recording and filtering approaches. This circumstance is further facilitated by the technological advancements of wearable devices, which have become smaller and less intrusive over time, thus enabling people to capture more and more aspects of their everyday lives. For instance, Hoisko [1999] initially intended to utilize a bulky wearable computer connected to a digital camera and a microphone to collect lifelog data. However, considering the higher likelihood of individuals to wear and use a less intrusive system in more situations, they instead opted for a portable audio recorder and a small camera that could only

capture pictures every 30 seconds but fit into a shirt pocket. With this setup, participants were able to record audiovisual data during their entire day. The resulting material was fed into a central database, which could be queried for retrospective purposes.

Gouveia and Karapanos [2013] extended this concept with location data and proposed the *Footprint Tracker* application. It also captured information in fixed time intervals but allowed individuals to review visual, regional, and temporal aspects of past activities on a timeline-based interface. However, this reviewing process can quickly become overwhelming due to the large volume and variety of data generated by lifelogging applications. Therefore, Lee and Dey [2008] proposed a system that lets caregivers preselect and annotate relevant memory cues before they are presented to end-users. While they benefit from not having to inspect all the collected data, this method still requires manual labor from caregivers. Niforatos et al. [2018] attempted to remove these dependencies within the context of work meetings by automatically generating memory cues from the recorded data. For that, they transcribed the contents of previously captured meetings to identify the most frequent topics and generate summaries for each of them.

An alternative method to reduce the amount of collected data is to detect and record only relevant events instead of automatically capturing the environment in fixed time intervals. One of the most frequently used devices for that was the *SenseCam* prototype developed by Hodges et al. [2006]. It is a small wearable device equipped with various sensors that collect data such as images, audio, acceleration, temperature, and light level. Due to its integrated processing unit, it can react to changes in sensor readings and use them to automatically capture certain events. Examples of such triggers include significant changes in light levels or the detection of nearby heat zones.

Sellen et al. [2007] utilized the SenseCam device in their studies and found evidence that it can help people connect to their past. They also compared its efficacy to user-captured images and showed that its performance is similar or even better without requiring manual effort. Kalnikaite et al. [2010] combined the device with a GPS tracker to study how different types of data (i.e., visual and locational) can affect memory recall. Their results indicate that visual cues might lead to more detailed recollections of past events, while locational information reminded people of general behavioral patterns. Chen and Jones [2010] also used the SenseCam device to collect event-based information but primarily focused on developing an interface to improve the retrieval and presentation of personal lifelogs. More precisely, their system provides targeted information to supplement inaccessible memories, offers cues that assist users in reliving past experiences,

and helps individuals to enhance their memory capabilities through repeated presentations of related events and information.

Real-time Memory Assistance

An alternative strategy for memory augmentation is real-time assistance. Unlike lifelogging, which focuses on recording and reviewing past experiences at a later point in time, real-time memory assistance aims to provide potentially forgotten information exactly when needed to support the interactions and daily routines of individuals with memory impairments. To this end, digital tools and algorithms are used that can recognize and deliver the desired information based on the combined data from sensors and personal databases. Regarding the types of provided information, there are two general categories: *retrospective* and *prospective* memories [Niforatos, 2018]. On the one hand, retrospective memories refer to past events and previously obtained knowledge that are currently unavailable to a person [Einstein and McDaniel, 1990]. Examples to support affected individuals with this kind of data include offering instant access to names, faces, and relevant details about the people they encounter, helping them maintain social connections and engage in meaningful conversations. On the other hand, prospective memories refer to future commitments and responsibilities that might have been forgotten [Brandimonte et al., 2014]. To provide assistance for this type of information, reminders and scheduled notifications are typically used. For instance, a system could prompt users to take their medications, attend appointments, or complete essential tasks, ensuring they stay on track with their daily routines and obligations.

		Sensors		
Publication	Strategy			
Sohn et al. [2005]	detect position and provide reminders	○	●	○
Osmani et al. [2009]	recognize activity and provide reminders	○	○	●
Kurze and Roselius [2011]	recognize faces and provide information	●	○	○
Utsumi et al. [2013]	recognize faces and provide names	●	○	○
Iwamura et al. [2014]	recognize faces and show last encounter	●	○	○
Lee et al. [2016]	recognize objects and provide information	●	○	○
Kianpisheh et al. [2019]	recognize faces and provide information	●	○	○
Li et al. [2019]	recognize objects and show last interaction	●	○	○

Legend: Camera Location RFID

Table 4.9: Real-time memory augmentation approaches.

While reminders for prospective use cases are often time-based, other triggers have also been explored (see Table 4.9). For instance, Sohn et al. [2005] proposed the Place-Its application, which lets people create notification messages based on their arrival at or departure from selected locations. To prevent users from ignoring notifications due to being engaged in other activities, Osmani et al. [2009] employed activity recognition techniques to identify their current actions and determine whether it is an appropriate moment to provide a reminder. This allowed them to use activities as triggers and enabled the context-appropriate delivery of prospective notifications.

In contrast, retrospective approaches primarily focused on recognizing people and objects and providing forgotten information about them in real-time. To this end, Utsumi et al. [2013] used a wearable camera to capture images of people's faces as input for detection algorithms. Once a person was identified, their name was shown on a head-mounted display. In addition to providing names, Kurze and Roselius [2011] also presented information based on a detected person's social network profile, such as their current affiliation and recent activities. This knowledge was included to assist users with potential conversation topics and to give them more context about the person in front of them. Apart from delivering general information, Iwamura et al. [2014] investigated the effectiveness of showing videos from the last interactions with the encountered individuals. For each recording they displayed the time and location where it was captured to give users more indicators for remembering previous interactions. Kianpisheh et al. [2019] followed a similar approach but focused on audio instead of video data from past encounters to make the system usable for visually impaired people. They also provided the option to record custom audio descriptions for each subject to complement the already captured information.

Besides recognizing people, tracking objects and their states has been another area of interest for previous research. For example, Li et al. [2019] investigated the potential of using video clips to assist older adults in determining whether they have completed specific actions involving the depicted items. To achieve that, they attached visual markers to objects of interest and assigned labels to each of them. Once the system detected a marker, a short video was recorded with the body-worn camera until the object was no longer within the field of view. Based on the assigned labels, users were then able to watch the recorded clips at any time to remember their last interactions with the objects and understand their current state. Instead of visual markers, Lee et al. [2016] employed image-based object recognition algorithms to detect personal items. The system consisted of a smart glass with an integrated camera to capture the current field of view and

a server to process the recorded images. In addition to classifying predefined objects, it enabled users to train the models with newly captured pictures of personal items.

4.3.3 Cognitive Augmentation

Since higher-order cognition involves various processes, it remains a challenge to identify distinct strategies for each one, as the amount of suitable approaches targeting a specific process is too limited to draw general conclusions. Therefore, we rely on a more universally applicable augmentation strategy proposed by Raisamo et al. [2019] that can be used to address various cognitive conditions. It is achieved by detecting the current state of the user or the environment with sensing hardware, processing and interpreting the collected data with analytical tools, and providing a response that matches the specific needs and requirements of the person in a particular situation. Although this strategy is relatively abstract, it includes all components that have been commonly used by cognitive augmentation approaches identified in our literature analysis.

In this regard, the majority of proposed systems (see Table 4.10) focused on providing interventions for certain negative cognitive states such as low attention (Kern et al. [2010]; Pielot et al. [2015]; Hutt et al. [2021]; Vadiraja et al. [2021]), stress (Pina et al. [2014]; Sharmin et al. [2015]; Flobak et al. [2017]; Howe et al. [2022]), or depression (Fletcher et al. [2011]; Peng et al. [2011]). For instance, Pina et al. [2014] explored the usage of wearable electrodermal activity (EDA) sensors to recognize stressful situations and provide behavioral intervention strategies for parents of ADHD children. Flobak et al. [2017] followed the same principles but focused on detecting and supporting stressful events of adults with ADHD. Instead of a wrist-based wearable, Sharmin et al. [2015] used a chest band equipped with accelerometer, respiration, temperature, galvanic skin response (GSR), and electrocardiography (ECG) sensors to collect data in the wild and train a stress-inference model. They used it to deliver and evaluate different visualizations for adaptive just-in-time interventions in stressful situations. Howe et al. [2022] also explored different intervention types and timing conditions in their research. For that, they combined multiple work-related indicators such as email volume, number of appointments, and time of day with behavioral (i.e., facial expressions) and physiological (i.e., heart rate) reactions into a stress score, which served as the trigger for interventions during a four-week evaluation study.

Apart from everyday or work-related stressors, Fletcher et al. [2011] investigated the effectiveness of using ankle-worn sensor bands to provide cognitive behavioral therapy (CBT) at appropriate moments for patients with drug addiction and post-traumatic

Publication	Strategy	Sensors										
Chang et al. [2008]	detect RFID tags and show navigation instructions	○	○	○	○	○	○	○	○	●	○	○
Kern et al. [2010]	detect attention switching and display indicator	○	○	○	○	○	○	○	●	○	○	○
Fletcher et al. [2011]	detect PTSD and display intervention	○	○	○	●	○	●	●	○	○	○	○
Peng et al. [2011]	detect depression and provide music therapy	○	○	○	○	○	●	○	○	○	○	○
Carriço et al. [2012]	detect position and perform therapy action	○	○	○	○	●	○	○	○	○	○	○
Wang et al. [2013]	detect faces and display names	●	○	○	○	○	○	○	○	○	○	○
Pina et al. [2014]	detect stress and provide parenting strategies	○	○	○	●	○	○	○	○	○	○	○
Pielot et al. [2015]	detect boredom and provide reading suggestions	○	○	○	○	○	○	○	○	○	●	○
Rubin et al. [2015]	detect panic attack and provide intervention	○	○	●	●	○	○	●	○	○	○	○
Sharmin et al. [2015]	detect stress and provide intervention	○	○	○	●	●	●	●	○	○	○	○
Itoh et al. [2016]	detect motion and display predicted trajectory	●	○	○	○	○	○	○	○	○	○	○
Boyd et al. [2017]	detect distance and display social appropriateness	○	○	○	○	○	○	○	○	○	○	●
Flobak et al. [2017]	detect stress and provide intervention	○	○	○	●	○	○	○	○	○	○	○
Leelasawassuk et al. [2017]	detect relevant moment and provide guidance video	●	○	○	○	○	○	○	○	○	○	○
Danry et al. [2020]	detect argument evidence and provide information	●	●	○	○	○	○	○	○	○	○	○
Hutt et al. [2021]	detect mind wandering and provide intervention	○	○	○	○	○	○	○	●	○	○	○
Khan et al. [2021]	detect context and provide habit-support intervention	●	●	●	○	○	○	○	○	○	○	○
Vadiraja et al. [2021]	detect low attention and provide text summary	○	○	○	○	○	○	○	●	○	○	○
Howe et al. [2022]	detect stress and provide intervention	●	○	●	○	○	○	○	○	○	○	○

Legend: Camera Microphone Heartrate EDA GPS IMU Temperature Eye Tracking RFID Phone Distance

Table 4.10: Cognitive augmentation approaches.

stress disorder (PTSD). Based on a similar idea, Peng et al. [2011] employed an electroencephalogram (EEG) sensor to detect depression and deliver feedback based on music therapy concepts. Their results showed a positive impact of musical interventions on a user's cognitive state that might contribute to the prevention of depression in the long term. In addition to therapeutic methods, Rubin et al. [2015] suggested using breathing and relaxation exercises to reduce the symptoms of imminent panic attacks. To achieve that, they monitored physiological signals such as heart rate, breathing rate, and skin temperature with a chest band and predicted upcoming episodes with a regression model to provide exercise or relaxation instructions at appropriate moments.

Another focus of approaches identified in our literature analysis was supporting low attention, boredom, and mind wandering. To this end, Vadiraja et al. [2021] used eye-tracking to detect regions within a document where the reader has shown decreased attention. Based on the recognized text portions, the system extracted relevant keywords and generated summaries, which were provided to the users. One reason that might be responsible for such periods of low attention is task switching. Independent of whether this process is performed intentionally or caused by interruptions, resuming a task often requires effort to restore the previous context. To ease the burden of switching attention, Kern et al. [2010] used eye-tracking to detect and highlight the last fixated region before another task was focused. Their results confirmed the effectiveness of this technique and showed improved completion times for resumed tasks. Hutt et al. [2021] applied the same principles to recognize the process of mind wandering, which occurs when attention shifts to task-unrelated thoughts. Once this state was detected, they generated and delivered interventions with an intelligent tutoring system, which successfully reduced mind wandering and improved knowledge retention of students. One approach that did not rely on gaze data was proposed by Pielot et al. [2015]. Instead, they analyzed usage patterns of mobile phones to infer boredom and provide reading suggestions automatically. Their findings indicate that individuals are more inclined to interact with recommended content when experiencing boredom.

Besides these more common cognitive augmentation targets, several approaches focused on specific use cases with individual solutions. For instance, Boyd et al. [2017] developed a system for autistic children that measures the distance towards other people and indicates the socially appropriate proximity. To achieve that, they iterated their prototype several times through design sessions with end-users and evaluated it in a feasibility study, which showed its potential to aid individuals in becoming more aware of their physical surroundings. Another approach that supports behavioral change but does so in a more universal way was proposed by Khan et al. [2021]. Their idea was to

develop a system that allows users to form habits through automated reminders. To trigger these messages in appropriate situations, they continuously analyzed and identified a person's context based on visual and physiological sensor data. Instead of providing reminders, Leelasawassuk et al. [2017] worked on a system to automatically capture and deliver guidance videos. For that, they combined a lightweight object detector with a head-motion-based attention recognition model to determine when users focused on a specific task. This allowed them to record relevant action sequences and identify moments when users needed assistance.

Furthermore, Wang et al. [2013] proposed an approach to support users with cognitive impairments. Their goal was to improve the daily lives of people with prosopagnosia, which is the inability to recognize faces. To help affected individuals, they developed an application that detects faces within the field of view and displays the names of identified people. Chang et al. [2008] also tried to improve the independence of patients with cognitive impairments. To this end, they proposed an RFID-based indoor wayfinding system intended for users with traumatic brain injury, cerebral palsy, schizophrenia, or Alzheimer's disease. The prototype consisted of multiple RFID tags placed at relevant locations (e.g., intersections, corners, or doorways) and a mobile tracking system to sense nearby beacons. Another location-based approach was developed by Carriço et al. [2012]. Their framework enabled therapists to specify messages and activities that were triggered when users would enter or leave certain zones. As indicated by their preliminary evaluation results, participants were comfortable with the system and could successfully use it to perform therapeutic procedures.

In addition to supporting shortcomings and impairments, some approaches focused on extending the cognitive capabilities of individuals beyond natural limitations. For instance, Danry et al. [2020] explored the possibility of using a mobile explainable AI system to enhance human reasoning. Their proof-of-concept *Wearable Reasoner* consisted of smart glasses with integrated speakers to provide acoustic feedback as well as a natural language processing component running on a smartphone, which analyzed whether evidence was provided for an argument or not. Initial results indicate the effectiveness of their approach, as users assisted by the system agreed more with statements supported by evidence than those without. Another process that benefits from cognitive augmentation is motion prediction. For that, Itoh et al. [2016] used a head-mounted display to show the future trajectory of moving objects in real-time. This enabled users to predict the landing position of objects much more accurately than without the system.

4.4 Design Dimensions

Based on the insights gained from examining related literature in Section 4.2, we performed an additional analysis to identify common design dimensions in order to support the development of new and the classification of existing systems for the assistive augmentation of cognitive processes. To this end, we applied the constant comparative method by Glaser [1965], which is a qualitative analysis approach where records are coded and constantly compared to each other to refine the identified properties until commonly applicable categories emerge that form the basis for new theories. The usage of this method was inspired by previous research, such as the work of An et al. [2020], who applied it to develop a theoretical framework for teaching augmentation.

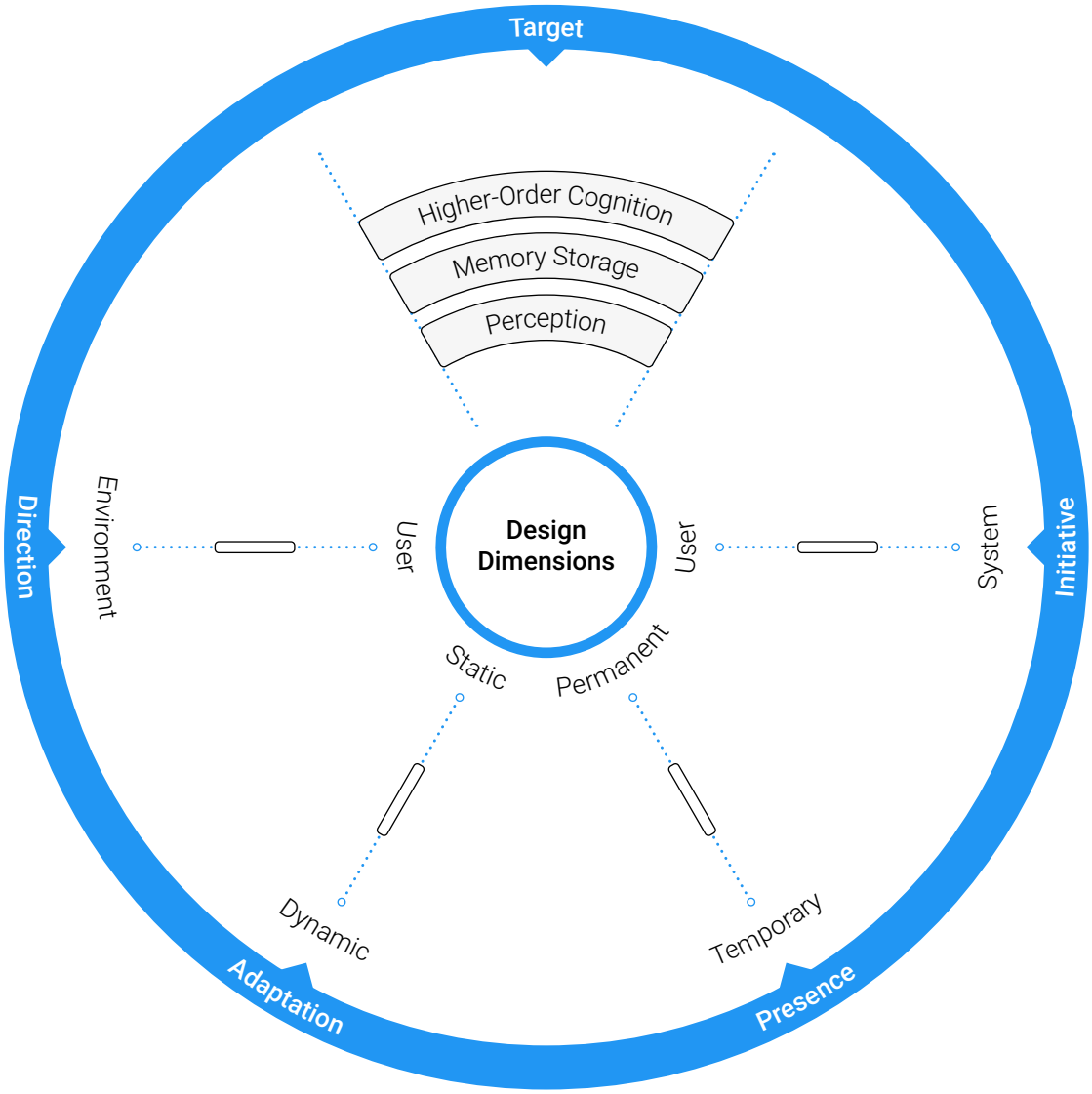


Figure 4.8: Generic template to indicate selected design dimensions.

In the first round of analysis, we annotated every publication identified in our literature review regarding various aspects such as their system structure, sensing hardware, processing steps, feedback type, and augmentation strategy while abstracting over specific hardware details and technical implementations. We then used these properties to find similarities between approaches and synthesized a preliminary set of design dimensions that captured the variations across all analyzed systems. For instance, we found that most sensory augmentation approaches targeted the environment with their sensors while methods for cognitive assistance mainly focused on analyzing the users. As a result, we represented these characteristics with the *direction* dimension (Section 4.4.4). After the initial consolidation, we examined each proposed system again to refine our preliminary categories and to cover the various properties as accurately as possible. An overview of the resulting dimensions is provided in Figure 4.8. It shows the identified categories arranged in the outside border of a circular shape and includes their potential variations on the inside oriented towards the center of the image. We designed the template to quickly indicate the different characteristics of assistive augmentation systems and use it throughout the research probes in Part III to showcase practical examples of its application. The details of each dimension are described in the following sections.

4.4.1 Target

The first dimension determines the targeted cognitive process that should be augmented (see Figure 4.9). Since assistive augmentation systems can have very different goals, such as making sensory stimuli accessible to people, supporting the storage and retrieval of knowledge, addressing certain cognitive conditions, and improving information processing, it is essential to define the intended augmentation target as the first step during the conceptualization phase. Depending on the selected process, certain strategies (see Section 4.3) are more suitable than others to achieve the desired goals.

Perception One potential target of assistive augmentation systems is to support users in perceiving information about their environment or the internal states of their bodies. This includes enhancing sensory experiences such as sight, hearing, touch, taste, and smell. The primary intention is to make unavailable stimuli accessible to individuals either by amplifying senses beyond natural limitations or by circumventing impairments of affected sensory organs. For example, Watanabe and Terada [2020] explored different transformation techniques to make inaudible sound frequencies perceivable by converting them into the audible spectrum. Based on a similar idea, Amini et al. [2020]

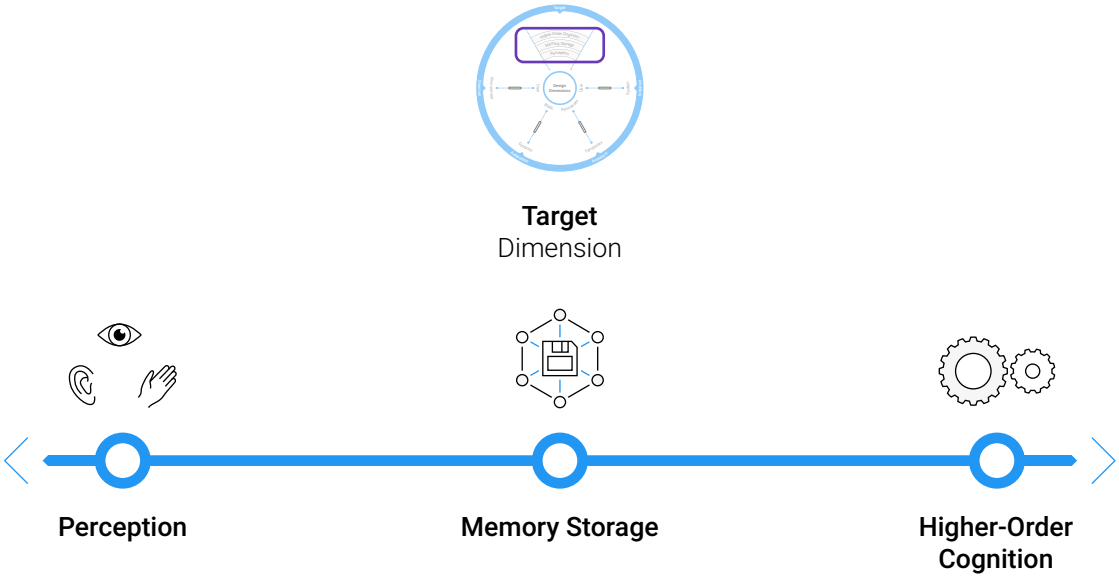


Figure 4.9: Spectrum of the *target* dimension.

transformed the field of view of people with partial vision loss by distorting the images at the position of blind spots and moving the hidden stimuli to visible areas.

Memory Storage Another augmentation target is to assist memory-related processes such as encoding, retaining, and recalling information. Systems that focus on supporting one of these functions try to compensate for memory failures by acting as surrogate information storage that is reliable and not prone to distractions. The general goal is to provide the desired knowledge at the right time in case users are unable to access these memories on their own. To achieve that, a person's experiences can be captured with various sensors and stored for retrospective purposes. For example, Gouveia and Karapanos [2013] proposed the *Footprint Tracker* application, which captured visual, regional, and temporal aspects of past activities and allowed individuals to review them through a timeline-based interface. Alternatively, reminders have been used to provide access to forgotten information (e.g., Sohn et al. [2005]).

Higher-Order Cognition The final target is to aid higher-order cognitive processes that might make use of knowledge gained from prior stages (e.g., perception or memory). Assistive augmentation systems with this intention are usually designed to support complex information-processing steps such as problem-solving or decision-making. They can also enhance certain cognitive conditions, including boredom, stress, and depression. To this end, a person's current state can be determined by capturing physiological and behavioral parameters with mobile sensing technology. Based on a continuous

analysis of the acquired data, the corresponding cognitive condition can be inferred and appropriate feedback can be provided. For instance, Flobak et al. [2017] explored the usage of wearable electrodermal activity (EDA) sensors to recognize stressful situations and provide behavioral intervention strategies for adults with ADHD.

Consideration: Assisted Cognitive Process

Selecting an augmentation target primarily depends on the specific needs and circumstances of the individuals it should support. Therefore, it is essential to consider the users' abilities and limitations to create tailored solutions for their conditions. However, this decision also severely affects the design and architecture of potential systems. Besides requiring specific strategies to achieve the intended augmentations, it also determines certain choices for other design dimensions. For instance, our literature analysis revealed that, on the one hand, most approaches focusing on supporting perceptual processes are directed at the environment since their goal is to make unavailable information about people's surroundings accessible to them. Similarly, the majority of approaches for memory augmentation try to capture knowledge from a person's environment for retrospective purposes. On the other hand, methods to assist higher-order cognition are often focused on the user to determine the targeted cognitive state and provide assistance at appropriate times. This is also the reason why most of these approaches automatically initiate the augmentation process once suitable conditions have been identified. In contrast, sensory augmentation methods are often manually initiated by the user since it is difficult to detect when they might be beneficial. Ultimately, the targeted cognitive state or process should be selected as early as possible based on the individual requirements and demands of the users. Once the objective has been determined, appropriate augmentation strategies (see Section 4.3) should be considered.

4.4.2 Initiative

The second dimension is concerned with the initiation of the augmentation process (see Figure 4.10). This responsibility can either be fulfilled manually by the user or automatically by the system. It is especially important to consider this dimension early on in the design phase of new assistive augmentation systems, as it can severely impact their technical architecture and the workflow of interactions with users. Although each direction has its own set of advantages and disadvantages, the decision primarily depends on the augmented condition and secondarily on the intended level of control.

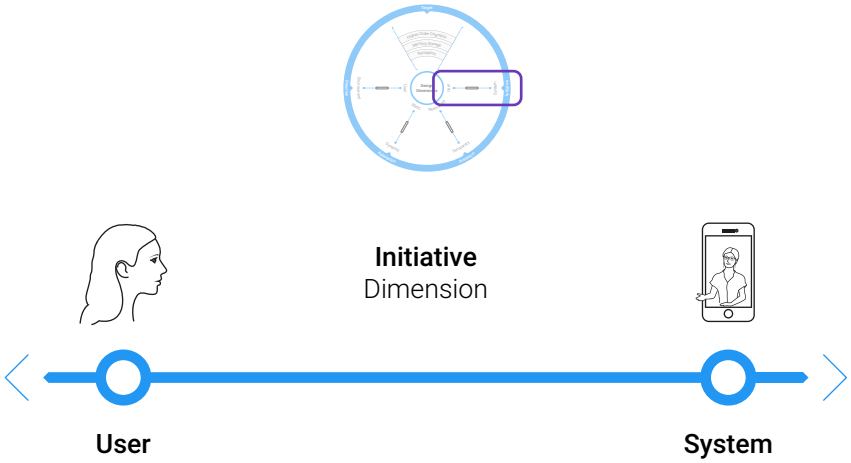


Figure 4.10: Spectrum of the *initiative* dimension.

User In case of manually initiated approaches, the system waits for explicit cues and commands from the user before any kind of assistance is provided. Consequently, the responsibility to recognize when aid is needed and support should be requested lies completely with the users. It puts them in control of the interaction and ensures that assistance is only provided when desired. For instance, Boldu et al. [2018] developed the *FingerReader 2.0* system, which consists of a finger-worn camera that can recognize objects and describe them to blind users at the press of a button. Based on a similar concept, Matthews et al. [2006] proposed a system for deaf people that can transcribe the acoustic landscape from the last 30 seconds on demand.

System For automatically triggered approaches, technology takes the active role and provides assistance without requiring direct input or initiation from its users. To achieve that, the system continuously monitors a person's actions, behavior, or context to determine whether aid is needed in a particular situation. The decision to proactively offer assistance is made based on an analysis of the captured data with predefined criteria or previously learned patterns. Examples in this category range from location- or activity-dependent reminders [Sohn et al., 2005; Osmani et al., 2009] to automatic interventions for certain cognitive states such as boredom [Pielot et al., 2015], stress [Pina et al., 2014; Sharmin et al., 2015; Flobak et al., 2017], or depression [Peng et al., 2011].

Consideration: Manual versus Automatic

In general, shifting the initiative towards the user through manual approaches reduces the complexity of potential systems since additional components to identify the current condition are not required. In the past, the processing capabilities to perform such an

analysis were also not available ubiquitously, which is the reason why early augmentation attempts primarily followed this methodology. Additionally, it gives users more control over the augmentation and is sometimes the only option if the targeted condition or situation can not be detected automatically. However, this method can also lead to circumstances where the augmentation could have been beneficial but was not requested because the user misjudged the situation or was not able to initiate the process due to physical or cognitive constraints. This is where automatic approaches show their biggest potential. By designing assistive augmentation systems so that they automatically recognize critical conditions and proactively provide appropriate interventions, such missed opportunities can be prevented. In turn, this requires careful consideration and tuning of approaches to avoid being overly intrusive or making incorrect assumptions about a user's needs. A compromise between both ends of the spectrum could be to combine them so that the system still anticipates when assistance might be needed, but a confirmation from the user is required before help is ultimately provided.

4.4.3 Presence

The third dimension addresses the intended duration and availability of assistance provided by augmentation systems. It considers when and how long the augmentation should be present to support users effectively. As shown in Figure 4.11, the spectrum for this dimension ranges from permanent to temporary augmentation availability. Each presence level has implications on a user's experience and comes with different requirements for suitable interactions.

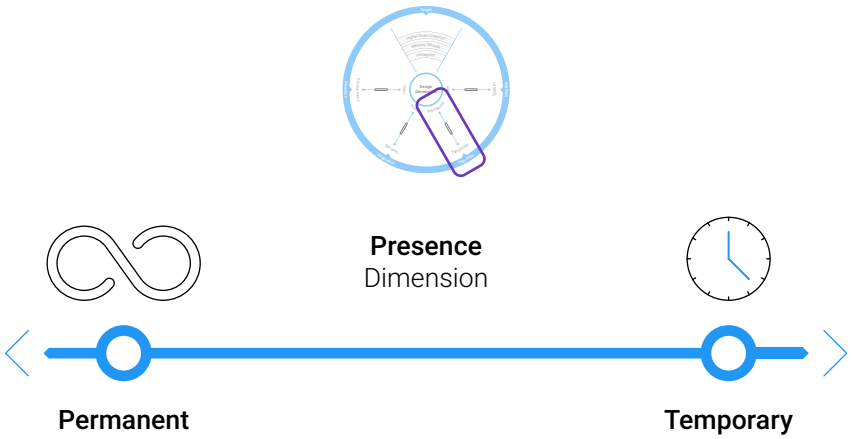


Figure 4.11: Spectrum of the *presence* dimension.

Permanent Assistive augmentation systems can be designed for permanent availability and continued usage over extended periods of time. The primary goal is to offer uni-

versal assistance that benefits users across various tasks and activities. In some cases, the technology to continuously augment fundamental processes can even be treated as an integral part or extension of a person's body. For instance, Langlotz et al. [2018] developed *ChromaGlasses*, which use head-mounted displays to compensate for color blindness in real-time. The system is intended to improve the general viewing experience of affected people and is not restricted to specific use cases. Another example was proposed by Jain et al. [2020], who converted sounds into vibration patterns of a wrist-worn device to improve the general awareness of deaf and hard of hearing users.

Temporary Alternatively, augmentation approaches can also focus on specific situations and provide highly specialized assistance for the duration of selected tasks and activities. Depending on the targeted position in the *initiative* spectrum, the system either waits until it detects a relevant situation or gets enabled by the user when assistance is needed. Thereby, the focus lies on the intended augmentation use case regardless of whether the underlying impairment is permanently present or caused by situational circumstances. For instance, Boldu et al. [2020] developed a wearable device that recognizes and describes grocery items for blind and visually impaired people. Since the system is primarily designed to support this specific use case, it can concentrate its resources on providing the most significant benefits in the targeted situations.

Consideration: Universal versus Situational

While permanent augmentation approaches are usually more universally applicable, they also have stricter requirements regarding their reliability and interaction experiences. Consequently, systems on this side of the spectrum need to ensure that the augmentations can be used for extended periods of time without becoming disturbing or annoying. Additionally, the provided assistance must be robust enough to work across various situations and activities since users often rely on these approaches to extend or replace fundamental cognitive processes. System failures or malfunctions during such interactions can severely impact a user's trust and acceptance of future augmentation technologies, which is why potential consequences should be considered and mitigated early in the design phase. In contrast, the requirements for temporary approaches are less strict due to their limited scope and interaction duration. For instance, a sound used to convey information might become distracting when played over longer periods but could be suitable for short durations in specific circumstances. One advantage of situational approaches is that the conditions in targeted use cases have less variance and can be determined more easily beforehand. This enables designers to incorporate these

factors and build specialized augmentation systems. Ultimately, the decision in this dimension depends on whether the goal is to address a specific situation or to provide universal assistance for a permanent condition.

4.4.4 Direction

The fourth dimension focuses on the sensing direction of assistive augmentation systems (see Figure 4.12). To indicate which aspects of reality are being analyzed, the following data sources can be employed: user-related properties, environment-based information, or a hybrid combination of both. The decision of which direction to choose depends on various factors, such as the augmentation target and the context where assistance is required. It also results in different implications for suitable sensing hardware and methods to process the captured information.

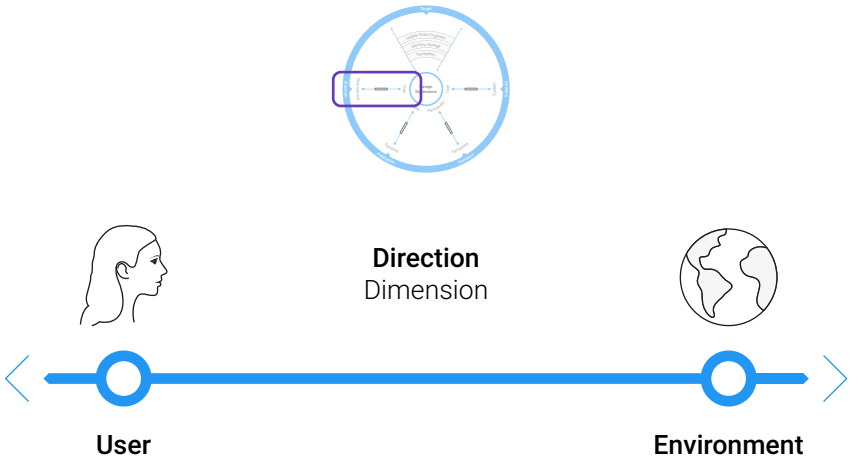


Figure 4.12: Spectrum of the *direction* dimension.

User Approaches at this end of the spectrum primarily focus on gathering information directly from the users. This involves capturing signals and data generated by a person's body, such as movements, facial expressions, or physiological reactions. While it is often less intrusive to only rely on external signals, recent advancements in wearable sensing technology also enabled the usage of internal body properties. For example, Rubin et al. [2015] used a mobile electrocardiogram (ECG) sensor to measure heartbeat symptoms and predict imminent panic attacks. Similarly, Peng et al. [2011] employed an electroencephalogram (EEG) sensor to recognize the degree of depression and adjust a music therapy system accordingly.

Environment At the other end of the spectrum, augmentation systems primarily rely on information captured from the environment. For that, various sensing devices such as cameras, microphones, or temperature probes can be used to monitor the surrounding conditions. The collected information then gets processed to provide the appropriate augmentation based on the captured context. For instance, Kayukawa et al. [2019] utilized cameras and vision-based algorithms to identify obstacles and support the navigational capabilities of blind and visually impaired people. Other approaches use the captured environmental data more directly, such as Carcedo et al. [2016], who converted color information into tactile signals.

Consideration: Individual versus Context

For systems that focus on gathering information from individuals, the collected data usually serves as the foundation to either control the augmentation or to determine a user's current state (i.e., whether assistance is needed or not). Meanwhile, the main goals of approaches at the right end of this dimension are to provide users with information about their current environment that would otherwise be unavailable to them and to deliver appropriate assistance based on the context inferred from the captured signals. A middle ground between both directions would be to analyze the users and their environment simultaneously. For example, Twardon et al. [2013] calculated the current gaze position with an eye-tracker, measured the distance towards the focused object with a depth sensor, and converted the captured information into acoustic signals. However, such hybrid approaches also increase the system complexity and processing requirements compared to methods focusing on a single direction.

4.4.5 Adaptation

The fifth dimension refers to the degree to which a system can be customized and adjusted after its initial development and deployment (see Figure 4.13). On the one hand, augmentation approaches can focus on fixed conditions and provide static assistance that does not require further changes for the intended purposes. On the other hand, certain preferences, circumstances, or capabilities might evolve over time and require systems that can be dynamically adapted. Depending on the characteristics of the targeted condition, a suitable direction for this dimension should be selected.

Static Towards the left side of this dimension, augmentation approaches are designed to provide assistance in a fixed way, exactly as defined during their initial conception.

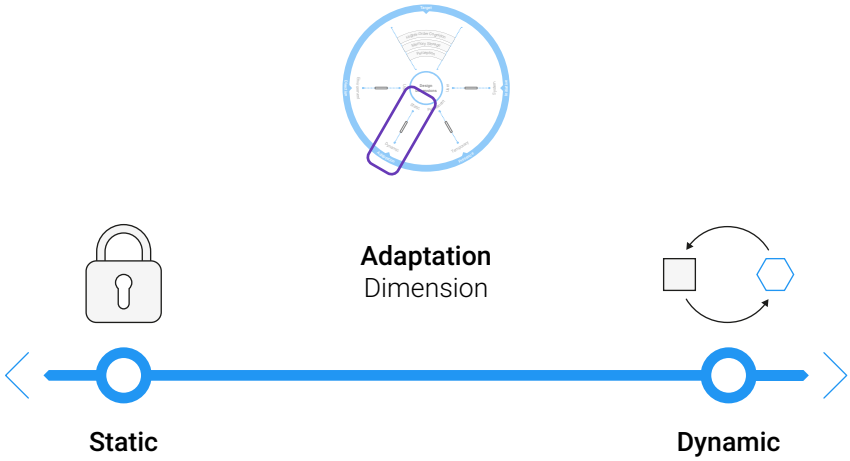


Figure 4.13: Spectrum of the *adaptation* dimension.

This methodology is especially suitable for permanent impairments and disorders (e.g., deafness or blindness), which are relatively stable and do not change or deteriorate over time. Since the circumstances of these conditions can be determined in advance, it is possible to tailor potential augmentation systems specifically to their targeted use cases. For instance, Zhao et al. [2018] initially conducted exploratory interviews with visually impaired participants to gather circumstances, challenges, and requirements from their experiences during social activities. Based on these insights, they developed a system to support the recognition of nearby people, which was the most frequently requested feature for the surveyed user group.

Dynamic Towards the other end of this spectrum, augmentation systems are characterized by their options for customization and their ability to dynamically react to changed conditions. These adjustments can either be performed manually by the user or automatically by the system. Thereby, various factors such as user inputs, context information, behavior patterns, or physiological data might be considered to adapt the approaches and provide the most suitable assistance in a given situation. For instance, Leelasawassuk et al. [2017] developed the *GlaciAR* platform, which automatically produces guidance videos based on the unsupervised observation of users performing certain actions. The system also learns to determine the appropriate conditions for providing these guides and can adapt to different circumstances. Another example that focuses more on the customization aspect was proposed by Lee et al. [2016], who worked on an object recognition system for memory augmentation that could be trained by users to detect personal items and provide associated labels.

Consideration: Constant versus Adaptive

When targeting stable and clearly defined conditions, static assistance is usually the preferred method to provide appropriate support. It does not require any adaptation or customization mechanisms, which reduces the complexity of potential systems. However, if a certain degree of personalization is desired or the intended circumstances evolve over time, adaptive approaches are more suitable. This especially applies to situations where frequent adjustments are required due to unstable conditions or changing needs and preferences. In these cases, dynamic systems can learn from past interactions with users and adapt their assistance in real-time. Alternatively, they can offer methods to perform these adjustments manually to give users more control over the augmentation and increase their agency. Thereby, the same considerations between manual and automatic solutions apply as in the *initiative* dimension. While manual adaptation often results in less complex systems with more customization options for users, it can also lead to missed opportunities due to judgment errors or bodily constraints. To prevent such situations, automatic approaches can be used that recognize the current conditions and adapt their assistance accordingly. In turn, dynamic approaches require more computational resources and precise tuning to avoid incorrect detection results.

4.5 Summary

This chapter established the conceptual and theoretical foundations of assistive augmentation. As part of that, it first categorized the paradigm within the larger context of human augmentation to provide a better understanding of related terms and concepts. While most of them focus on enhancing the abilities of the average population beyond natural limitations, assistive augmentation considers the personal needs and circumstances of individuals for the development of tailored solutions that recover and amplify their specific sensory, memory, and higher cognitive capabilities. After establishing the theoretical context, a comprehensive literature analysis was conducted to derive shared characteristics among related works and identify suitable strategies for each group of cognitive processes. One of the most prominent similarities across the examined approaches was their technical structure, which primarily consisted of sensing, processing, and feedback components. Even though some applications only utilized a subset of the identified modules, the shared architecture highlighted the requirements and expected capabilities of potential systems and served as a foundation for our universal framework in Chapter 6. Besides technical similarities, our analysis also revealed commonly applied sensory, memory, and cognitive augmentation strategies and included concrete

examples to inform the selection of appropriate methods. Based on these insights, we subsequently derived five general design dimensions that can be used to classify and compare existing approaches and guide the design process of future augmentation systems. For each dimension, we also illustrated the different variations and their implications with practical examples to facilitate appropriate design choices. The real-world application of these dimensions is demonstrated in Part III.

Chapter 5

Mobile Signal Processing

Analyzing, processing, and responding to behavioral, physiological, or environmental signals is a core aspect of assistive augmentation systems. It enables them to provide otherwise inaccessible information to users, determine people's current state, and infer whether they require support based on the identified circumstances. Without it, most approaches would not function properly and could not achieve the intended assistance. Consequently, this chapter provides an overview of the stages involved in processing signals on mobile devices. The first step typically focuses on selecting appropriate sensing and processing hardware since it influences all further decisions. Following that, one of three general methods can be used to process the captured signals: (1) algorithmic calculations, (2) feature-based machine learning, and (3) end-to-end deep learning. Depending on the targeted scenario and chosen procedure, different workflows need to be applied. The most straightforward approach is to perform algorithmic calculations directly on the sensor data. Despite its simplicity, this method can be highly effective in suitable situations and requires no further actions besides selecting appropriate hardware and implementing the respective computations. A basic example would be a system that captures environmental sounds with a microphone and amplifies the volume of specific frequencies within the audio signal to improve the acoustic perception of individuals affected by hearing impairments.

In contrast, the two other methods rely on machine learning techniques that necessitate more elaborate procedures. After selecting appropriate sensing devices, the next step typically involves recording and annotating data samples from participants or using existing datasets, which serve as the foundation for potential classifiers. At this stage, both remaining approaches start to differ from each other. While deep-learning models can

be trained directly on the collected data, traditional machine-learning algorithms require identifying and implementing effective features to extract relevant characteristics. After outlining common challenges of mobile signal processing systems and discussing potential solutions, the details involved in selecting appropriate hardware, collecting training data, and developing recognition models are described in the following sections. Additionally, the practical application of all stages is demonstrated in three research probes that showcase the processing of sensor data using algorithmic calculations (Chapter 7), feature-based machine learning (Chapter 8), and end-to-end deep learning (Chapter 9).

Parts of this chapter are based on the following publications:

Reference Seiderer, A., Dietz, M., Aslan, I., and André, E. (2018). Enabling Privacy with Transfer Learning for Image Classification DNNs on Mobile Devices. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs), Conference Proceedings*, pages 25–30. ACM.

Reference Dietz, M., Aslan, I., Schiller, D., Flutura, S., Steinert, A., Klebbe, R., and André, E. (2019). Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 149–158. ACM.

5.1 Challenges

While mobile signal processing systems can provide significant benefits to individuals, some approaches might also affect users in different ways. Neglecting these potentially adverse effects can lead to unintended consequences and undesirable situations, which would counteract the overall goal of supporting individuals [Raisamo et al., 2019]. For this reason, we examine common challenges and concerns related to mobile signal processing approaches and offer potential solutions to address them in future systems.

5.1.1 Privacy Concerns

Information privacy can be generally defined as *"the claim of individuals [...] to determine for themselves when, how and to what extent information about them is communicated to others"* [Westin, 1967]. While this principle should be applied whenever possible, assistive augmentation approaches usually rely on sensor data that contains sensitive information about the users, their environment, or both to provide the intended augmentations. Without it, systems might not operate reliably or function properly at

all, which is why a trade-off in favor of convenience is often made [Glenn and Monteith, 2014]. However, once personal data leaves the user's device, privacy issues and concerns start to arise. Examples include the ownership of transmitted data, third-party access, and unregulated usage.

Additionally, sensitive information about a user's personal life, habits, and relationships could be inferred or directly extracted from the acquired data [Seiderer et al., 2018]. Companies could then aggregate, store, and use these details for personalized advertisements, individually tailored offers, and predictive modeling of user behavior to elicit purchase decisions and maximize their profits. While these consequences might not seem too critical, access to such sensitive information can be extremely harmful to individuals [Raisamo et al., 2019]. This especially applies to cognitive augmentation, where a user's most private data, such as the current mental state, is analyzed. For instance, insurance companies could raise premiums or revoke contracts based on early signs of impairments or disorders [Christin et al., 2011]. Employers could monitor cognitive performance during work and deduct payments if certain thresholds are not met. Medical conditions and political views could even be inferred and used to justify suspensions or deny job opportunities. Corporations that benefit the most from monetizing personal data on a large scale further exacerbate these issues by actively advocating against privacy and portraying it as an expensive relic from the past that holds back efficiency and innovation [Glenn and Monteith, 2014].

In order to counteract these issues and let users regain control of their personal data, several techniques can be applied. One of the simplest methods for that is to reduce the data fidelity by adjusting the sensing granularity and recording intervals [Christin et al., 2011]. These measures can range from reducing the rate at which samples are collected (e.g., every hour instead of every minute) to automatically removing unnecessary details in the recorded data (e.g., blurring people's faces). Although sensitive information might still be disclosed with this solution, it improves the overall acceptance of users by allowing them to control when and how often sensing is performed.

Another approach involves encrypting the private data on the user's device before processing it on a server [Chabanne et al., 2017; Zhang et al., 2016]. There, model training and inference are performed on the encrypted data while the results are sent to and decrypted on the user's device [Gilad-Bachrach et al., 2016; Le Phong et al., 2018]. Through that, the more powerful computational resources of cloud services can be utilized without having to disclose unencrypted data. Nevertheless, there is always the risk that the encryption might be broken. For this reason, alternative approaches with the

same underlying idea have been proposed, such as using differential privacy instead of encryption to protect sensitive information [Abadi et al., 2016; Papernot et al., 2016]. The goal of these methods is to transform the data by adding noise and other perturbations in a way that ensures the same overall results but prevents the correlation to and identification of individuals [Dwork and Roth, 2014].

Finally, the safest approach would be to perform all necessary calculations directly on the mobile device. However, since the required computational resources for training complex neural networks from scratch currently exceed the performance of most mobile devices and would require large amounts of data from individual users, this is no feasible option for now. Instead, an alternative might be to start with pre-trained models, which can be adapted to the current task on the targeted devices. One method that uses this technique is called transfer learning and is often applied in situations where we want to solve a classification problem in one domain but only have a sufficient amount of training data in another domain [Pan and Yang, 2010]. Since modifying an already trained model requires fewer resources, it is possible to perform this task on mobile devices in a reasonable amount of time, as shown by Seiderer et al. [2018]. Federated learning takes this concept one step further by deploying a pre-trained base model to multiple client devices that perform additional training steps based on personal user data [Chen et al., 2020]. The parameters and weights of the resulting models then get uploaded to a central server, where they are aggregated and used to train a new base model. Subsequently, the new model is provided to all clients, and the process is repeated until it converges or the training data is exhausted [Rudovic et al., 2021]. The biggest advantage of this method is that only parts of the models (e.g., parameters and weights) are shared, while the private data never leaves the user's device [Bonawitz et al., 2021]. Further information about the different methods to protect the privacy of users can be found in the literature review by Boulemtafes et al. [2020].

5.1.2 Technical Limitations

The opportunities for processing signals on mobile devices are currently constrained by various technical limitations. One of them concerns the lower computational power compared to stationary desktop computers and dedicated servers. The primary reason for that are physical restrictions, which prevent the usage of more powerful hardware in portable devices. Besides space constraints, energy consumption is a major limiting factor since modern high-performance processors can require several hundred watts at full load, which currently can not be provided by reasonably sized batteries. Their high

power draw also leads to a lot of excess heat, which requires active cooling solutions to prevent the hardware from overheating and throttling. For these reasons, manufacturers equip mobile devices with much less powerful processing units that can be supplied by small batteries for at least a day of moderate use and cooled with less efficient but passive and silent solutions.

However, heavy utilization can reduce the typical runtime to a few hours, which would necessitate recharging the devices multiple times per day. For instance, Lu et al. [2012] observed a runtime drop from 32.6 hours during idle to 9.6 hours while processing audio signals and classifying stress. Since frequent recharging can get cumbersome and might lead to the reduced usage of potential systems over time, viable methods to prevent that from happening should be considered. One solution involves employing efficient algorithms to reduce the power draw and increase battery life. Alternatively, the sampling rate could be lowered to decrease the frequency of computationally intensive operations. If their results are not required immediately, heavy calculations could also be queued and only executed during regular charging intervals. While offloading complex processing tasks to remote servers would be another option, it is less advisable since people's privacy could be infringed once personal data leaves their devices (see Section 5.1.1).

One major disadvantage is that most of these solutions increase the processing latency and reaction time of potential systems (lower sampling frequency, postponed execution of heavy calculations, and transfer delays during offloading). This not only reduces responsiveness and perceived performance but also leads to situations where delayed feedback is provided at inappropriate moments or information becomes inaccurate by the time it gets delivered. Consequently, a balance between computational performance, energy consumption, and response times needs to be achieved. Another aspect concerns the synchronization of multiple signals with different data types from various sensors. The temporal alignment of these data streams is essential to draw reliable conclusions about the occurrence of specific events. For example, if audio and video signals are out of sync, wrong facial expressions might be associated with the corresponding vocal cues, which could lead to false assumptions. To prevent such mistakes, multimodal sensor data can be synchronized with various methods, including timestamp comparisons, sampling rate adjustments, and calibration events (e.g., clapping to match audible sounds with visual motion). Apart from processing-related challenges, the lower accuracy of mobile sensors compared to clinical devices is another limitation. While this discrepancy may decrease due to future technological advancements, the current generation of wearables and portable sensing hardware does not yet achieve the same level of precision. Additionally, the quality and consistency can vary widely between devices,

leading to inconsistencies and unreliable data. In order to minimize these negative effects and facilitate the selection of suitable mobile sensing instruments, their signal quality should be examined. For instance, Barrios et al. [2019] compared the accuracy of multiple heart rate sensors during different activities and showed that more intrusive devices like arm and chest bands are more reliable than wrist-based wearables.

5.1.3 Ethical Considerations

A major challenge of mobile signal processing approaches concerns user acceptance and trust. The primary reasons for that are potential reservations against the invasive nature of sensing devices and the fear of constant surveillance. Combined with the uncertainty of what happens to the collected data, these aspects can reduce peoples' trust towards and acceptance of ubiquitous solutions. For example, individuals might feel self-conscious and experience discomfort due to devices restricting their behavior or the awareness of being monitored, which could lead to the abandonment of augmentation systems despite their positive effects. To counteract these concerns, a compromise between accuracy and intrusiveness must be made. Additionally, approaches should be transparent about applied processing steps and provide comprehensible information regarding decisions related to personal data. Another method to increase people's trust is to directly involve them in the design and development process. This allows individuals to better understand the challenges, considerations, and selected solutions of mobile augmentation approaches. Moreover, it gives them the opportunity to shape conceptual decisions based on their feedback and experience the resulting changes in practice.

Besides acceptance and trust, cultural and demographic biases in mobile signal processing systems are further challenges to consider. They typically occur as a result of imbalanced datasets that only partially capture the respective characteristics and can lead to misinterpretations of signals from underrepresented individuals [Canali et al., 2022]. For example, if a recognition model learns to associate specific behaviors with respective conclusions based on a corpus with limited diversity and is then applied in another cultural context where the same gestures might have different meanings, the resulting predictions will be false. Similarly, models trained with data from young and healthy individuals might translate poorly to older adults affected by impairments of cognitive processes. For these reasons, it should be ensured that the intended user groups and individuals are adequately represented in the training data. To this end, publicly available datasets should be analyzed beforehand to identify the distribution of relevant characteristics and determine their suitability [Torralba and Efros, 2011]. If the intended range of

properties is not present, an alternative option is to directly record the signals of target users to ensure their proper representation. Unfortunately, capturing a sufficient amount of samples can be labor-intensive, time-consuming, and sometimes impossible. In these cases, data augmentation techniques could be used to generate artificial signals. However, to prevent these synthetic samples from introducing potential biases themselves, reporting mechanisms and regular system audits should be implemented.

5.2 Hardware Selection

The selection of suitable sensing hardware and recording devices is essential to accurately capture, process, and interpret the conditions and circumstances of individuals in real-world environments. It directly impacts the quality, scope, and reliability of the collected signals and affects the comfort, convenience, and intrusiveness of potential systems. However, due to the rapid technological advancements in this space, recommendations for specific devices can quickly become outdated. For instance, even though the Microsoft Band 2 was one of the few wearables that combined input and output capabilities with relatively accurate measurements, its production was discontinued after only one year, and the necessary activation servers were shut down shortly after, making further use of the device impossible[1]. Consequently, this section focuses on the general characteristics of mobile sensing hardware and provides universal guidelines for consideration during the selection process. As outlined in Chapter 3, the analysis and interpretation of non-verbal signals can yield valuable insights about a person's state and condition. To capture these parameters, various types of sensors can be utilized. Examples range from common devices like cameras and microphones to specialized equipment such as eye-tracking glasses and electrodes for physiological measurements. Based on the spatial distance to their target, these sensors can be classified into two broader categories: contact and remote devices.

The first category includes sensors that require direct contact with a user's body. This is typically achieved by attaching the devices directly to a person's skin or embedding them in clothing (e.g., shirts or pants) and worn accessories (e.g., wristbands or rings). Due to the spatial proximity to the origin of most physiological signals, contact sensors enable the measurement of parameters, such as heart rate, skin conductivity, muscle activity, and breathing rhythm, that otherwise cannot be captured at all or only with less accuracy from a distance. For instance, the electrodes of an electroencephalography

[1] https://support.microsoft.com/en-us/help/4467073/end-of-support-for-the-microsoft-health-dashboard-applications

(EEG) sensor must be placed directly on a person's scalp to detect voltage fluctuations generated by neuronal activations [Biasiucci et al., 2019]. With the current state of technology, there is no alternative option to acquire these signals remotely. Similarly, the electrodes of an electrocardiography (ECG) sensor must be attached near a person's heart to capture the electrical signals emitted by the cardiac muscle.

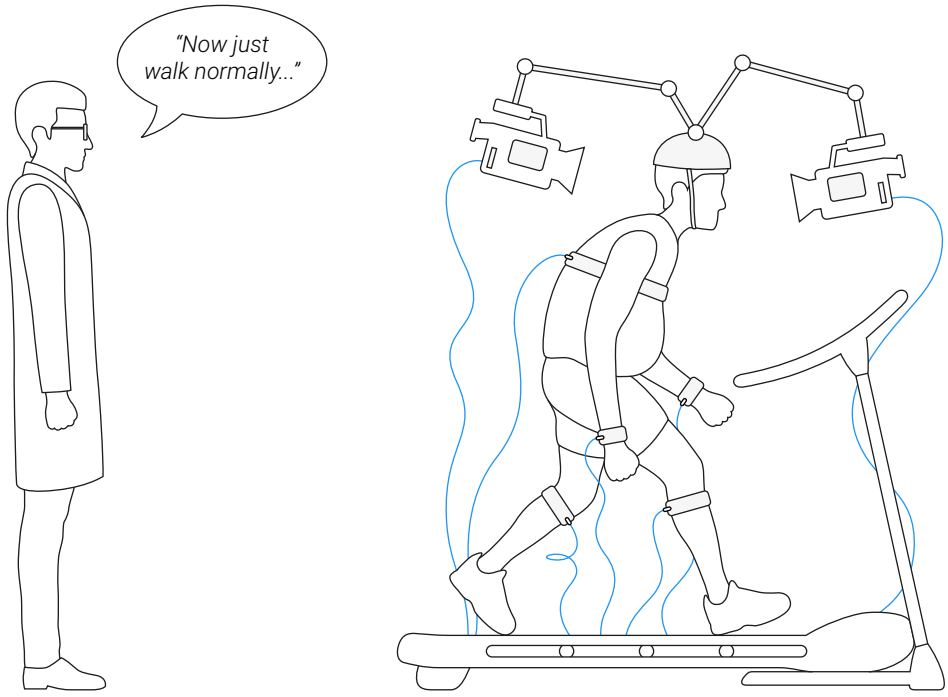


Figure 5.1: Example of obtrusive monitoring (based on cartoon by Wim Boost).

Another advantage of contact sensors is their increased robustness against environmental influences. For example, accelerometers placed on a person's limbs can be utilized to track their body movements without having to consider potential occlusion issues of camera-based solutions. However, the accuracy and data quality of contact sensors also depend on the methods used to acquire the respective signals [Canali et al., 2022]. While a person's heart rate can be inferred by illuminating the skin with an LED and analyzing the amount of reflected or absorbed light (photoplethysmography), measuring electrical impulses with an ECG sensor is typically more reliable and less prone to movement artifacts. In turn, the gained accuracy often comes at the cost of an increased obtrusiveness regarding the size, weight, or comfort of sensing devices (see Figure 5.1). Since physically attaching such objects to a person's body can lead to encumbrance, movement restrictions, and intrusions of their personal space, contact sensors are not always the best solution. Although some of these limitations could be tolerated over short periods, the primary reason against it is that all of these factors can influence people's

behavior and would falsify the recorded data [Ouwerkerk et al., 2008]. For instance, capturing a person's skin conductivity with electrodes attached to their fingers severely impacts their ability to interact with objects, leading to increased stress and altered behavior, which is not ideal when trying to measure their emotional state during everyday activities. In some cases, reducing these negative effects to a minimum still does not produce the desired outcomes since even the awareness of being recorded can influence people's decision-making process and resulting actions [McCambridge et al., 2014].

To prevent behavioral influences caused by intrusive devices, remote sensors can be used as a viable alternative. They are capable of capturing relevant signals from a distance and do not require direct contact with their target. While they are typically placed throughout the environment, they can also be embedded in mobile devices like smartphones and tablets. This integration enables the unobtrusive recording and analysis of individuals during natural interactions. For example, cameras can capture eye gaze, facial expressions, gestures, posture, and spatial positions of users within their vicinity. Similarly, microphone arrays can pick up sounds and vocal cues like tone, pitch, speed, and rhythm, which can serve as indicators to derive a person's emotional state. Wireless signals can also be used to measure physiological parameters, including breathing and heart rate [Adib et al., 2015]. However, the reduced intrusiveness of remote sensors is often associated with lower accuracy and higher susceptibility to noise and disturbances. Additionally, their coverage strongly depends on the positioning throughout the environment, which makes them less flexible and only usable in specific situations.

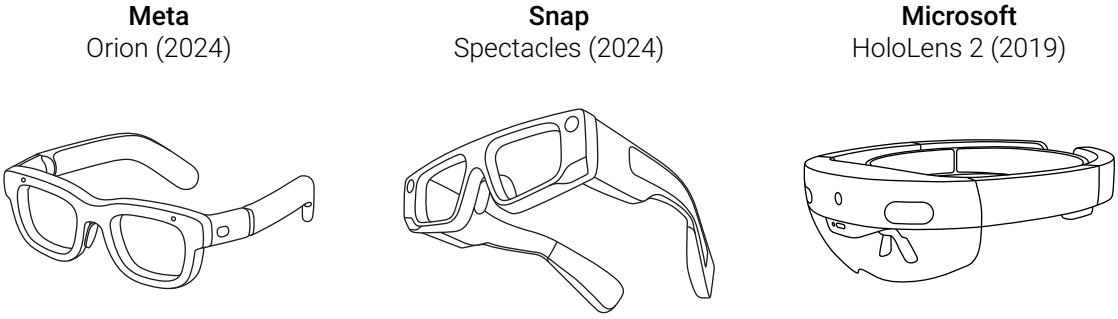


Figure 5.2: Examples showing trade-off between bulkiness and processing capabilities.

Overall, the decision between contact and remote sensors primarily depends on the intended scenario and its requirements. Since accuracy, mobility, robustness, and ubiquitousness play important roles in assistive augmentation approaches, mobile contact sensors are typically preferred. Although it is possible to use remote sensors if the intended augmentations should only occur in fixed locations, these environments still require specific preparations, which are often not feasible when aiming to support many

individuals. Besides choosing suitable sensing devices, selecting appropriate processing hardware also involves considering similar factors. While stationary computers and servers generally provide much more computational resources, smartphones and tablets have reached a point where most tasks can be performed directly on people's devices. These advancements have also led the development of new technology-enabled hardware like smart glasses, smartwatches, and smart rings. However, achieving such small form factors is currently only possible through trade-offs regarding processing power, battery life, or weight. For this reason, most augmented reality glasses are still relatively bulky since they require certain computational capabilities that can not be sacrificed in favor of a smaller chassis (see Figure 5.2). Another point to consider is the operating system of these processing devices. While most commercially available wearables like fitness trackers and smartwatches are compatible with both Android and iOS, some specialized sensors only support Android due to its openness and customizability. Combined with the large variety of available devices and their more affordable price points, these properties have made it the most widely used mobile operating system, which also makes it an ideal platform for assistive augmentation approaches.

5.3 Data Collection

Following the selection of appropriate sensing and processing hardware, another important step towards achieving effective recognition models is to acquire relevant training data. For that, the most common procedure involves conducting studies with participants and recording their emitted signals (see Chapter 3) during the targeted conditions. However, depending on the planning and execution of the recording process, the resulting data quality can vary significantly. Therefore, considering available guidelines and suitable options is essential when designing and conducting new recording experiments.

One of the first and most impactful decisions concerns the general study circumstances. In addition to traditional laboratory settings, mobile sensors and devices enable the recording of relevant signals in the wild. While this scenario typically leads to more natural and realistic data, it is also more prone to errors and unexpected situations. Nevertheless, it can reveal potentially overlooked challenges and conditions that could occur during real-world usage, which can be utilized to improve the robustness of the final system. In contrast, laboratory studies can be conducted and repeated in fully controllable environments, which mostly prevents external influences but requires suitable artificial stimuli to reliably elicit the indented conditions. Otherwise, the desired effects and responses might not occur, leading to unusable recordings and smaller datasets [Schmidt

et al., 2019]. In the worst case, improper stimuli could even result in unnatural reactions that do not appear outside the lab and would completely invalidate the experiments. Consequently, collecting data in field studies is typically preferable when building mobile systems for everyday use since they provide a more realistic foundation. However, in case field studies are not feasible due to specific limitations (e.g., resource constraints or environmental requirements), it is essential to mimic real-world circumstances and events as closely as possible when conducting laboratory studies instead.

Apart from defining the experiment conditions, other critical aspects include participant screening, selection, and recruitment. The goal of these steps is to find people who ideally represent the entire spectrum within the target user group and are willing to participate in the study. For that, interested individuals are typically assessed with screening questionnaires, which can result in exclusions if certain requirements are not met. Unfortunately, gathering a suitable number of participants is not always possible due to strict inclusion criteria or rare target conditions. To circumvent this issue, various data augmentation techniques can be applied that generate artificial samples based on existing data [Mumuni and Mumuni, 2022]. Other alternatives include training models on larger datasets and fine-tuning the acquired knowledge with a smaller sample size (transfer learning) or using actors to mimic the conditions of participants from the target group, which is a common practice in emotion recognition datasets. Although these techniques typically reduce the model quality, they often produce better results than solely training on smaller collections of samples.

5.3.1 Annotation

In addition to recruiting suitable participants and capturing their signals with mobile sensors, another essential aspect concerns the acquisition of ground-truth annotations for the targeted conditions. This information is necessary for models to learn the association between input signals and target labels. Thereby, the quality of annotations can greatly influence the resulting recognition performance. Depending on the study procedure and collected data types, different annotation methods are available. One of the most reliable techniques is the automatic labeling of signals based on scripted events, algorithmic properties, and context information. For example, a specific condition could be automatically triggered during a predefined time window, which would also mark the start and end of the corresponding label [Plarre et al., 2011]. Since this process does not involve any human interference, it can be performed relatively fast, does not require many resources, and produces the most objective results. One of its most significant

disadvantages, though, is that it can only be applied under very specific circumstances that are usually only available in laboratory settings.

Alternatively, the collected data can also be labeled manually by human annotators. Since the ground truth can only be estimated in this case and largely depends on the subjective opinion of the annotators, a common practice involves employing multiple labelers and aggregating their assessments to reduce the effects of individual outliers. Depending on the knowledge required to correctly evaluate the recorded data, this process either has to be performed by multiple experts or can be distributed to regular individuals (e.g., using crowdsourcing services like Amazon's Mechanical Turk). In both cases, specialized annotation tools like Anvil [Kipp, 2014], ELAN [Wittenburg et al., 2006]), GTrace [Cowie et al., 2013], and NovA [Baur et al., 2013] are typically used to assign labels at specific timestamps to the collected signals. While reviewing large amounts of data can be labor-intensive and time-consuming, it is considered the gold standard to achieve accurate models when automatic annotations are not possible [Artstein and Poesio, 2008; Snow et al., 2008]. However, this traditional labeling approach heavily depends on the availability of human-comprehensible data (e.g., video or audio) that provides insights into the situative context of the recordings. Consequently, this method is limited to the annotation of phenomena that can be observed externally by reviewing the captured signals (e.g., gestures, facial expressions, or environmental conditions), which might also lead to annotations that do not correspond with the self-perception of the recorded participants.

In an effort to reduce the workload required for the manual annotation of large datasets, a cooperative machine-learning workflow can be employed. During this process, only a fraction of the total data has to be labeled by human annotators. The reviewed parts are then used to train models that automatically produce labels for the remaining signals. To ensure sufficient prediction quality, the generated annotations are evaluated and corrected by human supervisors. Afterwards, the initial model gets updated with the revised labels to produce more accurate annotations in subsequent iterations. This process can be repeated until sufficient quality and performance are achieved [Baur et al., 2020; Heimerl et al., 2022]. Although cooperative machine learning can alleviate most of the efforts associated with the manual labeling procedure, it still requires externally observable and human-comprehensible data for the initial set of annotations. In cases where such signals are not available, the only viable alternative is to involve participants directly in the annotation process. Popular examples of this approach are diary studies, where individuals are asked to report on specific aspects of their daily lives using diary entries [Carter and Mankoff, 2005].

A variation of such diary studies is the experience sampling method (ESM) [Larson and Csikszentmihalyi, 1983], which distinguishes itself by prompting participants to report on their experiences during the current activity instead of retrospectively reflecting on them in a diary. In this regard, proactively notifying and reminding participants to provide annotations also reduces their burden compared to reporting the data on their own accord [Chang et al., 2015]. Thereby, determining the best moment to query users and collect annotations can depend on various factors, including personal preferences, mental load, and current circumstances. In general, there are three types of notification strategies available: (1) *signal contingent*, in which participants report when prompted (usually at random times); (2) *interval contingent*, where annotations are collected at a regular (time-based) interval; and (3) *event contingent*, during which individuals report experience samples in response to certain events of interest [Barrett and Barrett, 2001; Wheeler and Reis, 1991]. Independent of the chosen strategy, the close temporal proximity between an experience that influences a participant's current state of mind and the annotation helps to avoid incorrect situational assessments caused by erroneous reconstructions of memories [van Berkel et al., 2018]. Additionally, an increased labeling frequency allows for a much more fine-grained - and therefore accurate - assessment of a person's state throughout the day. However, requesting frequent reports over an extended period can become cumbersome and might require incentives to maintain a steady annotation quality. Otherwise, users might lose interest in answering future requests after a certain time if it yields no benefits for them [Dietz et al., 2019].

5.3.2 Existing Datasets

An alternative to manually conducting experiments and recording signals is using existing datasets. Since collecting large amounts of data is often labor-intensive and can require significant time and financial investments for suitable sensing devices, study preparations, and participant compensation, existing datasets represent a considerably less resource-intensive option. Due to their almost immediate availability, they allow researchers to quickly evaluate novel hypotheses and accelerate the overall research process. These properties can be especially beneficial when the intended circumstances are difficult to replicate or require a long-term deployment of sensing devices. Furthermore, they enable comparisons of new approaches with existing methods and act as performance benchmarks to validate and assess potential advancements. For instance, the ImageNet dataset by Deng et al. [2009] has become a standard for comparing the accuracy of image classification and object recognition approaches and is typically used to showcase performance improvements in visual tasks over existing methods.

Dataset	Users	Environment	Signals					
			♡					
WESAD [Schmidt et al., 2018]	15	laboratory	✓	✓	○	✓	✓	✓
SWEET [Smets et al., 2018]	1002	in the wild	○	✓	○	✓	✓	✓
CLAS [Markova et al., 2019]	62	laboratory	✓	✓	○	✓	✓	○
ECSMP [Gao et al., 2021]	89	laboratory	✓	✓	✓	✓	✓	✓
DAPPER [Shui et al., 2021]	88	in the wild	✓	○	○	✓	✓	○
Emognition [Saganowski et al., 2022]	43	laboratory	✓	○	✓	✓	✓	✓
VerBIO [Yadav et al., 2022]	55	laboratory	✓	✓	○	✓	✓	✓
UBFC-Phys [Sabour et al., 2023]	56	laboratory	✓	○	○	✓	✓	✓

Legend: ♡ Heartrate Electrocardiogram Electroencephalogram Electrodermal Activity Acceleration Skin Temperature

Table 5.1: Examples of existing datasets recorded with wearable sensors.

Table 5.1 provides an overview of example datasets that were specifically recorded with wearable and mobile sensors. Despite not being limited by stationary capturing devices, most datasets listed were still collected in a laboratory environment, which highlights the difficulty of recording signals in the wild. Although corpora acquired under natural conditions are typically preferable, several characteristics and limitations should be considered when selecting an existing dataset. On the one hand, it has to be ensured that the domain and scope of the recorded signals match the intended scenario. Otherwise, potential findings might not translate to the conditions found during real-world usage [Pan and Yang, 2010]. This conclusion also applies to the targeted population, which can influence recognition performance and lead to varying results based on demographic factors like age, culture, and health condition. Consequently, assessing whether the dataset contains a representative population sample can facilitate the generalizability of potential approaches within the target user group.

On the other hand, the data quality should be examined beforehand regarding artifacts and distortions, which commonly occur in recordings with mobile and wearable sensors [Lane et al., 2010]. Since removing noise can require extensive computational resources, efficient algorithms might be necessary to prepare the data for further processing. In this regard, it should also be ensured that the selected hardware used to capture the dataset matches the quality and accuracy of the sensors intended for later use. More precisely, even though two devices are supposed to provide the same type of signal in theory, their outputs might deviate and show discrepancies in practice. Besides

examining the sensors and produced signals, the included annotations should also be reviewed. This step is essential to identify potential labeling inconsistencies or class imbalances that might impact performance later on. Once all of these aspects have been verified, the dataset can be considered to achieve effective solutions. However, depending on the intended scenario, some characteristics might be more relevant than others, which can enable the usage of datasets recorded in unrelated domains or with different sensors. For instance, even though the AffectNet [Mollahosseini et al., 2019] corpus consists of images with varying quality and resolution, it can still be utilized to train emotion recognition models that work with the camera sensors of mobile devices.

5.4 Model Training

After obtaining a suitable set of samples, either by collecting them yourself or using a publicly available dataset, the next step involves training a recognition model that analyzes the input signals and predicts the targeted conditions. To this end, two general directions are available: (1) traditional feature-based methods and (2) newer end-to-end deep learning approaches. As their name implies, traditional techniques typically rely on hand-crafted features engineered by domain experts to extract meaningful information from sensor data. In contrast, deep learning models can directly handle raw signals and automatically derive internal feature representations [Saganowski et al., 2023]. However, this process also requires much larger datasets, more computational resources, and longer training times. Additionally, the resulting models are often more complex and less interpretable "black boxes" compared to traditional algorithms. In turn, deep learning models can achieve better performance if their requirements are met and even enable new use cases that were previously impossible (e.g., large language models). To facilitate the decision between these methods, especially considering the circumstances of mobile applications, both directions are described in more detail below.

5.4.1 Feature-based Machine Learning

Traditional machine learning typically involves preprocessing the input data, extracting meaningful features, and training recognition models that map the calculated values to the desired labels. In the first step, the raw sensor data is prepared and cleaned up for further processing. This includes synchronizing the streams from different modalities, applying denoising filters to improve the signal quality, handling missing or incomplete information, and transforming the data into normalized ranges to ensure comparable feature scales [Schmidt et al., 2019]. Afterwards, the data streams are split into fixed

segments, which serve as the foundation for subsequent processing steps. In this regard, the selection of appropriate window lengths and overlaps depends on various aspects (e.g., classification task, signal type, or response time) and can significantly impact the resulting recognition performance (as shown in Section 8.2.5).

The next step involves calculating features based on the previously defined segments. This process is performed to reduce the problem dimensionality by extracting only meaningful information from the respective windows and plays a critical role in achieving effective models [Saganowski et al., 2023]. Overall, there are various feature sets available, but depending on the signal type, some are more appropriate than others due to the nature of the underlying modalities. For instance, time-domain features may be suitable for chronological events, while frequency-domain features are more effective for periodic or oscillating data. Other common categories include linear, non-linear, unimodal, and multimodal relationships. In addition to these different types, the computational complexity can also vary from general statistical measures (e.g., mean, min, max, variance, or standard deviation) to complex modality-specific calculations (e.g., gaze-based wordbooks containing saccade direction occurrences). Consequently, identifying the most suitable combination of features for a given problem represents an integral part of traditional machine learning.

Once all relevant features are extracted, the final step is to train the recognition models. During this process, the model parameters are adjusted to learn and establish a mapping between feature values as inputs and corresponding labels as outputs. Examples of commonly used algorithms include support vector machines (SVM), naïve Bayes (NB), decision tree (DT), and k-nearest neighbor (kNN) classifiers. While each of these models has its strengths and weaknesses depending on the problem characteristics, data composition, and desired performance, all of them are suitable for mobile real-time applications due to their lightweight computational requirements and the increasing hardware capabilities of modern devices.

5.4.2 End-to-End Deep Learning

Deep learning is a machine learning paradigm that utilizes multilayer artificial neural networks, which mimic the structure of the human brain and consist of various interconnected nodes (neurons) to automatically learn relationships between raw signals and target labels. Due to recent technological advancements and increased availability of large-scale datasets, deep learning has experienced a surge in popularity over the past decade and has been widely adopted to solve complex problems. Despite requiring

more computational resources and longer training times, the resulting performance improvements and enabled opportunities significantly outweigh potential disadvantages. Further information regarding the challenges and possibilities of deep learning for mobile mental health applications can be found in the overview by Han et al. [2021].

Similar to traditional methods, deep learning approaches typically also start with a preprocessing stage. Its goal is to clean up and prepare the raw signals to achieve satisfactory model performance [Gu et al., 2022]. Apart from synchronizing different data streams, removing unwanted artifacts, and handling missing or incomplete samples, normalizing numerical signals (e.g., between $[0, ..., 1]$ or $[-1, ..., 1]$) is an important step that can be the deciding factor between failed and successful training attempts [Goodfellow et al., 2016, p. 448]. It improves the rate at which models converge and prevents them from focusing on specific features solely due to their wide range of values. Otherwise, large inputs could lead to larger weights, while features with narrow ranges would be neglected. Although removing artifacts is typically another essential preprocessing operation in traditional feature-based approaches, adding noise is a common step during the training of deep learning models to improve their robustness. For example, popular methods of adding noise to images include cropping, distorting, resizing, and rotating the original samples to mimic potential interferences that could occur during a model's later usage. In addition to improving stability, the modified noise samples also increase the amount of data available for training. Similarly, using smaller window sizes during segmentation is another method that leads to a larger pool of samples.

Architecture	**ImageNet Top-1 Acc (%)**	**# Parameters (M)**
MobileNet [Howard et al., 2017]	63.7 – 70.6	1.3 – 4.2
ShuffleNet [Zhang et al., 2018]	71.5 – 73.7	3.4 – 5.4
ShuffleNetV2 [Ma et al., 2018]	69.4 – 74.9	2.3 – 7.4
MobileNetV2 [Sandler et al., 2018]	72.0 – 74.7	3.4 – 6.9
MnasNet [Tan et al., 2019]	75.2 – 76.7	3.9 – 5.2
FBNet [Wu et al., 2019]	73.0 – 74.9	4.3 – 5.5
EfficientNet [Tan and Le, 2019]	77.1 – 84.3	5.3 – 66
MobileNetV3 [Howard et al., 2019]	67.4 – 75.2	2.5 – 5.4
GhostNet [Han et al., 2020]	66.2 – 75.7	2.6 – 7.3
EfficientNetV2 [Tan and Le, 2021]	83.9 – 85.7	22 – 120

Table 5.2: Example CNN architectures for mobile applications.

After the preprocessing stage, the resulting signals can be used directly as inputs for model training, which is one of the most significant differences compared to traditional machine learning approaches. Instead of requiring manually engineered and extracted features, deep learning models can automatically identify underlying patterns and internally derive suitable mechanisms for their detection. Depending on the intended application, various architectures and building blocks are available that excel at different tasks. One of the most commonly used architectures for processing grid-based structures such as images are convolutional neural networks (CNNs). They consist of convolutional layers that apply filters (or kernels) to detect features like edges, shapes, and textures within the data. These operations are typically followed by pooling layers, which reduce the resulting outputs while retaining the most characteristic features by summarizing the values within specific regions (i.e., calculating the average or maximum). This combination of layers allows CNNs to identify spatial hierarchies and understand complex relations within grid-based structures, making them particularly well-suited for applications such as object detection [Krizhevsky et al., 2017], image segmentation [Long et al., 2015], and video analysis [Karpathy et al., 2014]. Although their execution can become computationally intensive, especially for high-resolution images, several optimized CNN architectures have been proposed to enable their usage on mobile devices. As shown in Table 5.2, these approaches still achieve relatively high accuracy scores despite only using between two and seven million parameters on average. While current state-of-the-art models can reach more than 91% accuracy on the ImageNet dataset [Yu et al., 2022], they also require several billion parameters and currently only run on powerful desktop or server hardware. However, most architectures designed for mobile devices can be scaled to achieve better recognition performance by increasing the input size and number of parameters according to the intended use case and available computational resources.

Apart from CNNs, recurrent neural networks (RNNs) are another commonly used family of models. They possess cyclic connections, which allow them to retain previous information and identify temporal dependencies within the data. While CNNs are most suitable for processing grid-based structures, RNNs excel at handling sequential signals such as speech or accelerometer readings [Gu et al., 2022]. Consequently, popular use cases for RNNs include sentiment analysis [Tang et al., 2015], machine translation [Sutskever et al., 2014], activity recognition [Guan and Plötz, 2017], and time-series forecasting [Hewamalage et al., 2021]. Another group of deep learning models are generative adversarial networks (GANs). They consist of two competing neural networks with opposing goals. While the *generator* aims to learn the properties of a given dataset

and tries to produce realistic samples, the *discriminator* seeks to recognize whether the provided data is authentic or artificial [Goodfellow et al., 2014]. During training, both components work against each other and try to maximize their individual goals, which results in models that are capable of generating realistic samples. Besides extending smaller datasets with additional data, typical applications of generative adversarial networks include image generation [Brock et al., 2019], style transfer [Zhu et al., 2017], and super-resolution [Ledig et al., 2017].

Another closely related type of generative models are autoencoders. Similar to GANs, they consist of two opposing networks that can be used to extract efficient representations from complex signals. While the *encoder* part is responsible for compressing and transforming the raw inputs into essential features (bottleneck), the *decoder* part tries to reconstruct the original inputs from the low-dimensional representation. Their training process involves minimizing the difference between the original data and the reconstructed outputs, which incentivizes the encoder to identify and extract only the most relevant characteristics. Common applications of autoencoders include anomaly detection [Sakurada and Yairi, 2014], denoising [Vincent et al., 2008], and feature extraction [Li et al., 2014]. Lastly, the newest deep learning model type is the transformer architecture. Instead of relying on recurrence or convolutions, it combines an encoder-decoder structure with self-attention mechanisms to process input sequences in parallel [Vaswani et al., 2017]. This allows it to capture complex dependencies more efficiently and results in superior performance compared to previous approaches. Due to the significant improvements of models like GPT-3 [Brown et al., 2020], PaLM [Chowdhery et al., 2023], and LLaMA [Touvron et al., 2023], transformers have become the standard architecture for natural language processing applications.

5.5 Summary

Mobile signal processing plays an essential role in assistive augmentation approaches, enabling them to analyze, interpret, and respond to behavioral, physiological, and environmental signals. In order to better understand these processes, this chapter provided a detailed overview of the typical challenges, solutions, and procedures. One major challenge that needs to be addressed are privacy concerns. Due to the sensitive nature of the collected signals, appropriate security mechanisms must be implemented to keep individuals in control of their data. Otherwise, they might lose trust in potential solutions, which could lead to reduced acceptance and even abandonment of augmentation systems despite their beneficial effects. In this regard, developing transparent and compre-

hensible applications without cultural or demographic biases is another essential aspect that could impact these factors. Moreover, technical limitations such as computational power, memory, storage, battery life, and physical dimensions must also be addressed to achieve effective approaches. Besides outlining these challenges, this chapter also discussed common stages of mobile signal processing systems and provided details regarding available methods and procedures. The first step typically involves selecting appropriate sensing and processing hardware to capture relevant signals. Following that, the sensor data can be processed using simple algorithmic calculations, feature-based machine learning, or end-to-end deep learning. While this chapter focused on conceptual and theoretical details of these different approaches, their practical application is demonstrated with research probes in Part III.

Chapter 6

The SSJ Framework

In this chapter, we introduce the open-source SSJ[1] software framework for building and prototyping assistive augmentation systems. While the foundation was initially conceived by Damian [2017] with the intention to support social augmentation (see Section 4.1), the project's scope was extended early on through extensive collaboration to also incorporate the necessary capabilities and requirements for augmenting cognitive processes. This especially applies to the framework's modular architecture, which was inspired by the shared technical structure of previous approaches identified in Section 4.2. As shown in Figure 6.1, SSJ consists of reusable and easily exchangeable components that fulfill the roles of each commonly employed augmentation step (compare Figure 4.6). By following the general system structure of established approaches and combining the respective components into so-called processing pipelines, the framework provides all necessary tools to replicate, adapt, and extend the augmentation strategies described in Section 4.3.

Additionally, it supports all variations of design dimensions derived from the analyzed approaches in Section 4.4. For instance, SSJ contains various sensing components that can be utilized to analyze both users and their environments. Depending on the targeted cognitive processes, the captured data can be used to make otherwise not perceivable information accessible to people (sensory augmentation), remind them about forgotten details (memory augmentation), or recognize and assist specific cognitive conditions (cognitive augmentation). For that, SSJ supports the synchronized real-time process-

[1] The name SSJ was originally an abbreviation for *"Social Signal Processing for Java"* and dates back to its origins when the intention was to develop a Java version of the Social Signal Interpretation (SSI) framework by Wagner et al. [2013]. Although the project has matured beyond its initial scope, the original name was maintained to indicate the strong conceptual and technical connection to SSI.

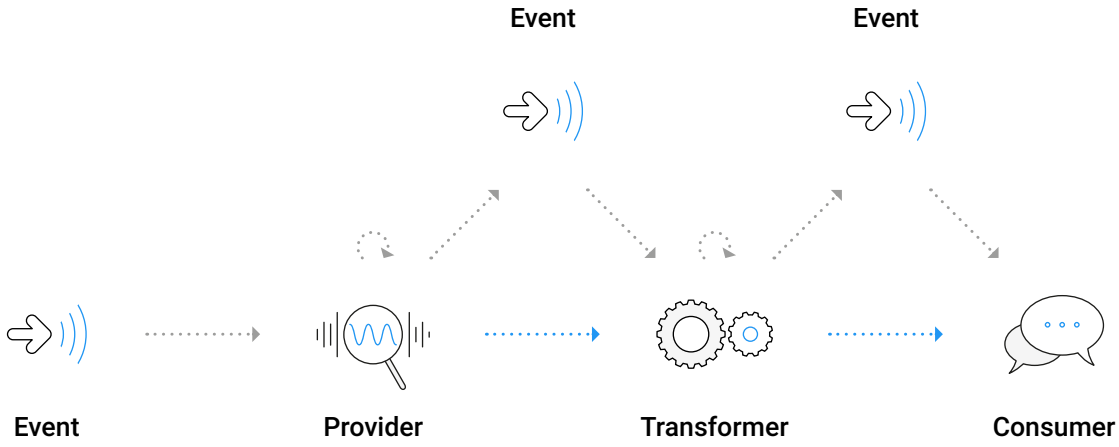


Figure 6.1: General components of the SSJ framework.

ing of signals on mobile devices, including advanced classification techniques, such as neural networks. Based on the results, assistive augmentation systems built with SSJ can initiate specific actions in response, provide appropriate feedback to their users, or dynamically adjust their behavior to the current circumstances. Furthermore, the framework's modular architecture enables all components to be rearranged, reused, repurposed, and replaced with minimal effort, which leads to shorter iteration times during prototyping and development. Overall, SSJ supports a wide variety of sensing hardware by default and can be easily extended with additional sensors. This flexibility also applies to the integrated filtering and feature extraction algorithms, which can process the captured signals directly on people's mobile devices. Combined with the latest machine learning techniques, the processed data can be classified in real-time, enabling augmentation systems to provide immediate responses and appropriate assistance through various output modalities (e.g., visual, auditory, or tactile).

In the following sections, we first introduce the basic concepts and core design principles that serve as the framework's foundation. We then provide an overview of existing mobile signal processing solutions, highlight the differences between approaches, and outline potential limitations we address with SSJ. Following that, we describe the framework's architecture and core components in more detail and demonstrate how they can be extended. Afterwards, we introduce the SSJ Creator application, which enables people without technical background or programming knowledge to rapidly build and prototype processing pipelines through a graphical interface. Finally, we showcase how the framework's capabilities can be used to implement approaches for the assistive augmentation of cognitive processes with a simple and easily understandable example.

Parts of this chapter are based on the following publications:

Reference Damian, I., Dietz, M., Gaibler, F., and André, E. (2016). Social Signal Processing for Dummies. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 394-–395. ACM.

Reference Damian, I., Dietz, M., and André, E. (2018). The SSJ Framework: Augmenting Social Interactions Using Mobile Signal Processing and Live Feedback. *Frontiers in ICT*, 5.

Reference Dietz, M., Aslan, I., Schiller, D., Flutura, S., Steinert, A., Klebbe, R., and André, E. (2019). Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 149–158. ACM.

6.1 Origins and Basic Concepts

SSJ is closely related to the Social Signal Interpretation (SSI) framework by Wagner et al. [2013], which enables the recording, analysis, and fusion of social signals in real-time. Since both projects were developed at the chair for Human-Centered Artificial Intelligence, they are fully compatible with each other and have several similarities, including a common data format and communication protocol. However, one significant difference is that SSI was primarily designed to run on stationary computers requiring the Windows operating system and can not be deployed on modern mobile devices. Due to the restrictions associated with adapting the code base of SSI to a mobile operating system (i.e., limited access to the full range of native platform capabilities, such as energy-saving mechanisms or Bluetooth connectivity), we instead built the SSJ framework from the ground up for portable devices. This decision enabled us to utilize the platform-specific features of the mobile ecosystem while also considering the requirements of assistive augmentation approaches and incorporating the insights gained from using SSI on desktop computers. Consequently, SSJ shares several core design principles and concepts with SSI, which are explained in the following sections.

6.1.1 Modular Design

Modularity is the core design principle behind the SSJ framework, which facilitates flexibility, scalability, and maintainability during the development of assistive augmentation approaches. It is based on the idea that complex systems and processes can be constructed from smaller units, each responsible for a specific function, such as filtering, feature extraction, or classification. By following clearly defined input and output

interfaces between components, each module can be developed, tested, and optimized independently, which not only streamlines the development process but also enhances the reliability and performance of the overall system. Furthermore, the principle of encapsulating functionality ensures that changes within a component do not affect other parts of the application or lead to unintended side effects. Due to the standardized external interfaces, this characteristic also allows developers to update the implementation of modules at any point in time as long as the designated functionality remains unchanged. For instance, if a more efficient algorithm becomes available, the back-end code of the respective component can be replaced and directly deployed to immediately benefit from the improvements without requiring any further changes. In order to handle the communication between these independent modules, SSJ uses a construct called *processing pipeline*. A pipeline consists of two or more components and defines how information is exchanged among them. Based on the standardized interfaces, each module can receive the outputs from multiple components and can provide its results as input to one or more subsequent modules. This structure enables the creation of complex acyclic processing chains with branches and junctions where data continuously flows through the specified sequence of components.

6.1.2 Sampling

Real-world signals, such as sounds, lights, or vibrations, are inherently continuous. They can be thought of as functions with infinite values that change smoothly across time and amplitude. Since capturing and processing these continuous phenomena, also known as *analog signals*, with digital systems would require an infinite amount of storage space, it is necessary to reduce their resolution and convert them into a discrete representation. This conversion process, called *sampling*, involves observing the analog signal at fixed time intervals to extract the closest discrete value at the targeted amplitude resolution, as shown in Figure 6.2. The frequency at which these data points are measured is the *sampling rate*, typically expressed in samples per second or hertz (Hz). While the sampling process always results in a loss of information, ensuring that the sampling rate is sufficiently high enough to accurately represent the continuous signal is still an essential aspect when capturing sensor data.

According to the *Nyquist–Shannon sampling theorem*, analog signals can be perfectly reconstructed if the sampling rate is greater than twice the maximum frequency component present in the original signal [Shannon, 1949]. In practical terms, this relation means that sampling a 4 Hz signal at a rate of 8 Hz or lower can lead to distortions and

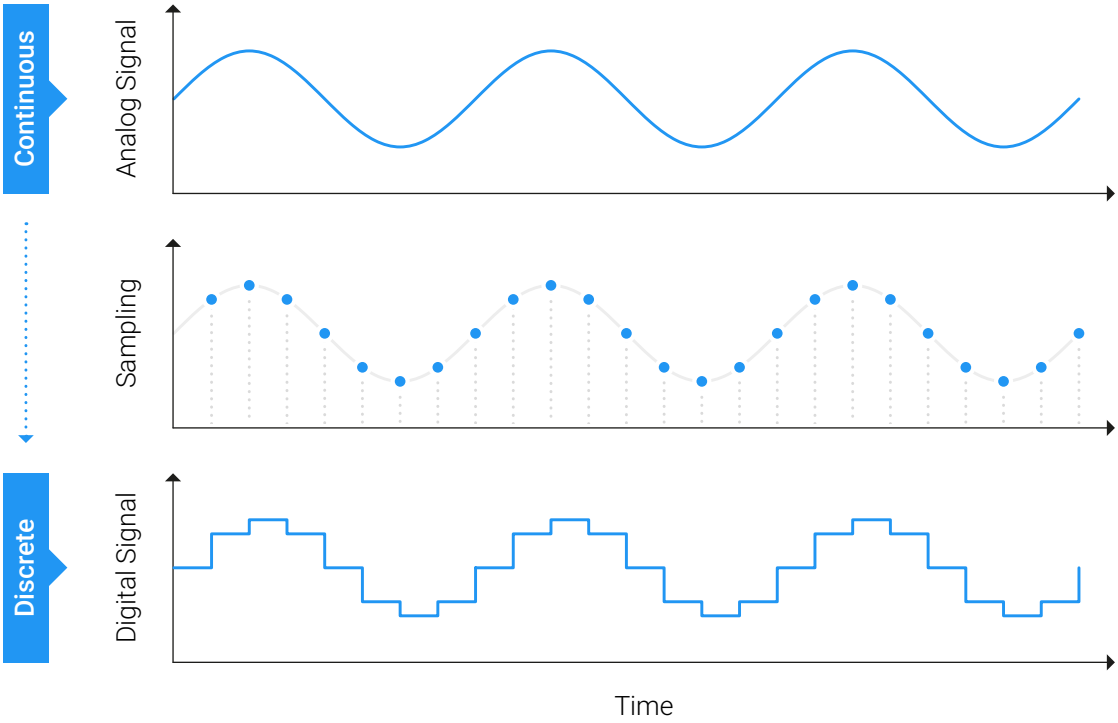


Figure 6.2: Sampling process of continuous to discrete signals.

inaccurate representations over time. However, reducing the sampling rate due to energy or resource limitations might be beneficial in some cases (e.g., to increase battery life or prevent overheating). These circumstances especially apply if the current sample value is sufficient for further processing and is not needed as often as it could be provided. To enable all of these application-specific scenarios, SSJ supports adjusting the sampling rate of every integrated sensor. This functionality is achieved by applying the sampling process to the discrete sensor data before providing it to other components.

6.1.3 Generic Data Handling

Modern sensing hardware typically uses various data types and structures to capture and accurately represent the wide variety of natural signals. Furthermore, the rate at which new samples are provided can differ significantly depending on the type of sensor. For instance, a sound wave is usually represented by thousands of individual float values per second that quantify the current amplitude. In contrast, a basic video signal only consists of around 30-60 samples per second, but each one can contain millions of bytes that define the color value for every pixel of the respective frame. For these reasons, a generic solution is required that can be applied to handle the wide range of different data types and sample rates. SSJ addresses this challenge by splitting the signals into

smaller parts with specific durations and storing the samples captured during these time windows in a universal data structure called *stream*. A stream can be visualized as a table where each row represents one sample that consists of the values contained within each column (*dimension*, see Figure 6.3). Internally, the data is stored in a generic byte array and can be converted to the actual type of the respective signal on demand.

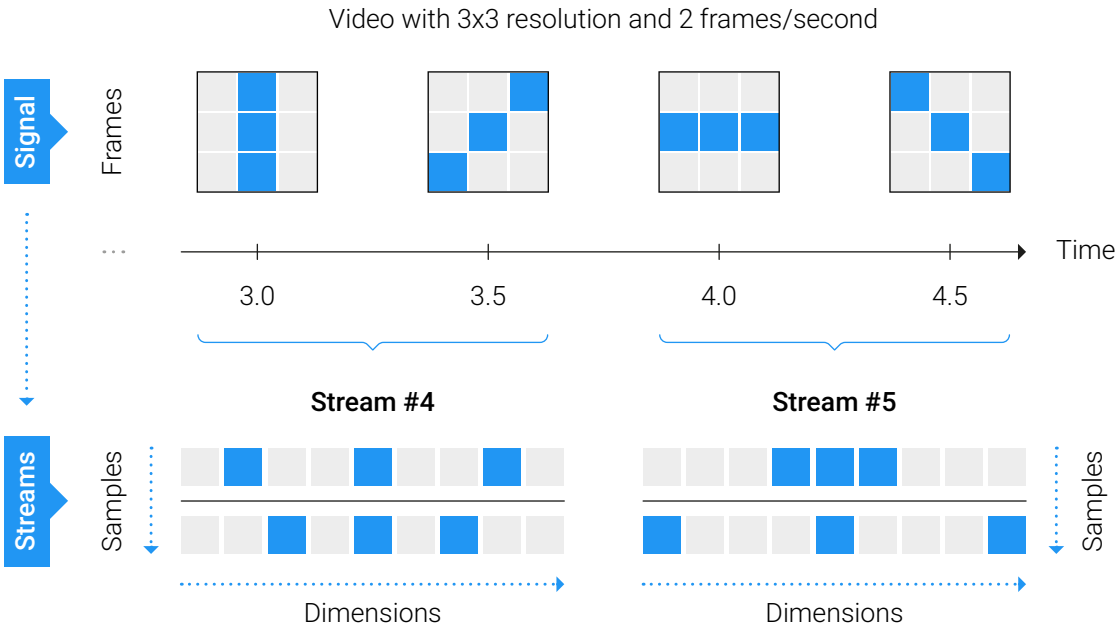


Figure 6.3: Mapping of signals to stream packages.

This universal structure provides the foundation to handle any kind of data produced by sensors and processing components. For instance, a 48 kHz signal from a microphone sensor can be represented by emitting a stream with two dimensions (stereo signal) and 48,000 samples every second, resulting in data packages with 96,000 float values each. The same signal could also be provided in streams containing 24,000 samples (48,000 float values) twice per second. On the other hand, a camera signal with a resolution of 1920×1080 pixels captured at 30 Hz could be represented by producing streams with 30 samples and 6,220,800 dimensions ($1920 \times 1080 \times 3$; RGB color values are stored separately) every second. Furthermore, this generic approach allows components to perform tasks independently at their individual frequency. For example, the update rate of a multimodal (audio & video) feature extractor could be set to 0.2 Hz ($\frac{1}{5\,\text{s}}$), at which point it would receive a video stream with 150 samples (30 Hz $\times$ 5 s) and an audio stream with 240,000 samples (48 kHz $\times$ 5 s) every five seconds.

6.1.4 Synchronization

Processing multimodal data generally involves analyzing and comparing parts of each observed signal that occurred at the same time. While the generic data handling approach described in Section 6.1.3 provides the foundation to enable such comparisons, it does not guarantee that the values contained within each stream package were recorded simultaneously on its own. Since the internal hardware clocks of sensors might not always be perfectly accurate, the rate at which new samples are provided can drift over time and lead to incorrect results. For example, if a sensor has a theoretical sampling rate of 10 Hz but suddenly starts to emit 11 samples per second, the resulting streams would contain 600 more samples than intended after only 10 minutes, which means the signal would be already offset by one minute compared to accurately sampled streams. For this reason, a synchronization mechanism is required to ensure that every sensor and processing component adheres to its specified sampling rate.

SSJ solves this problem by calculating the number of expected samples within a given time window and monitoring the actual output of every component. In case of discrepancies, synchronization is maintained by removing redundant values or inserting missing samples. For that, several strategies can be applied, including repeating the previous sample or filling the gaps with default values (e.g., zeros). While skipping or duplicating parts of the signal can affect the data quality, these measures are rarely required for properly working sensors. Additionally, another benefit of this approach is that the same mechanisms can be applied to handle sensor failures and lost connections. In these cases, SSJ fills the data streams with default values to maintain the specified sampling rates and synchronization with other components until the problem is resolved.

6.2 Existing Solutions

After introducing the basic concepts of SSJ, this section provides an overview of existing mobile signal processing frameworks and highlights the differences between proposed solutions. The detailed properties and capabilities of each discussed approach are listed in Table 6.1. In order to enable a valid and objective comparison, we only include systems that can perform some form of signal processing on mobile devices, which is essential for developing assistive augmentation applications. For this reason, frameworks with a primary focus on data collection, such as Open Data Kit (ODK) [Brunette et al., 2012], Mobile Sensing Framework (MSF) [Cardone et al., 2013], DataLogger [Ciliberto et al., 2017], UniMiB AAL [Ginelli et al., 2018], MyExperience [Froehlich

et al., 2007], Funf[2], and IoTool[3] are not considered. Similarly, tools that can process sensor data in real-time but do not work on mobile devices like Multisensor-Pipeline [Barz et al., 2021], Microsoft's Platform for Situated Intelligence (PSI) [Bohus et al., 2021], and OpenSense [Stefanov et al., 2020] are also excluded.

One of the first approaches for mobile devices is the CenceMe application by Miluzzo et al. [2008]. It runs on Nokia's Symbian operating system and supports capturing accelerometer, camera, microphone, and GPS signals. Additionally, it can classify the resulting data into different activities (i.e., sitting, standing, walking, or running) and sound contexts (i.e., conversation, silence, or loud environment). As shown in Table 6.1, several researchers adopted and extended this concept. For instance, Wang et al. [2009] proposed the Energy Efficient Mobile Sensing System (EEMSS), which targets the same platform and sensors (except the camera) but reduces their power consumption with dynamic processing cycles. Rachuri et al. [2010, 2011] also used a similar foundation and included the ability to recognize people's emotions (EmotionSense) and sociability with others (SociableSense). Around the same time, the group behind the CenceMe application introduced the Jigsaw continuous sensing engine [Lu et al., 2010], which runs on Apple's iOS in addition to Nokia's Symbian operating system and consists of robust processing pipelines for accelerometer, microphone, and GPS data. One solution that differs from the previous frameworks is the Auditeur platform by Nirjon et al. [2013]. It runs on Google's Android operating system and exclusively focuses on processing and classifying audio signals with local and cloud-based methods. However, all of these approaches are closed-source and not publicly available, making them unsuitable for developing open and widely accessible assistive augmentation systems that can be adapted to support people's individual needs and requirements.

In contrast, one of the first publicly available open-source toolkits is the BeTelGeuse platform by Kukkonen et al. [2009]. Despite requiring devices with support for the discontinued Mobile Information Device Profile (MIDP) API, it offers an extensible architecture that can collect and process signals from various internal and Bluetooth-connected sensors, including location, motion, and physiological data. Another open-source approach that instead targets the widely used Android operating system is the FieldStream framework by Ertin et al. [2011]. It is specifically designed to capture measurements from their custom physiological sensing device (AutoSense), supports calculating features with sophisticated signal processing methods, and can infer behav-

[2] `https://funf.org`

[3] `https://iotool.io`

Framework	Access			Usage		Platform			Sensors					Processing			Outputs			
CenceMe [Miluzzo et al., 2008]	○	○	○	●	○	○	○	●	●	●	○	●	●	○	●	●	●	○	○	●
BeTelGeuse [Kukkonen et al., 2009]	●	●	○	●	○	○	○	●	○	○	●	●	●	●	●	●	○	○	○	●
EEMSS [Wang et al., 2009]	○	○	○	●	○	○	○	●	○	●	○	●	●	○	●	●	●	○	○	●
Jigsaw [Lu et al., 2010]	○	○	○	●	○	○	●	●	○	●	○	●	●	○	●	●	○	○	○	●
EmotionSense [Rachuri et al., 2010]	○	○	○	●	○	○	○	●	○	●	○	●	●	○	●	●	○	○	○	●
FieldStream [Ertin et al., 2011]	●	●	○	●	○	●	○	○	○	○	●	○	●	●	●	●	○	○	○	●
SociableSense [Rachuri et al., 2011]	○	○	○	●	○	○	○	●	○	●	○	●	●	○	●	●	○	○	○	●
Dynamix [Carlson and Schrader, 2012]	●	●	○	●	○	●	○	○	●	●	●	●	●	○	○	○	●	○	○	●
Auditeur [Nirjon et al., 2013]	○	○	○	●	○	●	○	○	○	●	○	○	○	●	●	●	○	○	○	●
CRNTC+ [Spina et al., 2013]	○	●	○	●	●	●	○	○	○	○	●	●	●	●	●	●	●	○	○	●
AWARE [Ferreira et al., 2015]	●	●	○	●	○	●	●	○	○	●	●	●	●	○	○	○	●	●	○	●
mHealthDroid [Banos et al., 2015]	●	●	○	●	○	●	○	○	○	○	●	○	●	●	●	●	●	○	○	●
mobileSSI [Flutura et al., 2016]	●	●	○	●	○	●	○	●	●	●	●	○	●	●	●	●	●	●	○	●
mCerebrum [Hossain et al., 2017]	●	●	○	●	○	●	○	○	●	●	●	●	●	○	●	●	●	○	○	●
MediaPipe [Lugaresi et al., 2019]	●	●	●	●	○	●	●	●	●	●	○	○	○	●	●	●	●	○	○	●
SSJ [Damian et al., 2016, 2018]	●	●	●	●	●	●	○	○	●	●	●	●	●	●	●	●	●	●	●	●

Legend: Available · Open Source · Maintained · Code · GUI · Android · iOS · Other OS · Camera · Microphone · Physiological · Location · IMU · Filtering · Feature Extraction · Classification · Visual · Auditory · Tactile · Digital

Table 6.1: Existing frameworks for mobile signal processing.

ioral states from the resulting data. Similarly, the mHealthDroid framework by Banos et al. [2015] also focuses on collecting and processing physiological signals with wearable sensors. It facilitates the development of mobile health applications by providing various functionalities, including data acquisition, knowledge extraction, persistent storage, and visualization of measurements. Unfortunately, both FieldStream and mHealthDroid are primarily concerned with physiological signals and do not support processing camera, microphone, or location data.

One approach that does support various sensors is the Dynamix framework by Carlson and Schrader [2012]. It runs as a background service on Android devices and offers sensing information to other applications through numerous plugins that can be installed and updated during runtime. Despite its versatility, the flexible plugin-based architecture only provides basic sensing capabilities and does not support advanced processing techniques like filtering, feature extraction, or classification. This limitation also applies to the AWARE toolkit by Ferreira et al. [2015], which primarily focuses on collecting, inferring, and generating context information on mobile devices. Hossain et al. [2017] addressed these shortcomings and proposed the mCerebrum platform a few years later. It supports various types of sensors, can perform advanced data processing methods, and handles high-frequency signals more efficiently than AWARE and prior approaches. Another toolkit with similar capabilities is mobileSSI by Flutura et al. [2016], which is a mobile port of Wagner et al.'s [2013] Social Signal Interpretation (SSI) framework. While its universal C++ core enables very efficient data processing on Android tablets and mobile phones, it also prevents access to native platform features, including energy-saving mechanisms and Bluetooth-based communication with other devices. Additionally, it requires relatively advanced programming knowledge to fully utilize its capabilities and build effective solutions.

In contrast, Spina et al. [2013] tried to reduce the generally high entry barrier of signal processing approaches for end users and proposed the CRNTC+ framework. It is an extension of the Context Recognition Network (CRN) Toolbox by Bannach et al. [2008] and provides a graphical interface to visually configure component options and specify the data flow between them. Unfortunately, none of the previously introduced approaches are still being actively developed and maintained, which restricts their potential usage in future applications. Considering the rapid advancements within the mobile and wearable sector, abandoned software quickly becomes incompatible with new devices and updated operating systems. Furthermore, the lack of support for modern sensors and peripherals limits the usefulness of these outdated solutions. To the best of our knowledge, the only other active open-source project for mobile signal processing

besides SSJ is Google's MediaPipe framework [Lugaresi et al., 2019]. Its flexible architecture enables the cross-platform development of efficient processing applications and combines the advantages of most previous approaches. However, it primarily focuses on video and audio signals and does not support other sensors by default. Additionally, it only offers tools to visualize existing processing graphs but does not provide a user-friendly application or interface to create them.

In conclusion, although there are several mobile signal processing solutions, they require discontinued platforms, are not publicly available, only support specific sensors, can not perform advanced processing techniques, or are not actively developed anymore. The majority of these shortcomings can be attributed to the circumstances that led to the creation of the respective solutions. Usually, they were designed to solve a specific problem and only implement the necessary components for this purpose but do not consider the challenges of more universal scenarios. Additionally, most approaches either require advanced programming knowledge or an in-depth understanding of the underlying technology to develop appropriate applications. To address these limitations, we proposed the flexible and easily usable open-source SSJ framework for performing signal processing on mobile devices.

6.3 Architecture

The common technical structure of previous augmentation approaches identified in Section 4.2.2 served as the primary foundation for the architecture of the SSJ framework. To achieve the same functionalities as these existing systems and support future developments beyond that with a universal approach, we followed the same overall structure and implemented each augmentation step as an independent, exchangeable, and reusable component. Additionally, the strong conceptual connection to the SSI framework influenced several design decisions and resulted in various similarities across the proposed solution. For the target platform, we selected the open-source Android operating system. Due to its openness and customizability, it quickly became the most widely used platform for mobile devices. At the time of SSJ's development, the mobile operating system powered more than 2.1 billion devices worldwide and accounted for roughly 80% of all smartphone sales[4]. Moreover, it supports various other types of portable devices, including tablets, smartwatches, and smart glasses, which makes it an ideal platform for assistive augmentation approaches.

[4] `https://statista.com/statistics/385001/smartphone-worldwide-installed-base-operating-systems`

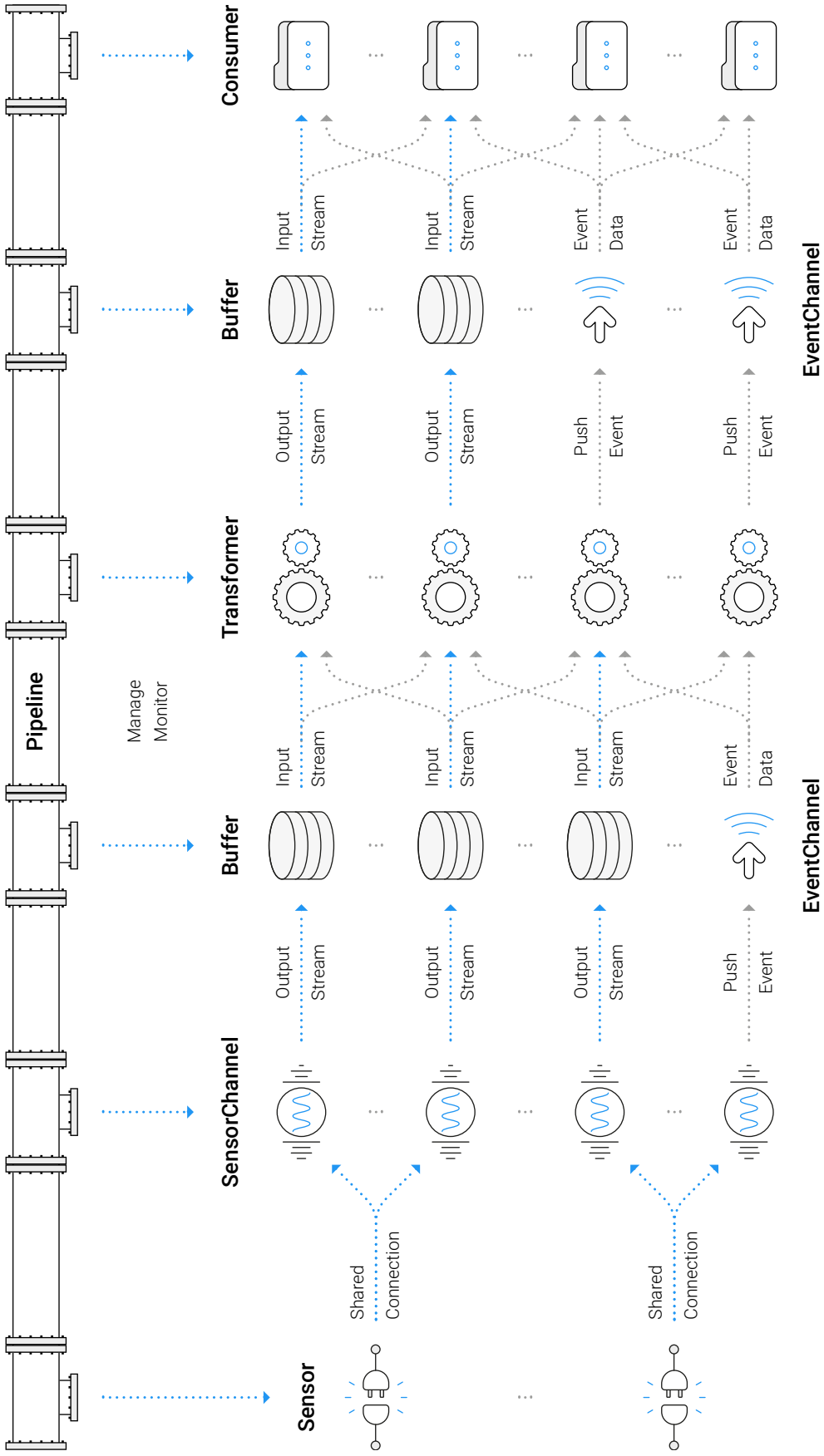


Figure 6.4: Architecture of the SSJ framework.

Figure 6.4 provides an overview of the general system architecture and illustrates the relations between involved components. Overall, the core construct of the framework is the `Pipeline` class. It represents a set of processing units and is responsible for managing the flow of information between them. To this end, it provides several methods for controlling the execution of the pipeline, adding different types of components to the chain of active elements, and exchanging data between connected components. Each `Component` represents an independent processing step and can be categorized into one of three basic types: (1) *providers*, that only output data to other components; (2) *transformers*, that receive data and output processed results; and (3) *consumers*, that only receive data as inputs. Since physical sensing devices typically include more than one type of measurement, we split the responsibilities of such providers across a `Sensor` and a `SensorChannel` component. While the sensor handles the connection to the external device, each sensor channel provides one of the signals captured with the associated sensing hardware. For instance, the framework includes a `Polar` sensor that handles the Bluetooth connection to the Polar H10[5] chest strap and multiple sensor channels that provide the transmitted heart rate (`PolarHRChannel`), acceleration (`PolarACCChannel`), or electrocardiogram (`PolarECGChannel`) data. This structure allows designers and developers to use only the desired signals from a given sensor without having to process all provided measurements. Additionally, since every channel is always connected to a specific sensor, it enables them to capture data from multiple sensing devices with different output configurations.

After adding an output-capable component (i.e., `Provider`, `SensorChannel`, or `Transformer`) to a pipeline, it internally creates a `TimeBuffer` instance to handle the generated data. From there, all subsequent components can access the stored information in parallel and perform their processing steps independently. As a result, each buffer connects exactly one data source with one or multiple information sinks. During runtime, packages of samples with flexible window lengths specified individually by each connected component are extracted from the buffer and provided to the respective units as input *streams* (see Section 6.1.3). Once an output-capable component finishes its current processing cycle, the resulting data is stored in an output `Stream` container and passed to the subsequent buffer, from where connected components can access it. Throughout this process, the framework monitors the inputs and outputs of every processing unit and ensures that they maintain the specified sampling rate to keep the signals synchronized (see Section 6.1.4). Apart from the synchronous exchange of information, every component in SSJ can also communicate asynchronously through *events*. For that, the

[5] `https://polar.com/en/sensors/h10-heart-rate-sensor`

abstract `Component` class offers public methods to access its output `EventChannel` and to register other components' channels as input. Each channel acts as a queue where events are processed on a first-in-first-out (FIFO) basis. Once an event is pushed to a channel, every registered listener gets notified and receives the transmitted information. Following this general overview, more details regarding the core components of the framework and possibilities for expansion are provided below.

6.3.1 Providers

Every processing unit that supplies data to other components within a pipeline but does not receive information from them is considered a `Provider`. Thereby, it does not matter where the signals come from or how they are acquired as long as they are made available in the correct output format. Due to the framework's modular architecture, the implementation-specific details for that are encapsulated within the respective component and do not affect other processing units. The most common type of providers are external sensing devices. As mentioned in Section 6.3, they typically offer multiple measurements, which is why the framework uses a `Sensor` component to handle the shared connection and one `SensorChannel` for each provided signal. To integrate a sensor into the framework, it is only required to inherit the abstract `Sensor` class (line 2) and implement the given methods for connecting (lines 7-12) and disconnecting (lines 15-18) the data source. For sensing hardware that provides a software development kit (SDK), this process typically involves executing specific functions according to their documentation. However, connecting to other data sources, including Bluetooth signals, network sockets, or file systems, is also possible.

```
1  // Extend abstract Sensor class
2  public class MySensor extends Sensor
3  {
4      [...] // Define potential component options, see Appendix A
5
6      @Override
7      public boolean connect() throws SSJFatalException
8      {
9          // Implement sensor connection, return true if established
10
11         return false;
12     }
13
14     @Override
15     public void disconnect() throws SSJFatalException
16     {
```

```
17          // Implement sensor disconnection
18      }
19  }
```

Each type of signal acquired from the respective data source is then processed in a separate component that inherits the abstract `SensorChannel` class (line 21) and has access to the corresponding `Sensor` instance through an automatically populated reference set by the framework during initialization (`_sensor` variable in line 30). This reference can be used to acquire the signals from the shared sensor connection in the predefined `process()` method, which continuously provides the data to other framework components based on the channel's sampling rate. Apart from that, the optional `enter()` and `flush()` methods can be implemented to perform specific actions immediately before and after the recurring processing loop (e.g., for setup or cleanup purposes).

```
20  // Extend abstract SensorChannel class
21  public class MyChannel extends SensorChannel
22  {
23      [...] // Define potential component options, see Appendix A
24
25      @Override
26      public boolean process(Stream stream_out) throws SSJFatalException
27      {
28          // Implement data output, return true if successful
29          float[] out = stream_out.ptrF();
30          out[0] = ((MySensor) _sensor).getData();
31
32          return true;
33      }
```

The last step involves specifying the channel output, which includes its sampling rate, number of dimensions, data type, and provided sample count for each `process()` call (lines 35-44). In case the `getSampleNumber()` method is not overwritten, the framework assumes that only one sample is returned after each processing cycle by default. As shown in line 35, it is also possible to expose these properties as dynamically adjustable options, thus increasing the channel's flexibility and areas of application. Additionally, the different sample dimensions can be described more precisely with a string array, which primarily informs developers about the provided sample values but is also used to automatically label the signals in dynamic graphs and plots (lines 47-51).

```
34      @Override
35      public double getSampleRate() { return options.sr.get(); }
36
```

```
37      @Override
38      public int getSampleDimension() { return 1; }
39
40      @Override
41      public int getSampleNumber() { return 1; }
42
43      @Override
44      public Cons.Type getSampleType() { return Cons.Type.FLOAT; }
45
46      @Override
47      public void describeOutput(Stream stream_out)
48      {
49          stream_out.desc = new String[stream_out.dim];
50          stream_out.desc[0] = "Dimension description";
51      }
52  }
```

Although the separation between sensors and channels was primarily intended to handle sources with multiple signals, we also use it for components with only one data type due to the increased flexibility in case another signal becomes available. In order to add sensors and channels to a pipeline, we simply create the respective instances (lines 53-54), set potential options (line 55), and call the pipeline's `addSensor()` method (line 57). All further steps are managed by the framework.

```
53  MySensor sensor = new MySensor();
54  MyChannel channel = new MyChannel();
55  channel.options.sr.set(10);
56
57  pipeline.addSensor(sensor, channel);
```

6.3.2 Transformers

Transformers are relatively similar to providers. The major difference is that they also receive information from prior processing units in addition to supplying data to subsequent components. While providers are always positioned at the start of a pipeline, transformers typically occur throughout the middle and always require a preceding data source. Since the inputs can either originate from a provider or another transformer, it is possible to create pipelines with a sensing device at the beginning that provides the signals to a chain of transformers for further processing. However, cyclic connections between components are not supported and result in errors during initialization. To integrate a new transformer into the framework, the abstract `Transformer` class needs to be inherited, and an implementation for the `transform()` method must be provided.

Its primary purpose is to convert the data from one or more input components into corresponding output `Stream` packages at a specified update rate. As with providers, the optional `enter()` and `flush()` methods can be implemented to perform specific actions immediately before and after the recurring processing loop.

```
1  // Extend abstract Transformer class
2  public class MyTransformer extends Transformer
3  {
4       [...] // Define potential component options, see Appendix A
5
6       @Override
7       public void transform(Stream[] stream_in, Stream stream_out)
8       throws SSJFatalException
9       {
10           // Implement data processing
11      }
```

Following that, the number of dimensions, data type, and sample count of produced `Stream` packages after each `transform()` call are specified (lines 13-22). Except for the sampling rate, these methods are identical to those of providers. The reason why this property is not defined inside transformers is that it can be dynamically adjusted when adding the respective component to a pipeline. Consequently, every `Transformer` can be used with different frequencies and input stream durations. Apart from that, the sample dimensions can be described in the same way as for providers to inform developers about the supplied values and transformation results (lines 25-29).

```
12      @Override
13      public int getSampleDimension(Stream[] stream_in) { return 1; }
14
15      @Override
16      public int getSampleNumber(int sampleNumber_in) { return 1; }
17
18      @Override
19      public Cons.Type getSampleType(Stream[] stream_in)
20      {
21           return Cons.Type.FLOAT;
22      }
23
24      @Override
25      public void describeOutput(Stream[] stream_in, Stream stream_out)
26      {
27           stream_out.desc = new String[stream_out.dim];
28           stream_out.desc[0] = "Dimension description";
29      }
30 }
```

Adding transformers to a pipeline is also similar to providers. First, a component instance is created, and potential options are set (lines 31-32). Afterwards, the pipeline's `addTransformer()` method is called, which expects references to the transformer itself and one or more source components that provide the input data for it (line 38). During this method call, the optional `frame` and `delta` durations can be specified (lines 35-38). While the sum of both values determines the window length of each received `Stream` package, the `frame` duration defines the interval between subsequent `transform()` calls. In the listing below (`frame=1`, `delta=4`), the method would be executed every second with the data from a five-second sliding window.

```
31 MyTransformer transformer = new MyTransformer();
32 transformer.options.myOption.set(true);
33
34 // Specify duration in seconds
35 float frame = 1;
36 float delta = 4;
37
38 pipeline.addTransformer(transformer, channel, frame, delta);
```

This example demonstrates how the framework calculates the update rate of transformers based on the `frame` duration. Consequently, specifying `frame=0.2` would result in an update frequency of 5 Hz. In case both values are not provided, the framework uses the same `frame` duration as previous components and sets the `delta` time to zero.

6.3.3 Consumers

Within the framework, providers and consumers have complementary functions. Instead of exclusively supplying data, consumers only receive inputs from preceding components and do not produce any outputs for other processing units. Due to their role, consumers are typically positioned at the end of pipelines and act as information sinks for providers or transformers. However, this does not prevent them from performing external actions and communicating with modules outside the framework. Examples include components that store the data on the file system (`FileWriter`), transmit information through network sockets (`SocketWriter`), or draw the signals in real-time graphs (`SignalPainter`). In this regard, creating new consumer components only requires inheriting the abstract `Consumer` class and implementing the `consume()` method, which continuously receives the input streams. Similar to providers and transformers, the optional `enter()` and `flush()` methods can also be implemented to perform specific setup or cleanup actions before and after the recurring processing loop.

```
1  // Extend abstract Consumer class
2  public class MyConsumer extends Consumer
3  {
4       [...] // Define potential component options, see Appendix A
5
6       @Override
7       public void consume(Stream[] stream_in, Event trigger)
8       throws SSJFatalException
9       {
10           // Implement data consumption
11      }
12 }
```

Furthermore, integrating consumers into pipelines follows the same rules as for transformers. After creating a component instance and specifying potential options (lines 13-14), the `addConsumer()` method is called, which expects a reference to the consumer and its data providing input sources (line 16). Like with transformers, the optional `frame` and `delta` durations can be specified to adjust the interval between subsequent `consume()` calls and the window length of received `Stream` packages. In the example below, these values are omitted, resulting in the same `frame` duration as the preceding transformer and a `delta` time of zero.

```
13 MyConsumer consumer = new MyConsumer();
14 consumer.options.myOption.set(true);
15
16 pipeline.addConsumer(consumer, transformer);
```

6.3.4 Events

In SSJ, asynchronous communication between internal components and external modules is achieved with events. They offer an alternative method of transmitting signals within the framework and exchanging information with outside solutions. Each `Event` instance contains a payload and basic meta information, such as a name, sender, time, duration, and state (lines 3-7). Depending on the primary data type (e.g., boolean, byte, short, integer, float, etc.), the framework internally uses a corresponding `Event` subclass and stores the values in a matching array (line 10).

```
1 // Create FloatEvent instance
2 Event event = Event.create(Cons.Type.FLOAT);
3 event.name = "MyEvent";
4 event.sender = "EventSource";
5 event.time = pipeline.getTimeMs();
```

```
6  event.dur = 100;
7  event.state = Event.State.COMPLETED;
8
9  // Set event payload
10 event.setData(new float[] {1.4, 3.2, 3.6, 7.9});
```

Communication typically occurs through event channels, which handle the distribution of information and act as a first-in-first-out queue. By default, each component has its own output `EventChannel`, which can be accessed with the `getEventChannelOut()` method. Alternatively, it is possible to create direct `EventChannel` instances like in the listing below (line 24). Interested components can then implement the `EventListener` interface (lines 12-19) and register themselves with the `addEventListener()` method to get notified about new events in a channel (line 25).

```
11 // Implement EventListener interface
12 public class MyListener extends Component implements EventListener
13 {
14     @Override
15     public void notify(Event event)
16     {
17         // Implement event handling
18     }
19 }
20
21 [...]
22
23 MyListener listener = new MyListener();
24 EventChannel eventChannel = new EventChannel();
25 eventChannel.addEventListener(listener);
26 eventChannel.pushEvent(event);
```

Furthermore, `Event` instances can be converted to XML representations for increased compatibility with external solutions. This capability is primarily intended for situations when the `EventListener` interface can not be used due to separated runtime environments or different programming languages. In these cases, the XML-based events can still be exchanged with incompatible modules through platform-independent methods, such as socket connections or file transfers.

```
27 <event sender="EventSource" name="MyEvent" from="10249"
28      duration="100" type="FLOAT" state="COMPLETED">
29      1.4 3.2 3.6 7.9
30 </event>
```

6.4 Graphical Interface

In order to extend the framework's potential user group beyond developers and researchers, we implemented an Android application called *SSJ Creator*. It makes all necessary tools to capture, process, store, and interpret sensor data on mobile devices accessible to people without technical background or programming knowledge. The graphical interface equips regular users with the same capabilities as professional developers working directly with the underlying code and facilitates the creation of personalized solutions tailored to each individual's specific circumstances. By allowing end users to easily build and adapt these processing pipelines, it also encourages them to experiment with different variations and alternative options, which might lead to the discovery of unexpected preferences or solutions that would otherwise not be found. The following sections provide an overview of the technical foundation and showcase the available features of the Android application.

6.4.1 Technical Foundation

The SSJ Creator allows users to add, configure, and connect processing nodes on a visual canvas, which gets internally converted to a corresponding pipeline. Since manually implementing and updating the mapping between code and visual representations can consume much time and resources, we applied an automatic approach instead. More precisely, we extensively used the *reflection* feature of the Java programming language, which enables applications to inspect themselves during runtime. For example, it allows us to automatically populate a component creation dialog by listing all subclasses of the respective types (e.g., `Sensors`, `Transformers`, and `Consumers`). Once an entry from this list has been selected, an instance of the corresponding class gets created and added to the background pipeline (see Figure 6.5). In case a new component gets implemented at a later point in time, it will automatically show up in the list of available subclasses.

The same principle also applies to the component options. In order to separate internal helper variables from externally configurable parameters, we use inner classes that inherit from the abstract `OptionList` class and expose their member variables as part of the framework's public interface. An example of this construct is shown in Appendix A. Through reflection, we are able to automatically read the name, data type, default value, and description of each option and can dynamically create dialog windows with appropriate interface elements that allow users to configure them directly in the application. Since these dialogs are automatically constructed during runtime, they always reflect the underlying component implementation even if options are added, modified, or removed

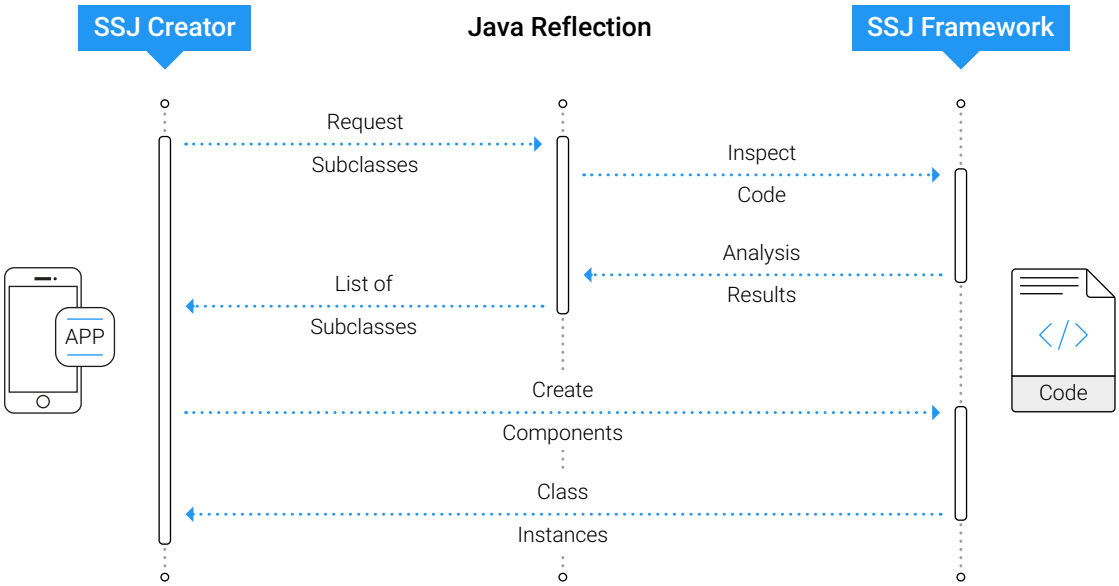


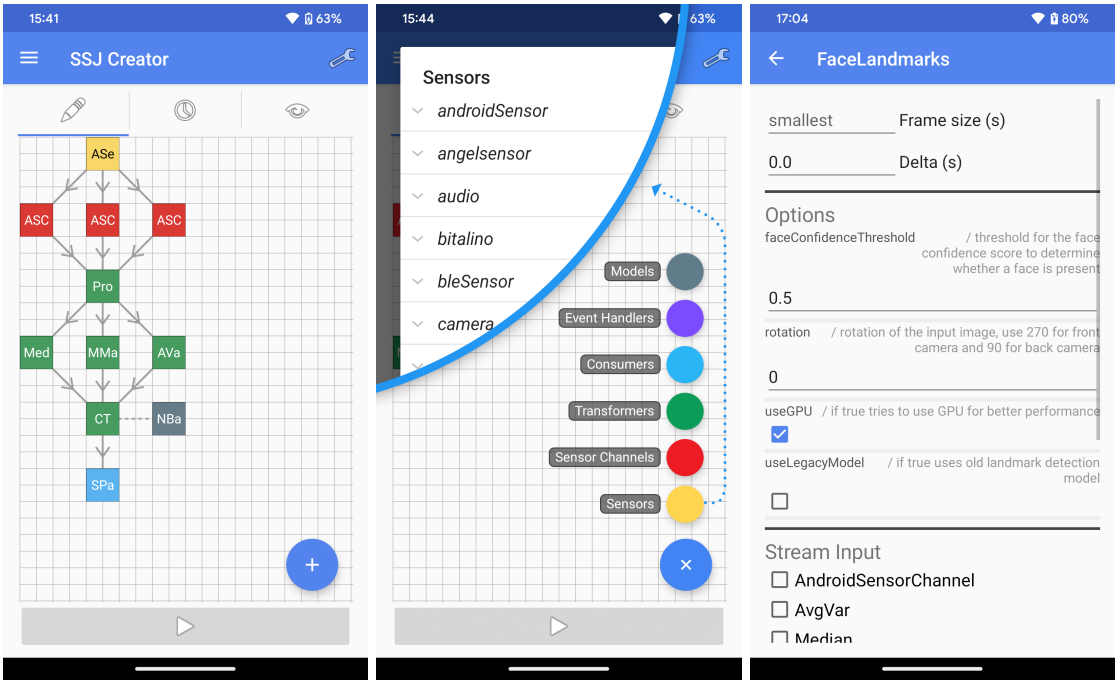
Figure 6.5: Example of Java reflection usage in the SSJ Creator application.

in the future. Apart from generic mechanisms that apply to all components, the application also contains specialized interfaces for specific processing units. For instance, a graph view is added for every `SignalPainter` node within the current pipeline to visualize the incoming data. The combination of automatically generated and manually crafted interfaces ensures a high degree of flexibility and enables the efficient improvement of the framework without having to adapt the application to most changes.

6.4.2 Feature Overview

As shown in Figure 6.6a, the graphical interface primarily consists of a visual pipeline editor. It allows users to build and modify signal processing applications by combining and manipulating nodes on a grid surface. Each node directly represents a component instance that performs a specific task. For example, the blue square labeled "SPa" in Figure 6.6a is a `SignalPainter`, while the yellow node named "ASe" is an `AndroidSensor`. Adding new components to a pipeline can be accomplished by tapping on the floating "plus" button in the bottom right corner and selecting the appropriate category (sensors, sensor channels, transformers, consumers, event handlers, or models). Following that, a list of all subclasses for the chosen type is displayed, which can be used to create instances of the intended components (see Figure 6.6b). After adding a new processing unit to the workspace, users can modify its options by short-tapping on the respective node, which brings up a parameter configuration dialog (see Figure 6.6c). Components can be connected with each other by long-pressing and dragging the data

source over the destination node. Alternatively, the inputs of a processing unit can be selected in the options dialog. Removing nodes is similar to connecting them, which can be achieved by dragging components over a "trash bin" icon in the bottom left corner. Once a pipeline is fully built, it can be started with the "play" button at the bottom of the screen. If the pipeline is already running, the same button can be used to stop it.

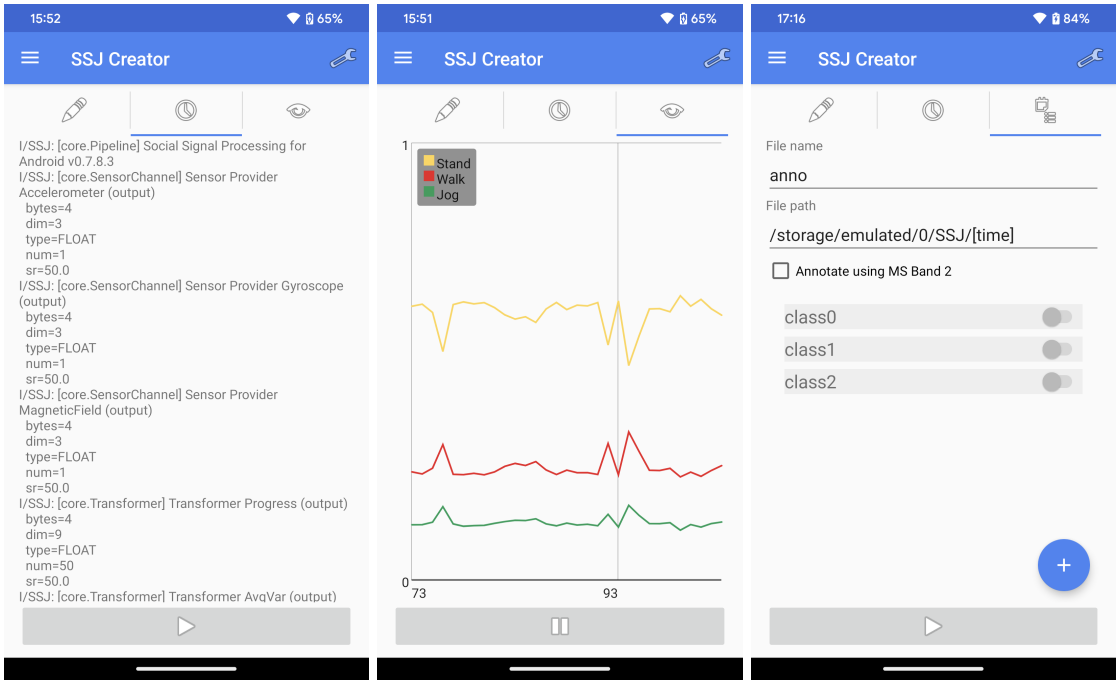


(a) Pipeline editor **(b)** Adding components **(c)** Setting options

Figure 6.6: Primary functions of the SSJ Creator application.

Apart from the pipeline editor, the application also contains other views that are accessible through different tabs. While most of them only appear on demand, the console log is always available. It displays debug information, warnings, and error messages and allows users to monitor the proper execution of their pipelines. One example of an optional tab is the graph view for data visualization (see Figure 6.7b). As mentioned in Section 6.4.1, it gets added for every instance of the `SignalPainter` component and draws the input signals in real-time. Another example is the `VisualFeedback` view, which displays user-defined content according to its configuration. It can be used to prototype and evaluate different visualizations until a satisfactory result is achieved.

Finally, an optional annotation view appears for every `FileWriter` instance added to the pipeline (see Figure 6.7c). It enables individuals to define custom classes and lets them indicate the start and end of their occurrence during recordings. The resulting annotations are stored in a human-readable format compatible with the SSI framework to



(a) Console log **(b)** Signal painter **(c)** Annotation view

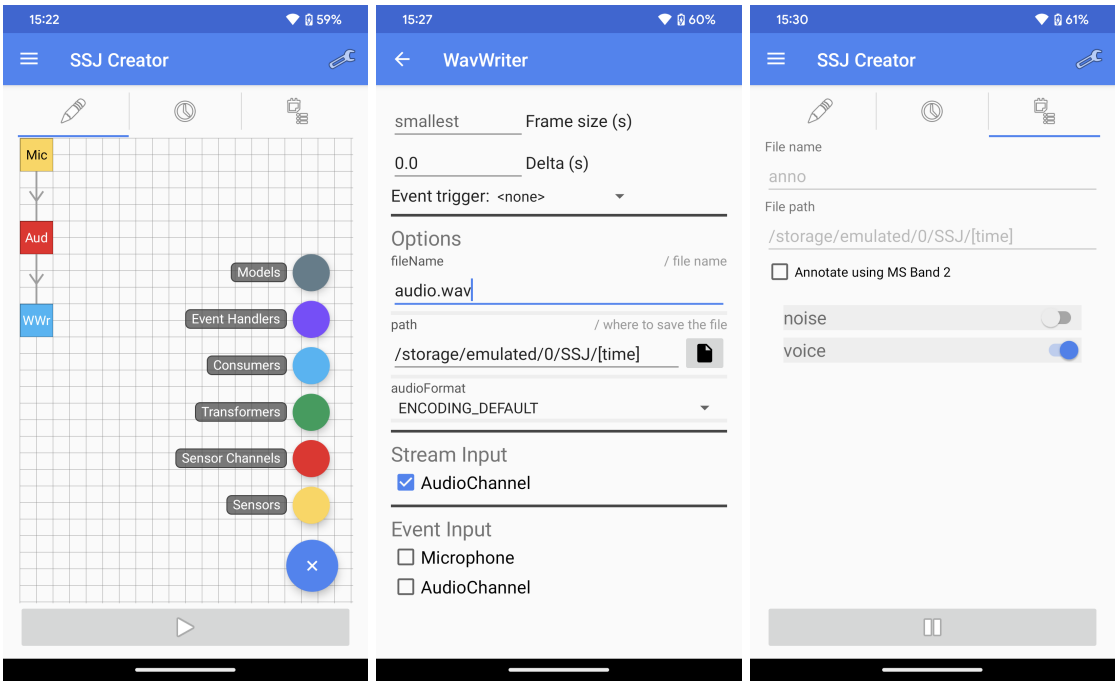
Figure 6.7: Additional views within the SSJ Creator application.

facilitate their unrestricted usage across platforms. Similarly, the pipelines themselves can also be saved and loaded from human-readable XML files. This feature enables individuals to effortlessly and reliably reuse the same processing configurations and allows them to share their creations across devices. Other than that, the application also provides limited functionality for on-device model training, although only naïve Bayes classifiers are currently supported.

6.5 Example Application

This section demonstrates how the capabilities of the SSJ framework can be used to quickly prototype and implement approaches for the assistive augmentation of cognitive processes. In this example, our goal is to build an application for individuals with hearing impairments that helps them recognize when someone in their vicinity is speaking. By informing affected people about the communication intents of others, we aim to improve the quality of their interactions and provide a foundation for future extensions (e.g., speech transcription). The first step involves recording appropriate data to train a voice activity detection model. For that, we can use the SSJ Creator application and add a `Microphone` sensor, `AudioChannel`, and `WavWriter` consumer to the pipeline, as

shown in Figure 6.8a. After connecting and configuring each component, we can switch to the annotation tab and add the target "voice" and "noise" classes (see Figure 6.8c). Once this step is completed, the pipeline can be started to record audio signals and labels that indicate the presence of speech. The resulting dataset can then be used to train a binary classifier for our intended purpose. Alternatively, publicly available corpora and pre-trained models, like VadNet[6], can be utilized as well.



(a) Add components **(b)** Configure options **(c)** Annotate data

Figure 6.8: Building an audio recording pipeline with the SSJ Creator.

The next step involves building a real-time classification pipeline. Although we focus on code examples throughout the rest of this section, the same results can be achieved with the SSJ Creator application. Similar to the data recording pipeline, we start by adding a `Microphone` sensor and an `AudioChannel`. Additionally, we create a `ConvertToDim` component to transform the audio signals into the correct input format for our recognition model. It converts the continuous one-dimensional data stream within a fixed time frame to a single sample with multiple dimensions that contain each value. Instead of processing each sample consecutively, this transformation allows the model to consider all signals from the provided period simultaneously. As shown in line 13, we specify `frame=0.1` and `delta=0.9`, which means the component receives data based on a one-second sliding window every 0.1 seconds (10 Hz).

[6] `https://github.com/hcmlab/vadnet`

```
1  // Create microphone sensor
2  Microphone microphoneSensor = new Microphone();
3
4  // Create audio channel and set sample rate to 16 kHz
5  AudioChannel audioChannel = new AudioChannel();
6  audioChannel.options.sampleRate.set(16000);
7
8  // Create transformer to prepare model input
9  ConvertToDim audioConverter = new ConvertToDim();
10
11 // Add components to pipeline
12 pipeline.addSensor(microphoneSensor, audioChannel);
13 pipeline.addTransformer(audioConverter, audioChannel, 0.1, 0.9);
```

Following that, we create a `ClassifierT` transformer component to perform the voice activity recognition task. For the sake of reproducibility, the example below shows how to load a Tensorflow Lite[7] version of the previously mentioned publicly available VadNet[6] model, but a custom-trained classifier could be used as well. During runtime, it receives the converted audio signals as inputs and provides the recognition results as outputs (a two-dimensional stream with probabilities for the noise and voice class).

```
14 // Create TensorFlow Lite model and select model file
15 TFLite vadModel = new TFLite();
16 vadModel.options.file.set(new FilePath("/model/vadnet_lite.trainer"));
17
18 // Create classifier and select model
19 ClassifierT vadClassifier = new ClassifierT();
20 vadClassifier.setModel(vadModel);
21
22 // Add components to pipeline
23 pipeline.addModel(vadModel);
24 pipeline.addTransformer(vadClassifier, audioConverter);
```

The final step involves providing feedback to potential users in case speech has been detected. For that, we add a `Selector` transformer to isolate the voice probability dimension from the classification results. We then create a `ThresholdEventSender` and configure it to trigger an event if the classifier recognizes voice activity with at least 50% confidence. At this point, we can decide how users should get notified about the occurrence of such events. For the initial prototype, we select vibrations as a feedback modality and add the `AndroidTactileFeedback` component. After setting the intended vibration pattern (line 35, repearing array of “pause” and “active” durations)

[7] `https://tensorflow.org/lite`

and registering it as an event listener (lines 42-43), we can start the pipeline and test the augmentation system with actual users. Depending on their responses, the pipeline can be adjusted to match people's preferences and requirements. For instance, the tactile notifications could be replaced or complemented with a `VisualFeedback` component. Additionally, the pipeline could be extended to not only detect when someone is speaking but also transcribe and display the message contents.

```
25 // Create transformer to select voice dimension
26 Selector voiceSelector = new Selector();
27 voiceSelector.options.values.set(new int[] {1});
28
29 // Create event sender and set voice probability threshold to 0.5
30 ThresholdEventSender voiceThreshold = new ThresholdEventSender();
31 voiceThreshold.options.thresin.set(new float[] {0.5f});
32
33 // Create component for tactile feedback and set vibration pattern
34 AndroidTactileFeedback tactileFeedback = new AndroidTactileFeedback();
35 tactileFeedback.options.vibrationPattern.set(new long[] {0, 100});
36
37 // Add components to pipeline
38 pipeline.addTransformer(voiceSelector, vadClassifier);
39 pipeline.addConsumer(voiceThreshold, voiceSelector);
40
41 // Add feedback component as receiver for threshold events
42 EventChannel voiceEvents = voiceThreshold.getEventChannelOut();
43 voiceEvents.addEventListener(tactileFeedback);
```

6.6 Summary

Based on the identified challenges, requirements, strategies, solutions, and general structure of previous works (see Chapters 4 and 5), this chapter introduced our open-source software framework for building and prototyping assistive augmentation systems using mobile signal processing techniques. To this end, it first discussed the framework's origins, basic concepts, and fundamental principles, such as sampling, modular design, generic data handling, and multimodal synchronization. Following that, an extensive overview of existing mobile signal processing solutions was provided to highlight the common capabilities and limitations that needed to be supported and addressed with our approach. Afterwards, the proposed technical architecture and its core components were presented in more detail. Due to its modular design, each component can be exchanged, extended, reused, and rearranged with minimal effort, which increases flexibility, reduces iteration times, and encourages experimentation with available al-

ternatives. Additionally, the framework supports all proposed design dimensions (see Section 4.4) and contains necessary tools to replicate, adapt, and extend the augmentation strategies identified in Chapter 4. Moreover, it addresses the common challenges of assistive augmentation approaches through various technical features and solutions, including efficient algorithms, energy-saving mechanisms, on-device processing, local data recording, and comprehensive input and output capabilities. In this regard, it also attempts to improve people's trust and acceptance of potential systems by directly involving them in the design and development process through a user-friendly application that even enables people without any technical background or programming knowledge to rapidly build and prototype processing pipelines with a graphical interface. Finally, this chapter also included a basic application example with step-by-step instructions, showcasing how the framework's capabilities can be used to implement effective solutions. To demonstrate and evaluate its practical feasibility, we designed and developed three assistive augmentation approaches in Part III, each targeting a different group of cognitive processes (perception, memory storage, and higher-order cognition).

III Research Probes

7 Assisting Visual Impairment 157

7.1 Augmentation Design
7.2 System Overview
7.3 Evaluation
7.4 Discussion
7.5 Summary

8 Assisting Memory Decline 173

8.1 Augmentation Design
8.2 Condition Analysis
8.3 System Overview
8.4 Evaluation
8.5 Discussion
8.6 Summary

9 Assisting Emotional Disorder 209

9.1 Augmentation Design
9.2 Condition Analysis
9.3 System Overview
9.4 Evaluation
9.5 Discussion
9.6 Summary

Chapter 7

Assisting Visual Impairment

Vision is one of our primary senses to perceive the real world, and thus, the diagnosis of visual impairment presents a great challenge for affected people. According to the latest World Health Organization (WHO) report, at least 2.2 billion people around the world had a visual impairment in 2019 [World Health Organization, 2019]. Since a rising number of affected people are not or only partially able to perceive the environment with their eyes, making the visual world more accessible to them presents an ideal opportunity for assistive augmentation approaches. As outlined in Section 4.3.1, one of the most commonly adopted notions to deal with this problem is sensory substitution. It involves transforming the stimuli from one sensory modality into another to compensate for a defect of the initial modality. To this end, a very promising concept

Figure 7.1: Visually impaired users participating in our user study.

is the automatic generation of semantic descriptions based on image contents, similar to how sighted people explain what they see to blind users. Even though researchers have achieved significant progress in this domain (e.g., Karpathy and Fei-Fei [2015] or Vinyals et al. [2015]), there is one considerable drawback to this method. It takes away the direct perceptual experience and the impressions of actively exploring the images from the visually impaired. Besides that, the generated descriptions only give a rough overview of the image contents, while details such as the visual appearance of individual objects, their position, and color effect are usually not included.

As an alternative, several approaches have proposed the usage of touch-based interfaces that convert specific visual features from the region underneath the fingertip into acoustic representations (e.g., Yoshida et al. [2011] and Banf and Blanz [2013]). This selective and controlled transformation process allows blind and visually impaired individuals to naturally explore images on devices with touch screens, such as tablets and smartphones. However, it also restricts people's ability to use their hands for other purposes while utilizing potential systems and requires them to take pictures of their surroundings, which might not reflect the latest conditions in changing environments. For these reasons, Twardon et al. [2013] proposed the usage of a head-mounted eye-tracker that converts the distance towards objects at the current gaze point of individuals into acoustic signals in real-time. Although their evaluation yielded interesting results, it was only conducted with sighted people and did not investigate whether the eye-tracking device or the applied sonification method might irritate users if the system is operated for an extended period. Unfortunately, such limited evaluations are relatively common among assistive approaches for blind and visually impaired people, which calls their effectiveness and viability to support this user group into question. Due to the previous visual experience of normally sighted people, potential insights from such studies might not translate to real-world usage and lead to inaccurate results.

Consequently, this chapter illustrates how the previously introduced concepts and framework can be used to develop a sensory augmentation system that enables blind and visually impaired people to explore and perceive the environment through their remaining senses. To achieve that, we utilized various independent components to build a processing pipeline that transforms certain image aspects, such as colors and texts, from the users' field of view into acoustic signals while it is still their task to analyze and interpret them. In order to give the users the ability to decide which information is relevant at any point in time, we analyzed their eye movements to control the interactive exploration of the field of view. This enabled a perception experience similar to that of

sighted people. Finally, we evaluated the feasibility of our concept in a user study with seven blind and visually impaired participants (see Figure 7.1).

Parts of this chapter are based on the following publication:

> **Reference** Dietz, M., Elgarf, M., Damian, I., and André, E. (2016). Exploring Eye-Tracking Driven Sonification for the Visually Impaired. In *Augmented Human (AH), Conference Proceedings*, pages 1–8. ACM.

7.1 Augmentation Design

The first step towards building the intended sensory augmentation system was to specify its general properties based on the dimensions identified in Section 4.4 (see Figure 7.2). Since the main goal was to support the perceptual processes of blind and visually impaired users, this selection had certain implications on other dimensions. For instance, the system should be primarily directed at the environment to capture information that would otherwise be unavailable to visually impaired individuals. However, it should also analyze the users to enable eye movements as input method, which is why a hybrid approach was selected. Regarding the initiative dimension, we decided to give users complete control over the interactions as it is extremely challenging to automatically identify suitable situations when this type of augmentation might be beneficial. In turn, this decision meant that assistance could be requested at any moment for varying durations, which resulted in the requirement to make the augmentation universally applicable and usable for extended periods. While we initially considered providing dynamic customization options, we instead opted for a static approach to focus on evaluating its feasibility first with the possibility for extensions in future iterations.

Considering the particular target user group we decided to address, getting user input at an early development stage was another top priority. To this end, we contacted a local association for the blind and visually impaired and conducted a design workshop with them. More specifically, one administrative personnel of the association and two visually impaired individuals who were also involved in the association took part in the meeting. The workshop itself was structured into two sessions. First, we presented our concept using a very basic prototype of the system. The prototype consisted of a simple color sonification demo using a head-mounted camera. The aim of this first session was to give the participants a general impression of the capabilities of sensory substitution systems as well as to gather information regarding the perception of such systems by the visually impaired. Furthermore, we discussed possible incompatibilities of medical

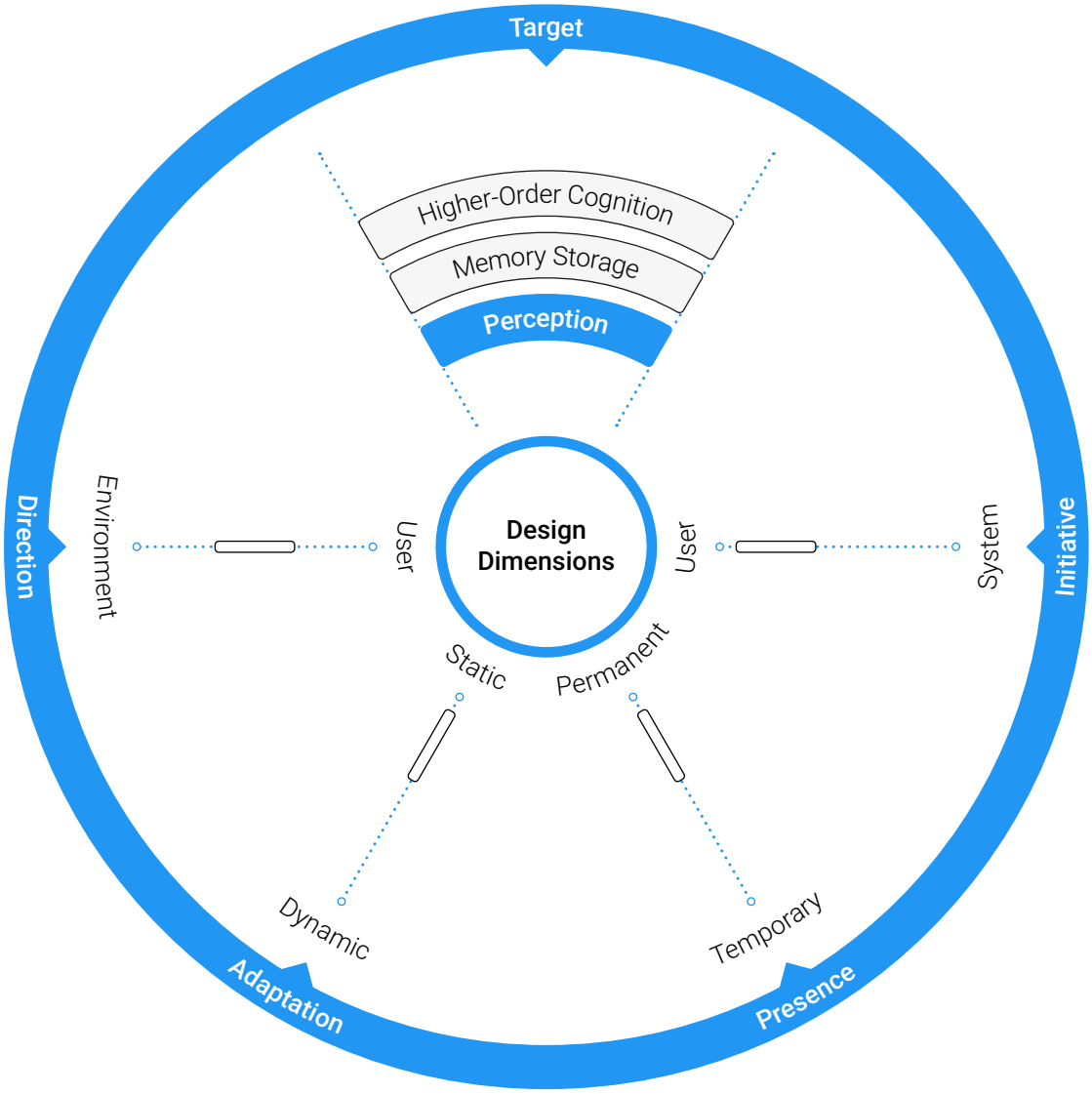


Figure 7.2: Selected design dimensions for augmenting visual impairment.

conditions with eye-tracking solutions. The second session consisted of a brainstorming exercise to identify, on one hand, daily activities visually impaired people struggle most with and, on the other hand, which of those activities could be realistically assisted with sensory substitution approaches.

The workshop yielded valuable insights. First, all three stakeholders showed great interest in sensory substitution solutions. However, concerns were vocalized regarding the visual appearance of the system. According to our stakeholders, many visually impaired fear the social stigma associated with their condition, a reason for which many also refuse to use white canes or other mobility-supporting instruments. While this is indeed a valid concern for technology-enhanced sonification systems, the rapid advancement of

wearable devices in the recent years has shown that the development of inconspicuous solutions is only a matter of time. We also learned that a large part of our target user group may develop pathological nystagmus, or more commonly called “dancing eyes”, which causes the user to lose oculomotor control. Because this condition can strongly impact the accuracy of eye-tracking systems, we decided to restrict our target group to blind and visually impaired people who do not suffer from pathological nystagmus. The second session gave us some clear examples of daily activities visually impaired people struggle with. More specifically, all participants pointed out activities, such as reading text, identifying objects, avoiding obstacles, or navigating unknown streets, as most encumbering. Based on these discussions, we chose to implement two modules to handle the sonification of color and text information.

7.2 System Overview

In order to explore the feasibility of eye-tracking as an input method for blind and visually impaired people, we implemented a sonification system that uses the gaze position to control which part of the user’s field of view should be sonified. After considering the outcomes of the participatory design workshop and the technical requirements to achieve them, we decided to create two independent processing pipelines for color and text sonification (see Figure 7.3). While the original system was not completely mobile due to its reliance on eye-tracking glasses (we used SMI Eye Tracking Glasses[1] because no fully mobile alternatives were available), which at the time could only be connected to notebooks, the same functionality can now be achieved with phone-based solutions

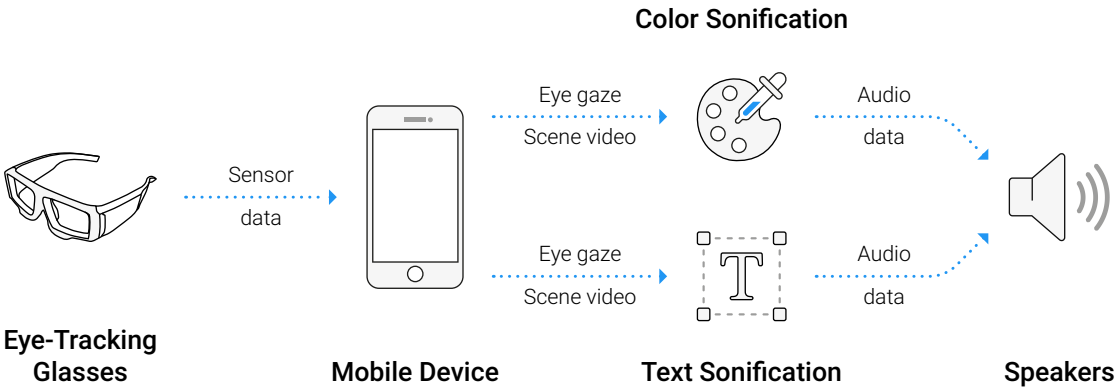


Figure 7.3: Architecture of the sensory augmentation system.

[1] https://smivision.com

(e.g., Pupil Labs Invisible[2]). Due to the modular architecture of the framework, the processing steps and components described in this chapter are still valid and produce the desired augmentation result, regardless of which eye-tracking sensor is used.

In order to capture the eye-tracking data from the respective device (in our case SMI Eye Tracking Glasses), we first added an `SMIETG` sensor component to each pipeline and connected it to the sensor channels `EyeGaze` and `SceneVideo`. While the sensor component is responsible for establishing the connection to the sensing hardware, both channels provide the actual data (gaze coordinates and video stream containing field of view) to the other components within the pipeline. For development purposes, we also added an `FFMPEGReader` and a `Mouse` sensor component to simulate the scene view and gaze coordinates. This enabled us to quickly switch between real data from the actual sensing device and test data based on recorded videos and mouse movements, which significantly accelerated the development process. Since the further processing steps for converting color and text information into acoustic signals are quite different from each other, we describe them separately in the following sections.

```
// Create sensor and channels
SMIETG etgSensor = new SMIETG();
EyeGaze gazeChannel = new EyeGaze();
SceneVideo sceneChannel = new SceneVideo();

// Add components to pipeline
pipeline.addSensor(etgSensor, gazeChannel);
pipeline.addSensor(etgSensor, sceneChannel);

// Create color sonification consumer
ColorSonification colorSonification = new ColorSonification();

// Create text sonification consumer
TextSonification textSonification = new TextSonification();

// Add components to pipeline
pipeline.addConsumer(colorSonification, new Provider[] {
    gazeChannel, sceneChannel
});
pipeline.addConsumer(textSonification, new Provider[] {
    gazeChannel, sceneChannel
});
```

[2] `https://pupil-labs.com`

7.2.1 Color Sonification

In order to implement the functionality for transforming colors into sounds, we created a new component called `ColorSonification` that receives the current gaze position and video frame as inputs and produces the corresponding sounds as output. The module is based on the idea that sounds can be mixed similarly to colors. To this end, we create an "audible color space" by mapping specific color values of the HSL color space to an appropriate counterpart within the sound space, as proposed by Banf and Blanz [2012, 2013]. Through that, the primary colors are represented by their respective sounds, while mixed colors can be identified by the mixture of two primary sound components. In combination with eye-tracking as an input method, the users should be able to explore the environment by moving their eyes. For instance, it enables them to differentiate between red and green apples while buying groceries and allows them to determine the color of their clothes when doing laundry.

Furthermore, with a bit of training, it might even be possible to recognize objects through the color differences of their contours, as shown by Abboud et al. [2014], Banf and Blanz [2013], and Covaco et al. [2013]. However, while those approaches only use static images for processing, we wanted to be able to sonify the video stream of the user's field of view in real-time. Therefore, we could not use those concepts since they contain certain image operations that require too much processing time and would cause performance problems if applied to every frame of a video stream. As a result, we came up with our own approach based on the work of Banf and Blanz [2013].

In general, one of the challenges of color sonification is the fact that color values in images often change rapidly from one pixel to another, even though the overall color of the material or texture is roughly the same. The reason for that is mostly due to image noise caused by the camera, which leads to faulty pixels with differing color and brightness values. Since the sonification of those pixels would lead to wrong impressions and could confuse the users, we needed to remove them first. However, doing that for the whole image region would have caused performance issues and was not necessary in our case. Instead, we first extract an area of 100×100 pixels around the current gaze point (x, y) from the video frame and apply a bilateral filter to it. As a result, the noise gets removed while the edges within the image are still preserved [Tomasi and Manduchi, 1998]. After that, the image section is converted to the HSL color space to extract the smoothed values for hue $h(x, y)$, saturation $s(x, y)$, and lightness $l(x, y)$.

The combination of these values is then mapped to a corresponding sound. As shown in Figure 7.4, we used the following assignment of MIDI instruments to colors similar to

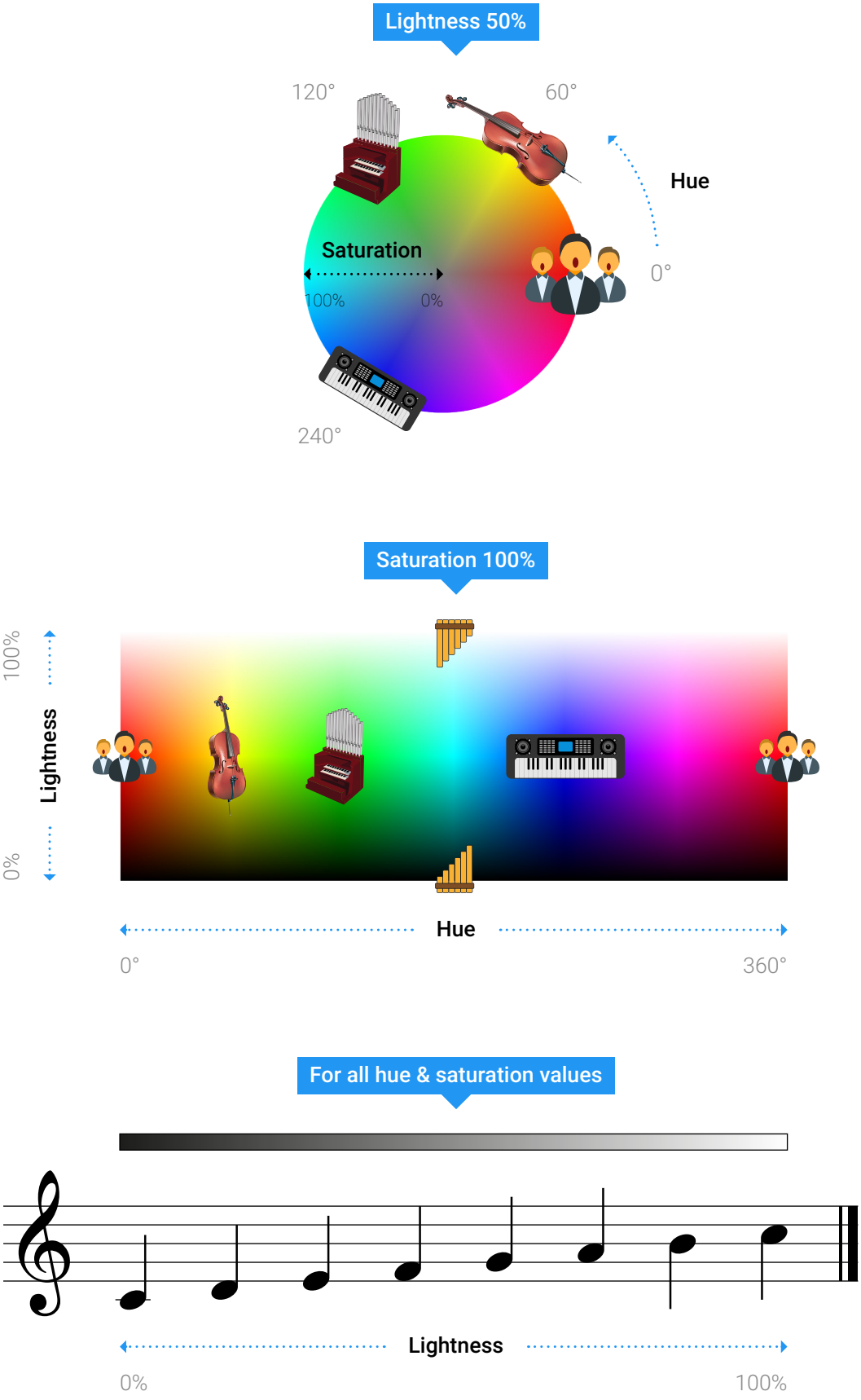


Figure 7.4: Assignment of colors to MIDI instruments.

Banf and Blanz [2012]: *flute* (black, white, gray), *choir* (red), *organ* (green), *synthesizer* (blue) and *cello* (yellow). However, this assignment was just an initial suggestion, which can be adjusted according to user preferences in future iterations. For instance, we tried using birds' twittering for green and wind noises for blue to simplify the mental mapping of colors. Although this was a bit more intuitive, we found that those sounds will likely irritate individuals if the system is used for extended periods, which is the reason why we reverted to MIDI instruments. In conjunction with that, the sonification of secondary colors is achieved by playing the instruments assigned to both involved primary colors simultaneously with a certain sound level. Thereby, the amount of each instrument is controlled through a specific volume shape ($v(h,s,l)$ with $v \in [0,...,1]$), which maps each combination of hue h, saturation s, and lightness l to a value between 0 and 1 [Banf and Blanz, 2013]. For instance, the volume shape $v_{choir}(h,s,l)$ returns 1 for $h = 0°$, $s = 100\%$, and $l = 50\%$, while the volume shapes for all the other instruments return 0. Therefore, only the choir is played with maximum loudness. In addition to the volume, we also adjust the pitch according to the current lightness. For that, each lightness value l between 0 and 1 is mapped to one of the eight tones of a musical scale from C4 (261.6 Hz) to C5 (523.2 Hz). The resulting tone is then played by all instruments even if they can not be heard due to the value of their volume shape. This enables a more precise sonification experience, allowing users to distinguish between different colors as well as dark and bright color variations.

7.2.2 Text Sonification

One of the most frequently mentioned problems during our interviews with blind and visually impaired people was the loss of the ability to read texts. While there are already some approaches (e.g., Alt et al. [2010, 2013] or Pfleging et al. [2012]) that convert text information into audio signals, most of them only focus on the intention of the sentences rather than the actual content. For the sonification of nameplates, street signs, billboards, or shop signs this is not very helpful, as in those cases, only the meaning of the text is relevant to the user. Therefore, we came up with the following approach:

Similar to the `ColorSonification` module, we created a new component that also receives the current gaze position and video frame as inputs. In the first step, this data is used to analyze if any visible texts are present within the user's field of view. Since this is a computationally intensive operation, it can not be done for every frame of the video stream. Instead, we only execute this process as soon as the previous run has finished, which results in an average execution rate of once every 2-3 seconds. For the

Figure 7.5: Conversion of text position to sound.

image data analysis, we use the Stroke Width Transform (SWT) algorithm by Epshtein et al. [2010] due to its high precision while only requiring a relatively short processing time compared to other text detection algorithms. It is also language-independent and works with many different fonts and sizes. After every execution, the algorithm returns a set of rectangles designating the areas in which texts have been detected. The rectangles are then sorted by the y-value of their center points in descending order so that the sonification of multiple texts is always performed from top to bottom. In this regard, the x- and y-coordinates of each detected text area are used to generate a sound that allows the users to locate its position [Brock and Kristensson, 2013]. As shown in Figure 7.5, the x-coordinate determines the sound's stereo panning, while the y-coordinate is mapped to a two-octave musical scale from C3 (130.8 Hz) to C5 (523.2 Hz). Initially, we only used one octave, but after a few tests, we found that a two-octave musical scale allows for a more precise localization. However, using more than two octaves did not lead to any further measurable benefits. The sound itself is generated through a vibraphone MIDI instrument. It was chosen since it resembles a pleasant notification tone and distinctively differs from the sounds of the color sonification module.

Once the user moves his gaze into one of the text areas, the text-to-speech conversion is triggered. More precisely, the corresponding section is first gets extracted from the image and is then processed using a binary threshold function to simplify the text recog-

nition [Otsu, 1979]. After that, we extract the actual text with Tesseract[3], an open-source optical character recognition (OCR) framework. Subsequently, the output is passed to the Microsoft Speech API[4], which reads the text to the user. Until that is done, the sonification of new texts is disabled to prevent any potential disturbances and distractions. The biggest advantage of this approach is that the user has complete control over the text sonification at any given time. This is especially helpful in situations when multiple texts are visible at once since the sequential sonification of all detected texts would otherwise overwhelm the user. Furthermore, individuals might not be interested in all the text information in their field of view. However, with our method, users can always decide when which text should be read to them.

7.3 Evaluation

In order to get an accurate impression of the system's effectiveness and usefulness for the target group, we conducted a user study with blind and visually impaired people with intact oculomotor control. Since this is a very special user group that can not be reached without direct contact, we cooperated with the Bavarian Association for the Blind and Visually Impaired (BBSB)[5]. With their help, we were able to find seven users

ID	Age	Gender	Visual impairment	Input method
P1	68	male	Cataract	head movement
P2	49	female	Cataract (early stage)	gaze position
P3	43	female	Optic atrophy	gaze position
P4	73	male	Congenital blindness	head movement
P5	68	male	Optic nerve damage	head movement
P6	87	female	Macular degeneration	gaze position
P7	70	male	Retinal degeneration	gaze position

Table 7.1: List of participants.

who met all requirements and agreed to participate in our study, as shown in Table 7.1. Even though the average age of the subjects was above 65 years, they were very open-minded towards and interested in new technologies. The study included a series of tasks and scenarios that simulated real-world situations where the system would be used,

[3] https://github.com/tesseract-ocr/tesseract

[4] https://msdn.microsoft.com/en-us/library/ee125663.aspx

[5] https://bbsb.org

such as identifying objects or reading texts. To evaluate the system's performance, we collected both qualitative feedback from questionnaires and quantitative data based on task completion metrics.

7.3.1 Experiments

Within the user study, each module of our sonification system was evaluated separately to prevent any mutual influences and to enable reliable conclusions from the individual results. Consequently, each experiment was adjusted to the designated use case scenario of the corresponding module. However, in order to ease the arrival of the participants, the study was conducted at the premises of the BBSB, and therefore, no complex setups could be utilized. Instead, only portable objects and tools were used. After each experiment, participants were asked four questions regarding the difficulty and usefulness of the system and the pleasantness of the sounds, which they could answer on a five-point Likert Scale. At the end of each trial, we also asked the users whether they would prefer to always run the modules in parallel or to activate them on demand.

Experiment 1: Color Sonification

In this experiment, we examined the ability to recognize objects by their color and shape using the color sonification module. As preparation, we first presented the mapping of colors and corresponding sounds to the participants. Once they had memorized them, we started a training phase with the examples shown in Figure 7.6 (left). During that, users were asked to move their gaze from left to right and to repeat this process vertically to explore the images from top to bottom. With those examples, we wanted to make the users aware of the color and sound differences between the colored shapes and the black backgrounds. Moreover, we taught participants that they could use the duration of each sound to identify the shape of objects. In the case of the square, assuming constant eye gaze speed, the sound for green always has the same duration when scanning the image, while in the case of the triangle, the sound for yellow is played shortly at the top and longer in the bottom region. Once participants were familiar with this concept, we started the experiment. For that, an apple was placed in the user's line of sight, as shown in Figure 7.6 (right). Participants were told that the object was either a red apple, a banana, or an orange. Now, the task of the participants was to identify which object was in front of them only by using the color sonification module.

Figure 7.6: Training examples (left) and experiment set-up (right).

Experiment 2: Text Sonification

The second experiment was used to evaluate the text sonification module. For that, we trained participants with two examples to clarify the transformation of text positions to acoustic signals. In the first example, the text area appeared in the top right corner of their field of view, resulting in a high-pitched sound coming from the right speaker. As soon as participants looked in the direction of the text, it was automatically read to them. In the second example, the text was shown in the bottom left corner, and the corresponding low-pitched sound from the left speaker was played. Once users were familiar with locating the text positions, we began with the actual experiment. Similar to the examples, users had to locate the text and move their gaze into the corresponding area. For each participant, we measured whether the text position was recognized correctly and how much time was required for the experiment.

7.3.2 Results

Each of the seven participants performed both experiments successively in one session with an average length of about 40 minutes per user. Generally, all of them completed the tasks without any major problems. However, in three cases, we noticed that the eye-tracker could not detect the gaze position correctly. In order to still obtain results from those users regarding the system itself, we adjusted the pipelines to use the center point of their field of view instead of the actual eye gaze. More precisely, we replaced the `EyeGaze` channel with a component that always provided fixed values $(0.5, 0.5)$

as coordinates for the current gaze position. This modification is in line with one of the goals of assistive augmentation, which strives to make technology accessible to as many people as possible, regardless of their conditions, and showcases the ability of the SSJ framework to achieve that. In our case, it enabled affected participants to control the sonification by moving their head in a certain direction. Table 7.1 shows for which individuals this was done. With the adjustments in place, six users were able to correctly identify the object in the first experiment only by using the color sonification. In the second experiment, it was even possible for all participants to detect the text position.

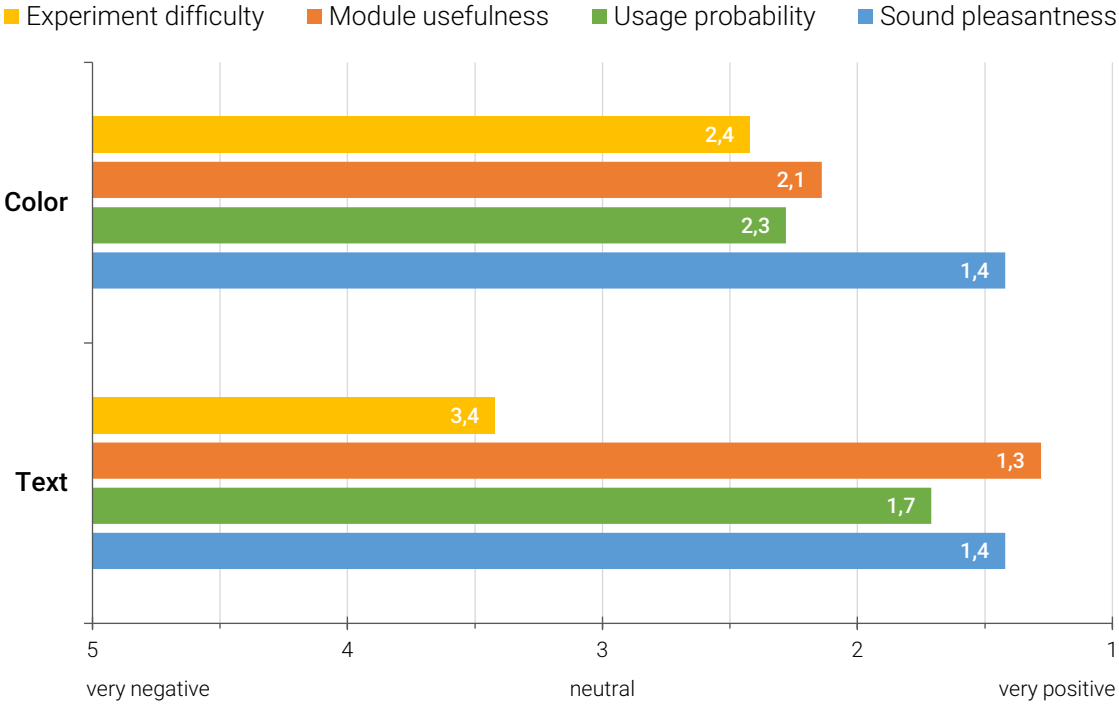


Figure 7.7: Questionnaire results.

Figure 7.7 shows the results from the questionnaires, where each number represents the average value across all participants. Generally, most of the results are in the positive area. The only exception to that is the difficulty of the text sonification experiment. In return, the module usefulness, usage probability, and sound pleasantness were rated more positive in this case than for the other module. A further result, which is not shown in the diagram, was regarding the question of whether the modules should always run in parallel or be activatable when needed. With 85.7%, the majority of all participants voted for an activation on demand.

7.4 Discussion

Overall, the evaluation study yielded positive results. Both color and text sonification modules proved to be useful and usable by blind and visually impaired individuals. From the questionnaire data, we can conclude that the likelihood of such a system being used by members of our target group is reasonably high. We were also pleased by the positive ratings for the sound pleasantness, which suggest that the system could be used over extended periods of time. Nevertheless, some users stated that they would have needed a longer training session to be more proficient in recognizing the different audio cues. P4 was more critical, stating that while the color sonification was helpful for recognizing simple objects (the apple), it might be more challenging for complex objects. Some participants pointed out that the usefulness of the modules might be influenced by the activity they wish to perform: *"I could imagine using it* [the color sonification] *in certain situations"* (P3).

We also observed some technical limitations of the system. For the color sonification module, we found that under certain conditions, the camera we used would falsify the colors. More specifically, during our user study, a brick wall outside the window significantly shifted the colors of objects within the room into the red spectrum. Here, a different camera might resolve the issue. The text sonification module currently also suffers from a relatively slow update rate, allowing text fields to be recognized only once every 2-3 seconds. This problem could be addressed by utilizing more efficient hardware in future iterations of the system. In our study setup, we also used a pair of stationary speakers. However, wireless bone conductance headphones could improve the usability as well. This would make the system more portable while not limiting the user's ability to hear and react to outside events. Furthermore, it would be interesting to evaluate the system in a more real-world setting over a longer period, as opposed to the relatively controlled environment in our present study. Such an opportunity could also be used to investigate how the system handles more complex visual scenes.

Although the eye-tracking technique proved to be successful, it did not work correctly in three of the seven cases, where the center point of the image had to be used instead of the eye gaze to identify the information to be sonified. This was necessary because of specific medical conditions that accompanied the visual impairments of those individuals. For instance, one user was diagnosed with cataract, a condition that causes the lens of their eye to turn cloudy and misty rather than being clear and transparent. This interfered with the eye-tracking glasses' ability to track the position of the pupil, as it was designed to detect dark-colored pupils instead. In another case, congenital eye blindness

caused one participant to have difficulty controlling his eye movements because he had never purposefully used his eye muscles before. The third participant suffered from optic nerve damage, which only permitted him to partially open his eyes. This did not give the eye-tracker a clear view of the pupil. However, there were no prominent differences in the results between participants who were able to use the eye-tracking method and those for which the center of the field of view was used instead. While all these issues were unavoidable with the eye-tracking glasses we employed, the majority of problems can be attributed to the fact that these technologies are designed for and tested with normally sighted people. Explicitly accounting for such variations in the human visual system during the design and development of future eye-tracking solutions might make them more robust toward the blind and visually impaired.

7.5 Summary

This chapter introduced a concrete example of a sensory augmentation system, which can assist affected individuals in perceiving specific visual information that would otherwise be inaccessible to them. Through the use of the SSJ framework, we explored the feasibility of eye-tracking as an input method to control the sonification for blind and visually impaired people. In order to identify the most useful applications for that, we conducted a design workshop with members of the target user group. Based on their feedback, we implemented an eye-tracking-driven sonification system capable of converting colors and texts from the users' field of view into acoustic signals. Through that, individuals are able to decide which elements should be sonified at any point in time just by moving their eyes. To evaluate the effectiveness of our approach, we conducted a user study with seven blind and visually impaired people. Generally, all modules of the sonification system worked as intended and were perceived rather positively by the participants. Although we limited our target user group to individuals who can move their eyes and do not suffer from pathological nystagmus, three participants were still unable to use their eyes to control the system. The reasons for that can be mostly attributed to the nature of their visual impairment. For example, the cataract of one participant had progressed so far that the pupil was almost white and thus could not be detected by the eye-tracker. In another case, the visual impairment caused by an accident only allowed for restricted eye movements. Therefore, future work should focus on identifying which conditions enable the usage of such a system. For people who can actually utilize our approach, eye-tracking appears to be a very promising input method to control the sonification of visual information.

Chapter 8

Assisting Memory Decline

With increasing age, it is common to experience a decline in working memory as part of the normal aging process [Salthouse and Babcock, 1991]. This can lead to forgetfulness and is generally associated with an overall decrease in quality of life. Over 15% of older adults can even develop more severe memory-related issues, including memory loss, confusion, and other cognitive impairments [Weyerer and Bickel, 2007]. Consequently, forgetting a name, an object, or an appointment can lead to very unpleasant situations. To prevent such circumstances, memory augmentation approaches can be utilized to recognize problematic conditions and offer appropriate support. One major use case for such a system is the automatic detection of instances when users search for misplaced objects like keys or wallets. Such episodes are often

Figure 8.1: Typical examples of memory lapses: visual search for misplaced objects.

experienced by older adults due to the cognitive decline of their working memory, which can be very frustrating and time-consuming (see Figure 8.1).

In order to support affected individuals, this chapter introduces an approach that uses the SSJ framework to detect visual search episodes in real-world scenarios and provides assistance for remembering the location of misplaced objects. The system described in this chapter was developed as part of the nationally funded Glassistant project[1], which aimed to support older adults by creating an autonomous assistant using smart glasses and wearable sensors. To achieve our goal, we first identified suitable behavioral signals that could be used to detect visual search instances caused by memory lapses. Based on our findings, we developed a completely mobile eye- and head-tracking device to capture the necessary sensor data. The resulting system was specifically designed to meet the requirements of older adults and was used to collect realistic data from 30 participants. We then trained and integrated a classification model into a real-time recognition pipeline that continuously analyzes eye and head movement data. Once the system detects a potential visual search episode, assistance is offered to aid individuals in finding the desired object. Finally, the system was evaluated in a study with eight older adults who showed indicators of mild memory impairments.

Parts of this chapter are based on the following publications:

Reference Dietz, M., Schork, D., and André, E. (2016). Exploring Eye-Tracking-Based Detection of Visual Search for Elderly People. In *Intelligent Environments (IE), Conference Proceedings*, pages 151–154. IEEE.

Reference Dietz, M., Schork, D., Damian, I., Steinert, A., Haesner, M., and André, E. (2017). Automatic Detection of Visual Search for the Elderly using Eye and Head Tracking Data. *KI - Künstliche Intelligenz*, 31(4):339–348.

Reference Seiderer, A., Dietz, M., Aslan, I., and André, E. (2018). Enabling Privacy with Transfer Learning for Image Classification DNNs on Mobile Devices. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs), Conference Proceedings*, pages 25–30. ACM.

[1] `https://interaktive-technologien.de/projekte/glassistant` – Funded by the German Federal Ministry of Research and Education (BMBF).

8.1 Augmentation Design

Imagine you have an appointment soon and are about to leave the house. You look for the keys, but they are not where you thought they should be. In a hurry, you start looking around, opening drawers, and checking your pockets, but the keys are not there. After a quick glance at your watch, you get even more stressed and frustrated since you are already late. As time goes by, you start to search in more unlikely locations until you finally find them where you never assumed they could be in the first place. Many people can probably relate to this scenario and might experience similar situations from time to time. For older adults, this happens even more often and negatively affects their daily lives. Therefore, we propose the following concept to recognize this situation and support affected individuals accordingly.

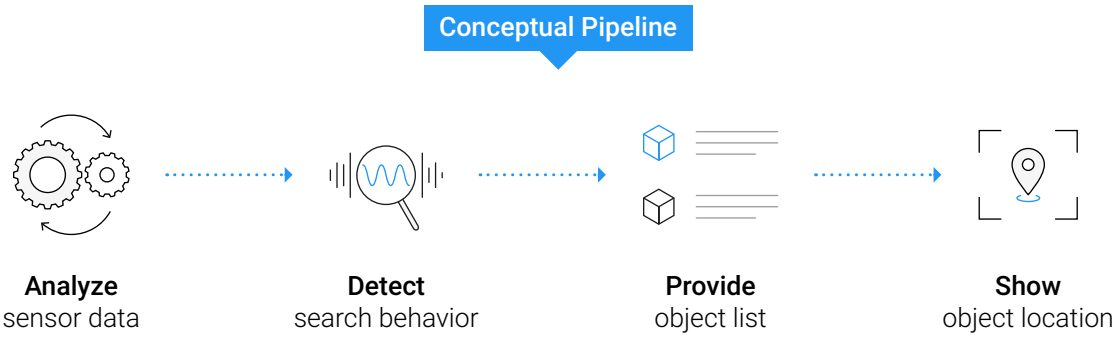


Figure 8.2: Conceptual pipeline to support visual search for objects.

As shown in Figure 8.2, an important step of the proposed pipeline is to determine whether the user is searching for something. This information is required to identify the point in time when the person needs assistance. Otherwise, the system could not act proactively and would require an explicit action or trigger from the user. However, such an approach would lead to situations where the system could be helpful but is not used because the person refuses to admit that they are in need of support. Since it is easier to accept help rather than to ask for it in the first place, we decided that the system should initiate the interactions and offer assistance whenever a critical situation is recognized with the option to decline if the user still does not wish any support. In terms of the presence dimension (see Figure 8.3), this meant that the augmentation could be tailored to the specific use case and only had to be available temporarily during that time. While recognizing episodes of visual search would likely only require sensors directed at the user, identifying desired objects also necessitates an analysis of the environment, which is why a hybrid approach was selected. Regarding the tracking of desired items, our concept intended that a list with all previously configured objects should be suggested

to the user once the system detects a suitable situation. For that, the system was designed with adaptive customization options in mind, allowing individuals and family members to adjust the list of potentially desired objects beforehand. After selecting an item, its respective location should be indicated through appropriate visualizations.

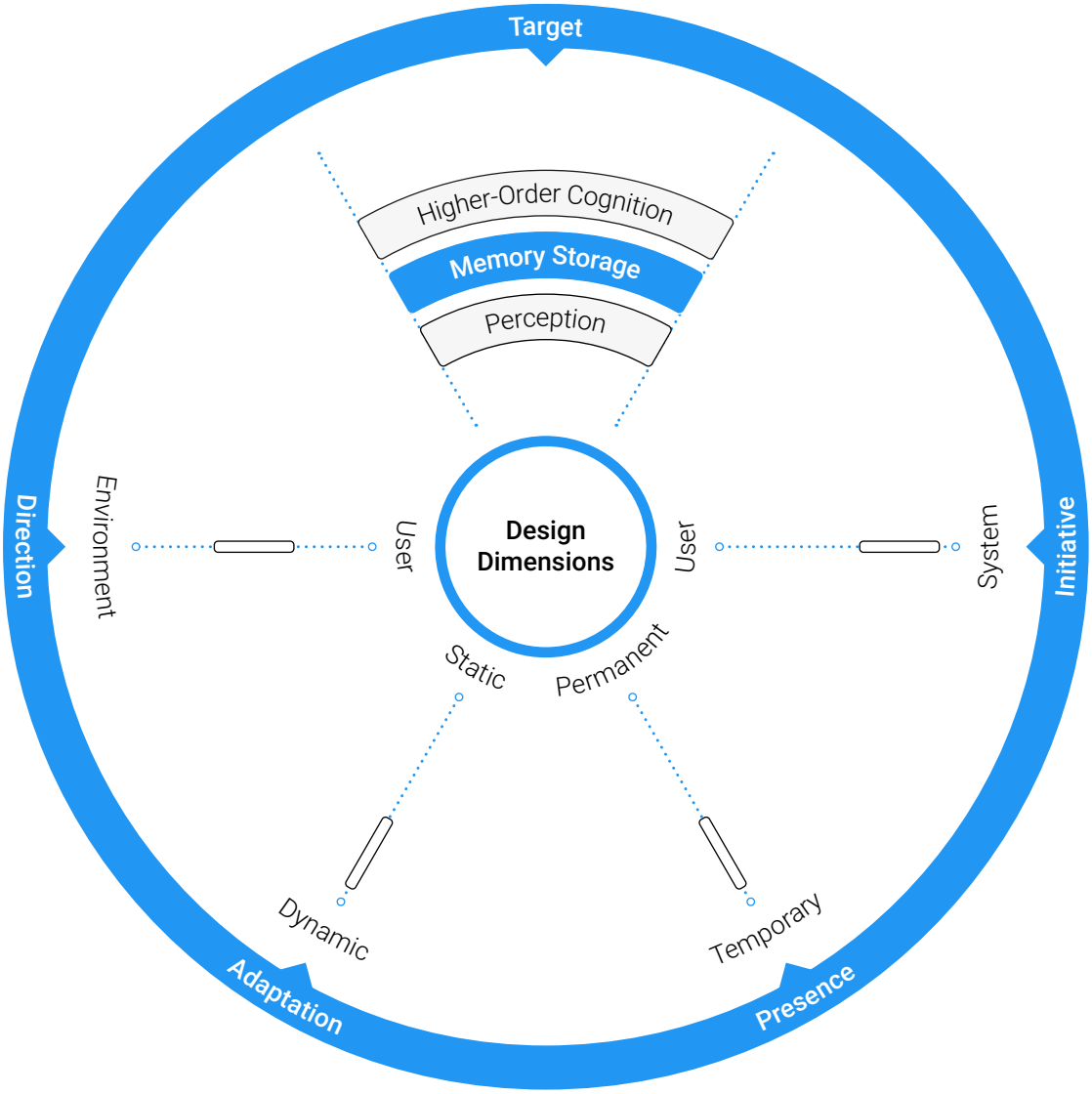


Figure 8.3: Selected design dimensions for augmenting memory decline.

8.1.1 Object Localization Visualizations

In order to create effective visualizations that help individuals find desired objects, we followed the design process proposed by Jain et al. [2015], who evaluated different categories of visualizations on head-mounted displays (HMDs) to support the spatial localization of sounds for users with hearing impairments. The first step of their ap-

proach involved developing a set of goals to guide further design decisions regarding the appearance and functionality of potential visualizations. Due to the similarities between use cases, we derived most of our goals from Jain et al.'s results and came up with the following list:

- *Helpful:* The main goal was to develop a system that would benefit its users.
- *Accurate:* The visualizations should precisely indicate the object position.
- *Glanceable:* The information should be easily comprehensible at a glance.
- *Responsive:* The interface should update and react in real-time.
- *Complementary:* The visualizations should augment the user's abilities without replacing them.
- *Universal:* The visualizations should support every type of object and work regardless of the user's position.

Based on these goals, we identified four general design dimensions, as shown in Figure 8.4. For each dimension, we created at least two different visualizations through variations of the following properties: size, perspective, shape, layout, and detail. Since the focus at this stage was to determine the most suitable design, potential concepts were unrestricted by technical considerations and corresponding limitations. The details of each dimension are described in the paragraphs below.

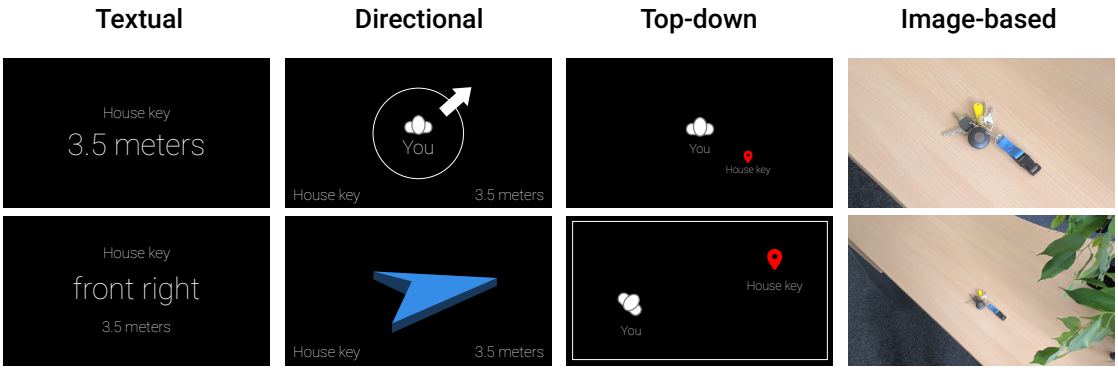


Figure 8.4: Overview of visualization candidates.

Textual This dimension refers to visualizations that primarily convey information in text form. With our concepts, we evaluated whether it is sufficient to only display the distance towards the object or if additionally providing directions is preferable.

Directional Visualizations in this category mainly use arrows to indicate the direction of objects relative to the user's position. Our first design presents the information in a top-down view with a 2D arrow that rotates around the person. In this regard, the upper

area of the visualization corresponds to a person's front, while the lower area represents their back. The second concept uses a 3D arrow that rotates around a vertical axis to indicate the direction from an egocentric perspective.

Top-down Similar to the first directional design, this type of visualization presents the information in a top-down view. One major difference is that this dimension visually represents the distance to the object rather than providing a textual measurement. While the first variant indicates the distance and direction of the object relative to the person's current position, the second design displays the absolute positions within the room.

Image-based This category serves as an alternative to the other dimensions since it provides a picture of the object instead of the distance or direction. The idea behind the image-based concept is to show the item at its last seen location, with the intention of triggering episodic memories about the object's position. Alternatively, the environment captured in the image can help users find the right location, which is why we evaluated different zoom levels that contain more or less contextual details.

8.1.2 Design Probe and Evaluation

To identify the most helpful visualization for our target user group, we first implemented a prototype version of all concepts on the Google Glass head-mounted display. Depending on the position and head orientation of the users, visualizations changed accordingly to convey the impression of a working system. This was achieved by analyzing the accelerometer and gyroscope data from the Google Glass in real-time and adjusting the visualization parameters correspondingly. For instance, the arrows were dynamically rotated in the direction-based design to always face north. After creating the prototype, we conducted a within-subjects study with ten elderly participants (50% female) aged between 61 and 86 ($M = 71.9$) years to evaluate the designs and gain a deeper understanding of the problems faced during the process of searching for misplaced objects.

Procedure

At the start of each session, we explained the study procedure to the participants and gave them a short introduction on how to use the Google Glass (see Figure 8.5). Once the subjects were comfortable wearing and operating the device, we asked them to imagine the following situation: *"You are currently at home and have an important appointment soon. As you are about to leave the house, you suddenly notice that your keys are*

Figure 8.5: Participant (left) and experimenter (right) during the design probe.

missing. How do you react?". Due to the open nature of this question, we recorded all of their responses and asked them relevant follow-up questions, such as *"Where would you search?"*, *"How would you feel?"*, or *"How often do you experience such a situation?"*. After that, we asked them to imagine the same situation again, but this time, they should react to what was displayed on the Google Glass, where the dialog shown in Figure 8.6 appeared a few seconds later.

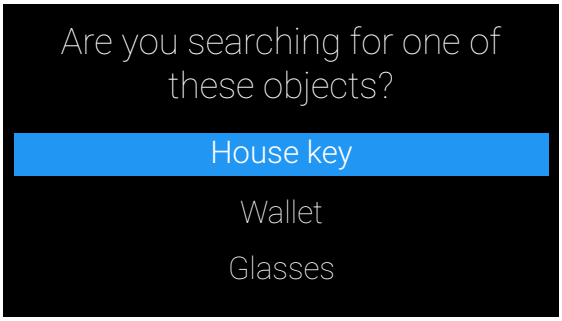


Figure 8.6: Object selection screen displayed on Google Glass.

Once the participants selected the *house key* option, one randomly chosen (counterbalanced) visualization type was displayed. We then gave the users time to explore the different designs within the selected category and asked them which option they preferred, what was positive/negative, and how it could be improved. Afterwards, we asked them

the following questions, mostly adapted from Brooke's [1996] System Usability Scale (SUS) questionnaire, which could be answered on a five-point Likert scale.

- *Helpfulness*: How helpful do you find this visualization for locating misplaced objects?
- *SUS*: I felt very confident using the system with this visualization.
- *SUS*: I would imagine that most people would learn to use the system with this visualization very quickly.
- *SUS*: I found the visualization unnecessarily complex.
- *SUS*: I think that I would like to use the system with this visualization frequently.

Following the within-subjects design, this procedure was repeated for each of the four visualization types shown in Figure 8.4. At the end of the study, we asked the participants which one of the visualizations would be the most helpful to them in the imagined situation, how they would rate the system as a whole, if they would use such a system in their daily life, and whether they had any further positive or negative comments.

Findings

Most participants reacted similarly when confronted with the imaginary situation at the beginning of the study. Four users stated that they would start to search in their pockets, three users said that they would look for the keys outside or in the vicinity of the door, while the others mentioned that they would search in the hallway, the kitchen, or in a dedicated key storage box. Additionally, four participants expressed that they would try to remember when they had used their keys the last time. Although all users intended to look for the keys in a nearby area, the different starting locations would have led to varying experiences in terms of search duration and frustration level. Therefore, providing some sort of guidance can unify the search process and reduce the required time to find the object. When asked about their feelings, six users mentioned that they would be nervous, stressed, or even in a slightly panicked state during that situation. In accordance with previous findings, this confirms that supporting the visual search for hidden or misplaced objects can be beneficial to this user group.

The results of the questionnaires regarding the four visualization types are summarized in Table 8.1. When looking at the helpfulness of each visualization, participants gave the textual and directional variants the highest average rating of 4.1, closely followed by the image-based version with 4.0. However, after seeing all visualizations, 60% of participants found the image-based variant the most helpful. Only 20% each preferred the directional or the top-down view, while no one favored the textual version for the given

Visualization	Helpfulness	SUS score	Most helpful
Textual	**4.1**	72.50	$n = 0$
Directional	**4.1**	71.87	$n = 2$
Top-down	3.6	61.25	$n = 2$
Image-based	4.0	**77.50**	$\mathbf{n = 6}$

Table 8.1: Design probe questionnaire results.

application scenario. A similar trend can be observed in the projected SUS score based on the four questions adapted from the SUS questionnaire. There, the image-based variant also received the highest score of 77.50, followed by the textual (72.50), the directional (71.87), and the top-down visualization (61.25). Overall, participants gave the image-based visualization the highest ratings and preferred it in their comments over the other options (*"I liked the version with the image the most"*). Eight participants even mentioned that they would use such a system in their daily lives, and almost all subjects evaluated the general system as good ($n = 4$) or very good ($n = 5$). Consequently, we decided to implement the image-based visualization method.

8.2 Condition Analysis

After considering the design decisions outlined in Section 8.1, one major challenge was the automatic recognition of visual search episodes to identify the point in time when people need assistance. To achieve that, we first reviewed related research to better understand the properties of visual search and gain insights from previous approaches. In general, visual search is commonly defined as the act of looking for a target object among several distractors [Verghese, 2001]. During this process, attention is focused sequentially on each element of the visual scene, resulting in specific eye movement patterns [Findlay and Gilchrist, 1998]. The first one to analyze these patterns was Buswell [1935]. He showed that eye movements differ distinctively during a visual search task on an image compared to a free viewing task with no instructions. Several years later, Yarbus [1967] confirmed that the visual task indeed plays an important role in the observed scan paths and patterns. Since then, a lot of research has been conducted regarding the analysis of eye movement patterns in visual search tasks.

For instance, Castelhano et al. [2009] compared various eye movement measures, such as the fixation duration, saccade amplitude, or percentage of fixated area between a visual search and a memorization task. Thereby, 35 photographs of real-world scenes

were shown to the participants, who were asked to either search for a specific target or to memorize the objects in the corresponding image. As the results show, most of the examined features yielded distinctive values for each of the tasks, enabling the usage of a binary classifier for their detection. Similarly, Mills et al. [2011] examined the influence of visual search, memorization, scene rating, and free-viewing tasks on spatial and temporal characteristics of eye movements. For that, they conducted a user study with 53 participants and asked them to perform the four tasks on 67 images of computer-generated natural scenes. In compliance with previous works (e.g., Buswell [1935]; Castelhano et al. [2009]; Torralba et al. [2006]; Yarbus [1967]), they identified several eye movement characteristics, which can be used to distinguish between these tasks and are therefore considered in our work as well.

Based on these findings, Henderson et al. [2013] tried to infer the viewing task from eye movement measures with a naïve Bayes classifier. In their study, they recorded the gaze patterns of 12 participants while performing a scene memorization and a visual search task on photographs presented on a display. As the results show, they were able to identify the viewing task with an accuracy of up to 83%. Likewise, Coco and Keller [2014] used eye movements to classify three visual activities. These consisted of a visual search, a scene description, and an object naming task, which were performed on 24 photographs of indoor scenarios. Using a support vector machine (SVM), they achieved a maximum accuracy of 88% for the visual search task. Although these are promising results for the detection of visual search, most of the previous research has been conducted using static images on displays. Due to the restriction of the target area to a limited screen space compared to the wider view of a room or a building, these results might differ in real-world scenarios. Besides, head movements could also be valuable indicators to identify the visual search process in such a setting, but were previously not considered because of the restricted target area. For these reasons, we investigate the visual search task in a completely mobile and real-world scenario.

8.2.1 Signal Selection

In order to validate whether previous findings regarding visual search patterns on photographs and displays also translate to real-world settings, we conducted a short evaluation. For that, we employed the Pupil Pro head-mounted eye-tracker from Pupil Labs[2], which consists of a scene camera that records the user's field of view and an infrared camera that captures one of the user's eyes [Kassner et al., 2014]. With these sensors,

[2] `https://pupil-labs.com`

we were able to record people's gaze position and other related data, such as pupil dilation or blink frequency. Additionally, head movements were also captured indirectly and could be determined by detecting and tracking prominent features within the scene image. The relative position of these features was then compared across consecutive frames and used to calculate the user's head movements. For our qualitative analysis, a 77-year-old woman without visual or cognitive impairments was instructed to engage in four distinct activities. First, we hid an object and told the participant to search for it in her immediate surroundings. To compare this recording to other everyday activities, the subject was also instructed to read texts on a sheet of paper, watch videos on a television, and converse with another person.

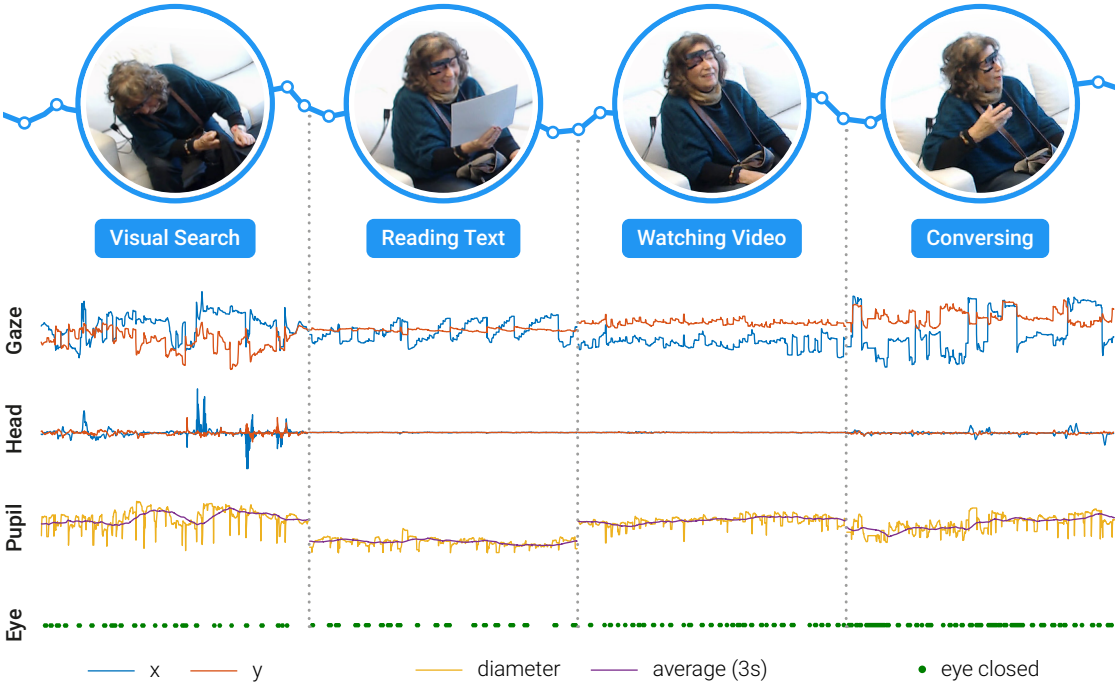


Figure 8.7: Comparison of sensor data during different activities.

Figure 8.7 provides an overview of the collected signals. As illustrated, the recorded activities can be differentiated by simply looking at the raw, unprocessed sensor data. Visual search and reading texts show the highest amount of saccades per second. However, the saccade distance is much smaller when reading. Searching and conversing have a similar saccade distance, although the saccades occurred less frequently while the subject was in a conversation. While head movements during reading and watching a video were almost nonexistent, some motion could be observed during the conversation. As expected, the user had to look around quite a lot during the visual search task. In contrast, pupil dilation remained almost constant when looking in a fixed direction

(i.e., during reading, watching a video, and conversing). A change in pupil size was noticeable when turning towards darker or brighter areas and was mostly seen during the searching task. One could also argue that the dilation changes show the user's distress as part of their affective processing. Based on these observations, we concluded that a differentiation between searching and other tasks in real-world scenarios could be achieved by calculating eye-tracking features like path length, average saccade distance, blink/fixation count, average blink time, or changes in pupil size. Additionally, head movement metrics, such as average distance along each axis or standard deviation of movement, could be used to train binary classifiers for the detection of visual search.

8.2.2 Tracking Device

Apart from Pupil Lab's eye-tracker employed in the previous section, several other commercially available devices, such as the Tobii Pro Glasses or the SMI Eye Tracking Glasses, could be used to record the targeted data. However, these devices are not capable of providing feedback to users and require an additional output component to support them, which could be too intrusive for older adults. Since no commercially available device fulfilled this requirement, we decided to build our own prototype. Through that, we were able to consider the special conditions and requirements of our elderly user group. For instance, the majority of older adults rely on prescription lenses. Therefore, it must be possible to wear the device in addition to glasses without disturbing the user.

Figure 8.8: Google Glass-based eye- and head-tracking device.

This requirement also implies that the device should be as small and lightweight as possible. Furthermore, the prototype should not impact the mobility of the users and must work in a completely mobile setting to increase the acceptance of this technology.

Considering these requirements, we decided to use the Google Glass as the basis for our prototypical device since it was one of the lightest head-mounted displays at the time and could be worn on top of prescription lenses. Besides that, it already has a built-in accelerometer and gyroscope sensor that can be used to track users' head movements. In order to record the eye movements as well, we created a custom mount with a 3D printer and attached a small infrared camera (30 Hz, 640×480 resolution) taken from a Pupil Labs eye-tracker to the frame of the smart glass, as shown in Figure 8.8. The camera is connected to a Raspberry Pi 2, which can either record the eye video locally or stream the data to another processing unit. Afterwards, an algorithm based on the open-source Haytham Gaze Tracker is applied to the video stream of the eye camera to determine the pupil position. Combined with the video from the scene camera of the Google Glass, we receive the same data as with a regular head-mounted eye-tracker, but with the added benefit of being able to support the user through instructions and visualizations on the head-mounted display.

8.2.3 Data Collection

Since the goal was to support visual search episodes of older adults caused by memory lapses, we conducted a large-scale study to collect test and training data for the automatic recognition of those situations. In order to obtain a rich dataset for user-independent machine learning models, we recruited 30 participants aged between 65 and 80 years ($avg = 71,7$) with a female ratio of 50%. During the study, each subject performed several activities, including the visual search for objects, while being equipped with our eye- and head-tracking device. Even though the study was not exclusively designed for the sole detection of visual search, the recorded data can be used for this purpose because all other tasks were similar to day-to-day activities and thus can serve as a comprehensive baseline.

Tasks

Overall, the study involved five tasks with distinct objectives and conditions. However, since this section focuses on detecting visual search, we mainly concentrate on the search scenario and only give a brief overview of the other tasks. Before each task, participants received detailed instructions and afterwards had to fill out a questionnaire

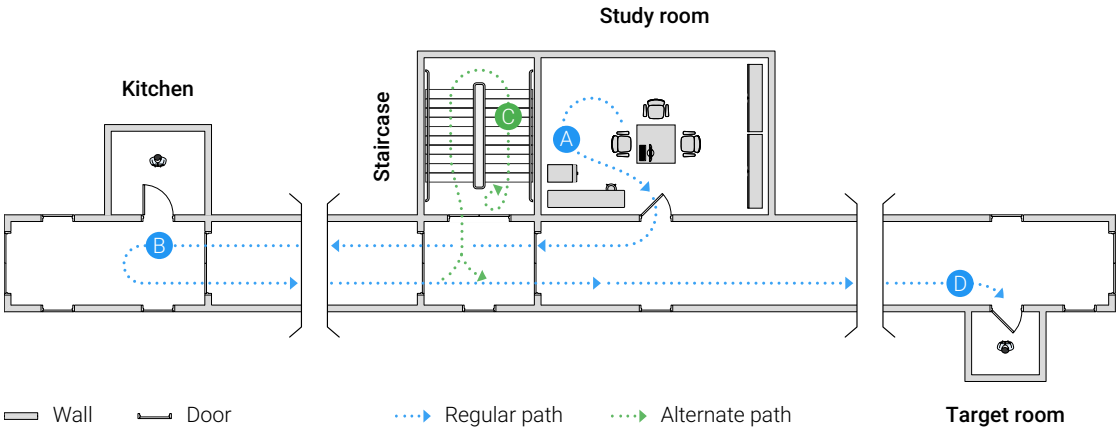


Figure 8.9: Overview of the study location.

regarding their experiences. In the first task, each participant was instructed to enter general demographic information into a smartphone app. For that, the system vocally asked the subjects basic questions, which they could answer using natural language. Due to the auditory nature of the interaction, users could look around freely during this task. In the second one, participants were asked to read and write texts on sheets of paper. After a fixed amount of time, an experimenter called them on a telephone and told them four terms, which they should memorize and recall at the end of the session. Participants were then instructed to work with a computer for the following two tasks. In task three, each user was asked to observe the screen for a specific visual condition and had to press a button every time it occurred. Similarly, in task four, an object was shown in the center of the screen for a few seconds while the users had to click on the corresponding button matching its condition. Between those tests, two videos were shown to the users for relaxation purposes.

Finally, the last task involved the visual search activity, which was investigated in the following two scenarios: the search for keys and the search for rooms. The reason why we selected these scenarios is that we wanted to capture the characteristics of visual search in a wide spectrum of occurrences, ranging from the search of a small item in a limited area to the search of a location in an open space. In order to create a realistic setting for both conditions, we told participants shortly before the end of the previous task that we had to leave them to prepare the study for the next participant and that they should meet us in a certain room. Additionally, they were asked to lock the door with a key located in one of the closets shown in Figure 8.9 Ⓐ once they were finished. However, the hidden key did not match the lock on the door. This caused some subjects to continue the search even after finding the key. Eventually, after a

certain amount of time, every participant gave up and started to search for the room in which they were supposed to meet the experimenter. Based on the room number we gave them, they assumed that it was located at the end of the hallway (Figure 8.9 Ⓑ), but upon arrival, they realized that there was no room with that number. Instead, they only found a person standing in the kitchen nearby, who they asked for the right way. The person was instructed to tell the participants the number of the correct room (Figure 8.9 Ⓓ) and with that information most of them were quickly able to locate it. In spite of knowing the room number, a few users got completely lost and used the staircase to search for the room on different floors (Figure 8.9 Ⓒ), which resulted in even more realistic search recordings. Nevertheless, all participants eventually found the target room, which marked the end of each session.

Sensor Setup

For the user study, we employed a completely mobile and wearable sensor setup. At the core of this setup was the mobile signal processing framework introduced in Chapter 6. It enabled us to interface with and extract data from multiple sensing devices in a synchronized fashion. Moreover, since SSJ has been designed and built specifically for mobile devices, participants were able to freely move around the room and building, increasing the authenticity of the search task. While our custom eye- and head-tracking device would have been sufficient to record the necessary data for the detection of visual search, additional sensors were used to recognize the other situations from our study. As a result, the complete setup consisted of two smartphones (Samsung Galaxy S4), our Google Glass-based eye-tracking system, a Raspberry Pi 2 and an Empatica E3 sensor armband, as shown in Figure 8.10. All devices were synchronized to each other and communicated via Wi-Fi. In order to avoid compromising the mobility of the system, a Wi-Fi hotspot was created using one of the two smartphones. The other one was operated by a researcher to control the entire sensor setup, which included synchronously starting and stopping the recording on all devices, triggering the calibration phase of the eye-tracker, and completely shutting down all involved devices. Moreover, the researcher also used this smartphone to label the start and end of the individual study tasks. The second smartphone was running an SSJ pipeline, which extracted data from the device's internal inertial measurement unit (IMU) and microphone, as well as the Bluetooth-connected Empatica E3 armband, and stored it to the local SD card. Similarly, a second SSJ pipeline was running on the Google Glass, which recorded IMU, audio, and video data. The eye-tracking camera data was captured using a custom program running on a Raspberry Pi 2.

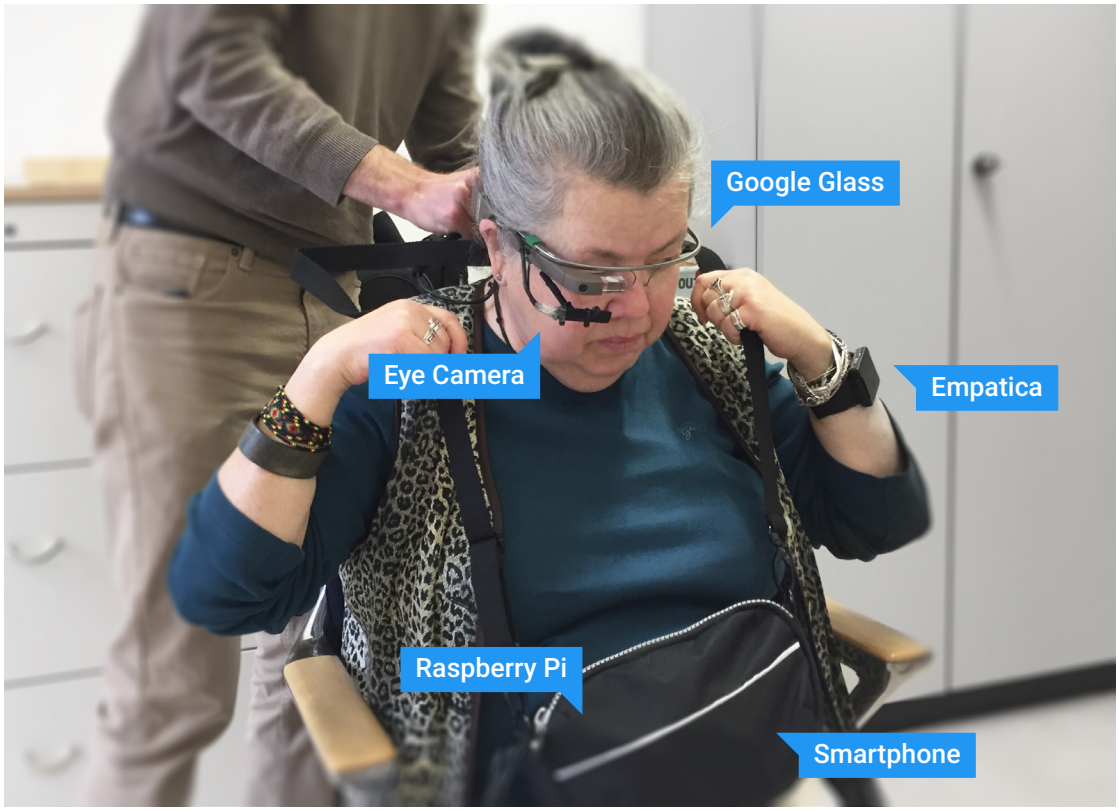


Figure 8.10: Participant wearing the sensor setup.

8.2.4 Data Analysis

Following the user study, we first analyzed the recorded data to prepare it for our classification approach. This step was necessary to ensure that the sensors worked correctly in all sessions and provided reliable data in each case. Otherwise, false or missing data streams could have negatively impacted the classification performance. Therefore, incomplete and corrupt session recordings had to be identified and removed before the data could be used for the automatic detection of visual search.

Signal Quality

Since the device used to record the gaze data consisted of a camera pointed at the participant's eye and another camera capturing the scene view (see Section 8.2.2), a calibration had to be conducted to map the pupil position from the eye camera to a gaze point in the field of view. After calibration, the device needed to stay in the same position to maintain the calculated mapping. However, some participants treated the device like a pair of glasses and readjusted its position multiple times after calibration. In most cases, this only led to a shifted gaze point, which left most feature calculations unaf-

fected. In some extreme cases, there was so much readjustment that the eye was no longer visible in the camera's view, which led to unusable data in later parts of these recordings. Another problem occurred because some users assumed the study was concluded after filling out the last questionnaire following the fourth task. In these cases, they took the eye-tracking device off before beginning the search task so that no data could be recorded. One participant even required too much time to complete the tasks, which led to the depletion of the Google Glass battery after one hour and forty minutes, resulting in incomplete data for the session. For these reasons, eight recordings had to be discarded, leaving 22 usable sets of data (*avg* = 71.2 years, 50% female).

Task Annotation

Based on the recorded audio and video streams, we refined the task annotations for every remaining session. During this process, the first four tasks were labeled as "Baseline", while both key and room search were annotated as "Search" to create a binary classification problem. The annotation for the key search began once the participants approached the closets and ended as soon as they left the room and closed the door. This event also marked the start of the room search, which continued until the users arrived at the target location. We did not exclude certain phases from these segments, such as the short conversations when asking for the right way, because even during these periods, participants were still looking around and trying to find the room. The resulting completion times for both search tasks are summarized in Figure 8.11.

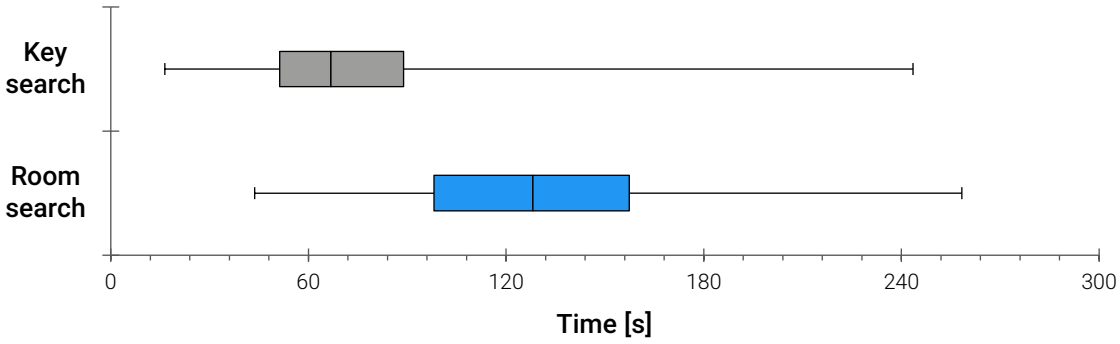


Figure 8.11: Visual search task durations.

To extend the baseline even further, we labeled one instruction phase where an experimenter explained an upcoming task to the participant with "Baseline" as well since it resembled a regular conversation. Besides that, we also included one questionnaire phase, which was similar to a common reading and writing task. As a result, the baseline consisted of the following day-to-day activities: reading, writing, speaking out loud,

talking on a telephone, memorizing terms, holding a conversation, working on a computer, and watching videos. We used this annotation set in our classification approach to accomplish the automatic detection of visual search with machine learning techniques. Additionally, we created a second annotation set with the same baseline but with individual labels for the key search and the room search. This separation enabled us to examine if there are any differences between these two scenarios.

8.2.5 Model Training

For the automatic detection of visual search, we selected a support vector machine (SVM) as classifier (*linear kernel*, $C = 1$, $\varepsilon = 0.1$, $\nu = 0.5$, $\gamma = 0.01$) since it was one of the most popular algorithms in the field of machine learning at the time [He and Jin, 2009] and also works efficiently on recent generations of mobile devices [Damian et al., 2016]. This was important because we intended to use the resulting classification model with our mobile eye- and head-tracking device in an online scenario. Although artificial neural networks might have yielded even better classification performances, they would have required much more training data and were therefore not considered. All evaluations were conducted using the leave-one-user-out (LOUO) method to achieve a subject-independent classification model. The procedure involves training classifiers with data from all users except one and performing tests on the excluded user. This process is repeated for every participant, and afterwards, the average values across all iterations are taken as the result. A key benefit of this method is that it simulates a real-time analysis based on the recorded data since the trained classifiers are always tested with signals from an unknown user, which is also the case in an online classification scenario. For the implementation of features, model training, and evaluation we used Wagner et al.'s [2013] Social Signal Interpretation (SSI) framework. It provides various tools to support all phases of machine learning and enabled us to utilize the computational resources of our workstations and servers to accelerate this process. Due to the prominent conceptual and technical connection between SSI and SSJ, the recorded data could be used for training without requiring any conversions, and the resulting classification models could be directly integrated into mobile processing pipelines.

Feature Extraction

All gaze features were based on the raw sensor data from our mobile eye- and head-tracking device. For a given window length, we processed the data and calculated the fixation duration, saccade duration, and saccade length. In our case, these metrics are

defined as follows: *Fixation duration* is the time in seconds of a single fixation, *saccade duration* is the time in seconds between two subsequent fixations, and *saccade length* is the Euclidean distance in pixels between two subsequent fixation points. For each of these three metrics, we then computed the mean, minimum, maximum, median, sum, standard deviation, skew, kurtosis, and range values, which were commonly used for visual search detection on displays and general activity recognition in previous works [Bixler and D'Mello, 2014; Castelhano et al., 2009; Coco and Keller, 2014; Greene et al., 2012; Henderson et al., 2013; Mills et al., 2011; Torralba et al., 2006]. Additionally, we applied a wordbook analysis proposed by Bulling et al. [2011] to identify repetitive eye movement patterns. Thereby, the eye movement direction of each saccade is mapped to one of 24 discrete characters. Depending on the length l of the wordbook, each saccade sequence is encoded into a string of l characters and added to the wordbook. If a pattern is already included, its occurrence count is increased by one. Similar to Bulling et al. [2011], we used four wordbooks with $l \in \{1,2,3,4\}$ and calculated the size, maximum, range, mean, and variance of all occurrence counts in each wordbook.

Furthermore, we analyzed the spatial distribution of fixations by computing the fixation dispersion, fixation coverage, and number of fixation groups. The *fixation dispersion* is calculated using the root mean square of the Euclidean distances between each fixation and the average position of all fixations within the current window [Bixler and D'Mello, 2014]. For the *fixation coverage*, we draw a circle with radius r based on the fixation duration around each fixation point and compute the ratio between covered area and total field of view [Castelhano et al., 2009]. Based on the fixation map from the previous feature, we identified the connected areas that represent *fixation groups* and counted their occurrences [Sadasivan et al., 2005]. Additionally, we calculated the number of saccades, fixations, and blinks as well as the ratio between fixation and saccade duration [Bixler and D'Mello, 2014]. Combined with six movement-independent features such as sum, mean, and variance of the blink duration and pupil size change [Bulling et al., 2011], this resulted in a total of 60 gaze features.

For the extraction of head movement features, we directly used the raw accelerometer and gyroscope data from the Google Glass. Since both sensors share the same sample rate and provide data for each axis (x,y,z), we applied the same features for both of them, as suggested by Rahman et al. [2015]. While most features were computed for each individual axis, some were based on pairs of axes or even factored in all three of them. The features calculated for each axis included the mean, variance, standard deviation, skew, kurtosis, interquartile range, mean absolute deviation, root mean square, energy, and frequency domain entropy values, which were previously used for activity

recognition purposes [Altun and Barshan, 2010; Bao and Intille, 2004; Chen et al., 2008; Huynh and Schiele, 2005; Lara and Labrador, 2013; Ravi et al., 2005]. Additionally, we applied a 1D Haar-like filter similar to Hanai et al. [2009]. For that, a small sliding window within the actual window is used to calculate the value differences between the left and right halves of each sub-window. Due to the variable filter parameters, this feature has shown promising results for various classification problems [Hanai et al., 2009] and was therefore adopted in our work as well.

Furthermore, we calculated the crest factor, spectral flux, spectral centroid, and spectral roll-off features, which were mainly used for classifying audio signals in the past [Lu et al., 2009; Yatani and Truong, 2012]. However, as demonstrated by Rahman et al. [2015], those features are also suitable for differentiating between activities based on acceleration and orientation data. For each pair of axes $\{(x,y),(y,z),(z,x)\}$, we then applied a biaxial 1D Haar-like filter [Hanai et al., 2009] and calculated the correlation between the corresponding axes. The *correlation* is determined by dividing the covariance by the product of the standard deviations and is especially helpful for detecting activities that involve movements in a single direction [Ravi et al., 2005]. Finally, we computed the *signal magnitude area*, which is defined as the sum of the absolute acceleration values from each of the three axes [Khan et al., 2010b]. It was used because it has proven to be a suitable indicator to distinguish between stationary and movement-related activities [Khan et al., 2010a]. Overall, 52 features were calculated for each of the two sensors, thus resulting in a total of 104 head movement features.

Feature Window Analysis

In order to explore the impact of window lengths on classification performance, we generated all features for different window sizes (1-10 seconds) and measured the accuracy of each feature set. For every window length, we also varied the overlap between each window from 0 to 90%. As it turns out, our results did not reveal an overlap ratio with significantly better performance than others. However, since previous works have shown the most success with a 50% overlap between each window, we selected it in our approach as well [Bao and Intille, 2004; Chen et al., 2008; He and Jin, 2009; Ravi et al., 2005]. Another interesting finding from our results is that the classification accuracy increased almost linearly with growing window sizes, as shown in Figure 8.12. Therefore, using a longer window size would make sense to achieve the highest possible detection rates. However, since the goal was to recognize visual search behavior in real-time, we could not use an arbitrarily large window as it would have slowed down the reaction

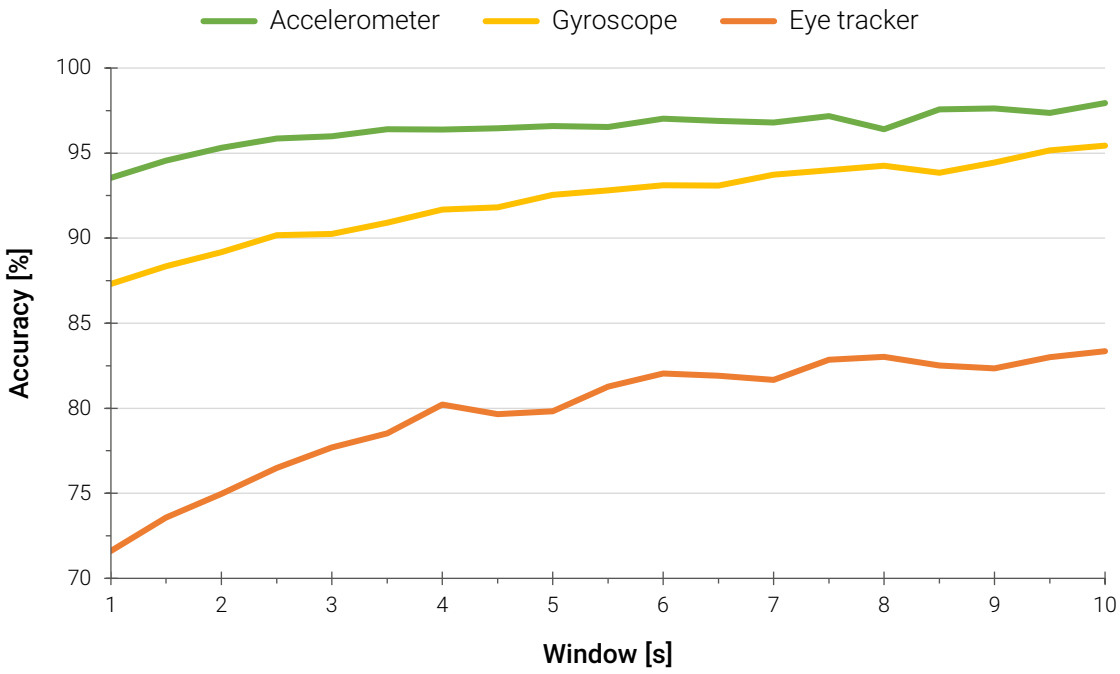


Figure 8.12: Relation between window length and accuracy (50% window overlap).

time of our approach. Instead, we needed to make a compromise between window size and detection rate, which is the reason why we chose a window length of four seconds.

Fusion and Feature Selection

After selecting a fixed window size, we applied various fusion techniques to combine the feature sets from the accelerometer, gyroscope, and eye-tracking sensors. During *early fusion* (feature level), the features from each modality are concatenated into a single feature vector before the classifier is trained [Snoek et al., 2005]. In contrast, late fusion is applied after individual classifiers for every modality have been trained by combining their predicted scores [Kächele et al., 2015]. For that, several methods can be used, including AdaBoost, Borda count, cascading specialists, Dempster-Shafer, stacked generalization, weighted majority voting, or even simple rules, such as the sum, minimum, maximum, median, and product [Kittler et al., 1998; Knauer and Seiffert, 2013; Lingenfelser et al., 2011, 2010]. In our case, the stacked generalization approach yielded the highest accuracy of those methods and is therefore used to achieve all further late fusion results. While both early and late fusion usually result in higher detection rates compared to the classification based on individual modalities, they also increase the required dimensionality of the input data. Consequently, all 164 features would need to be computed at the same time, which could cause performance bottlenecks in

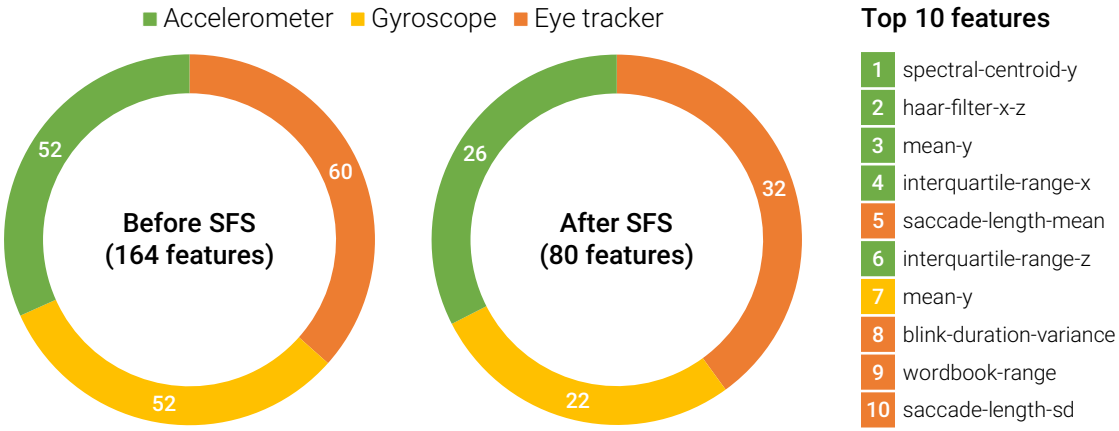


Figure 8.13: Feature composition for early fusion.

an online scenario. However, since not all features were equally helpful in detecting the visual search activity, we used the sequential forward selection (SFS) method to reduce the number of features and hence the required computational cost associated with it [Webb, 2002]. The feature selection was applied to the concatenated vector of all features for the early fusion and to each individual feature set for the late fusion. Using this technique, we were able to reduce the number of required features by more than 50%. As shown in Figure 8.13, the feature distribution across all sensors stayed nearly the same after applying the feature selection, which indicates the importance of using a multimodal approach.

Classification Results

The final results of our visual search detection approach are based on the reduced feature sets after applying the SFS feature selection. In compliance with all previous evaluations, we used the leave-one-user-out method to train and test our SVM models several times. Table 8.2 shows the average accuracy, precision, and recall values for ev-

Source	Accuracy	Precision	Recall
Accelerometer	97.39%	97.65%	97.11%
Gyroscope	92.18%	94.14%	89.97%
Eye-tracker	81.59%	82.67%	79.93%
Early fusion	**97.55%**	**98.11%**	96.97%
Late fusion	97.39%	97.47%	**97.29%**

Table 8.2: Classification results after feature selection for baseline vs. search.

ery modality, as well as the results after early and late fusion. Overall, early fusion yielded the highest accuracy with 97.55%, closely followed by late fusion with a value of 97.39%. From the individual modalities, the acceleration showed the highest accuracy, which is on par with the late fusion and only slightly lower than the early fusion results. Although this might lead to the assumption that the accelerometer alone can be sufficient for visual search detection, combining multiple modalities is more robust against signal fluctuations of individual sensors and more reliable in real-world applications. Surprisingly, the gyroscope model yielded a five percent lower accuracy compared to the accelerometer, even though both are based on the same initial feature set. The eye-tracking model resulted in the lowest accuracy of 81.59%, which can be mostly attributed to the signal quality, as described in Section 8.2.4. Generally, all modalities and fusion methods showed high precision and recall values. This means that in cases where visual search was detected, it was usually correct (precision) and that almost all instances of visual search were recognized as such (recall).

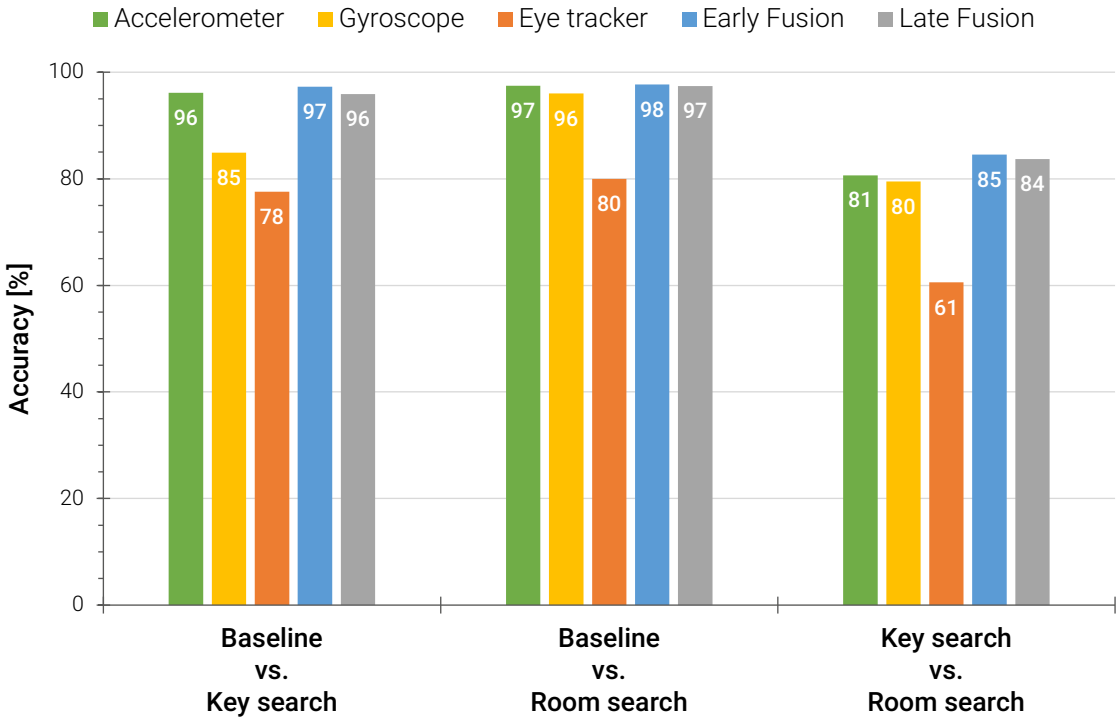


Figure 8.14: Additional results for different search scenarios.

In addition to the general detection of visual search, we also investigated whether there are any differences when recognizing either of the two search scenarios from our user study and whether it is possible to distinguish between them. The results of this analysis are summarized in Figure 8.14. Interestingly, when trying to detect the key or room search individually, we achieve similar accuracies compared to the general detection

of visual search. The only notable difference occurs in the accuracy of the gyroscope model, which is seven percent lower for the key search and four percent higher for the room search. This could indicate that the head orientation is more distinctive when searching for large objects that might not fit into the field of view and require more head rotations than when looking for smaller items, such as keys. Using the same features as before, we then tried to distinguish both scenarios from each other. As expected, the results were lower compared to the previous evaluations. However, we still achieved a reasonably high accuracy of 84.53% using the early fusion method, which could indicate that the target object type might have an influence on the search behavior. Additionally, information on the target object type would enable the system to provide more specific assistance to users after detecting visual search.

8.3 System Overview

Based on the results and findings described in the previous sections, we designed and developed a completely mobile system to support visual search episodes of older adults caused by memory lapses. Since explicitly requesting support in these stressful situations, especially under time pressure, can be too overwhelming and exhausting for our intended user group, we decided that the system should automatically identify these

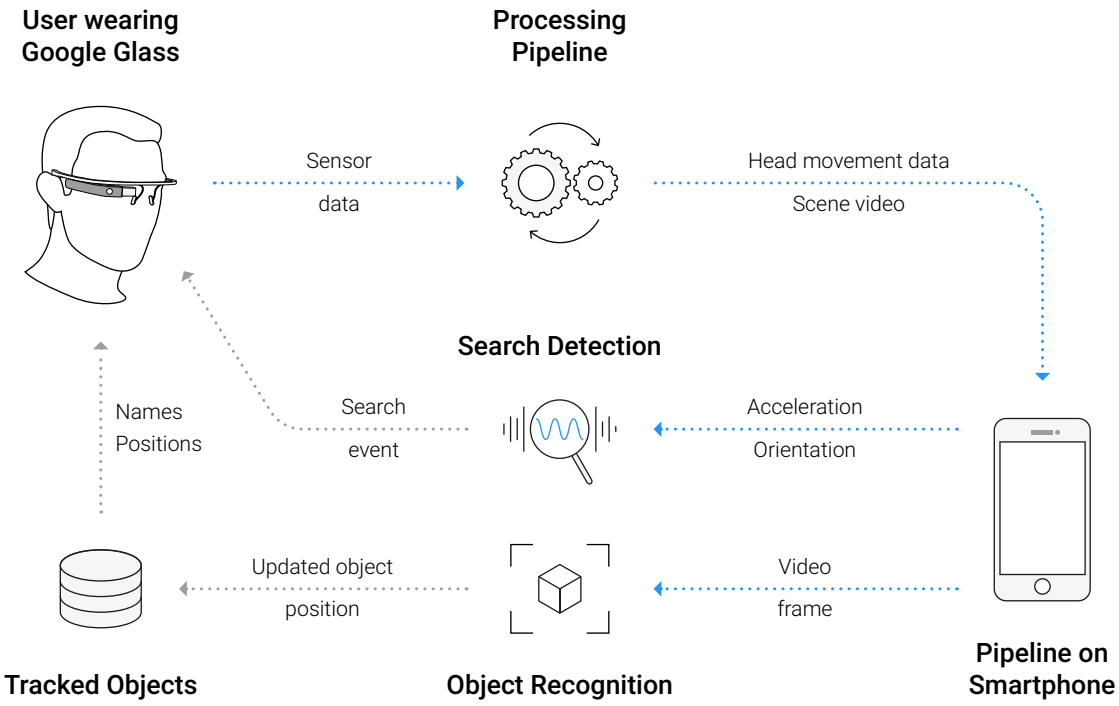


Figure 8.15: Architecture of the memory augmentation system.

critical conditions and proactively offer assistance. To achieve that, we continuously analyze peoples' head movements with the integrated accelerometer and gyroscope sensors of the Google Glass head-mounted display. Afterwards, the classification model described in the previous section is used to recognize the search behavior and identify appropriate moments when assistance is needed. As soon as the visual search activity has been detected, a list of tracked objects is shown on the head-mounted display and the person is prompted whether they are looking for one of them (Figure 8.6). Once the user selects an item, an image of the desired object at the location where it was last seen is displayed. The idea behind this approach is to trigger episodic memories about the item's last known position and to help users find the right location through contextual information about the target area captured within the images.

Figure 8.15 provides an overview of the system architecture. As illustrated, the system consists of two distributed processing pipelines: the first runs on the Google Glass and the second on a smartphone. To capture the necessary head movement signals, we first added an `AndroidSensor` component to the pipeline on the Google Glass and connected it to an accelerometer and a gyroscope `AndroidSensorChannel`. This allowed us to access the data from the integrated sensors in our pipeline. We also limited the sample rate to 40 Hz for both channels (lines 7 and 12) to reduce the required processing power, increase battery life, and prevent the device from overheating.

```
1  // Create android sensor to access integrated sensors
2  AndroidSensor androidSensor = new AndroidSensor();
3
4  // Create accelerometer channel with sample rate of 40 Hz
5  AndroidSensorChannel accChannel = new AndroidSensorChannel();
6  accChannel.options.sensorType.set(SensorType.LINEAR_ACCELERATION);
7  accChannel.options.sampleRate.set(40);
8
9  // Create gyroscope channel with sample rate of 40 Hz
10 AndroidSensorChannel gyrChannel = new AndroidSensorChannel();
11 gyrChannel.options.sensorType.set(SensorType.GYROSCOPE);
12 gyrChannel.options.sampleRate.set(40);
13
14 // Add components to pipeline
15 pipeline.addSensor(androidSensor, accChannel);
16 pipeline.addSensor(androidSensor, gyrChannel);
```

Additionally, we added a `CameraSensor` and a `CameraChannel` to capture the video stream of the user's field of view. We set the resolution of the camera image to 320×240

pixels (lines 19-20) and reduced the frame rate to 1 Hz (line 24) for the same reasons we limited the sample rates of the other components.

```
17 // Create camera sensor with resolution of 320x240 pixels
18 CameraSensor cameraSensor = new CameraSensor();
19 cameraSensor.options.width.set(320);
20 cameraSensor.options.height.set(240);
21
22 // Create camera channel with sample rate of 1 Hz
23 CameraChannel cameraChannel = new CameraChannel();
24 cameraChannel.options.sampleRate.set(1);
25
26 // Add components to pipeline
27 pipeline.addSensor(cameraSensor, cameraChannel);
```

The data from all three channels was then passed to a `BluetoothWriter` consumer component, which handled the communication with the other pipeline on the smartphone. We configured the component to act as a client and connected it to a server device with the specified MAC address (lines 30-31). However, instead of using a fixed string identifier, we implemented a custom method that dynamically resolved the address of the paired smartphone. In this regard, we also adjusted the frame size of the `BluetoothWriter` to a one-second window without overlap (line 37: `frame=1`, `delta=0`), which reduced the Bluetooth traffic by accumulating the data and sending it as a packet once per second.

```
28 // Create bluetooth writer as client
29 BluetoothWriter btWriter = new BluetoothWriter();
30 btWriter.options.connectionType.set(BluetoothConnection.Type.CLIENT);
31 btWriter.options.serverAddr.set(getSmartphoneMacAddress());
32 btWriter.options.connectionName.set("ssj_stream");
33
34 // Add component to pipeline and set sensor channels as input
35 pipeline.addConsumer(btWriter, new Provider[] {
36     accChannel, gyrChannel, cameraChannel
37 }, 1, 0);
```

On the smartphone side, a `BluetoothReader` component received the signals and passed them to three separate channels to reconstruct the original sensor data. This step involved specifying the properties of the data streams contained in the transferred packets, including data type, dimension, sample rate, and sample number (lines 9-12). Due to space restrictions, the following listing only shows the code for one channel as an example, while the complete pipeline is available in Appendix B.

```
// Create bluetooth reader as server to read data from Google Glass
BluetoothReader btReader = new BluetoothReader();
btReader.options.connectionType.set(BluetoothConnection.Type.SERVER);
btReader.options.connectionName.set("ssj_stream");

// Create channel to read acceleration from Google Glass
BluetoothChannel glassAcc = new BluetoothChannel();
glassAcc.options.channel_id.set(0);
glassAcc.options.dim.set(3);
glassAcc.options.type.set(Cons.Type.FLOAT);
glassAcc.options.sr.set(40);
glassAcc.options.num.set(40);

[...] // Full code available in Appendix B

// Add components to pipeline
pipeline.addSensor(btReader, glassAcc);
```

8.3.1 Visual Search Detection

For the detection of visual search, we implemented a new transformer component called `AccelerationFeatures` that receives the raw signals along each axis (x, y, z) as input and calculates the feature values outlined in Section 8.2.5 as output. While the component can compute all features that were initially considered, we configured it to calculate only the subset with the best classification performance by default. The features for both accelerometer and gyroscope were computed on a four-second sliding window, which was updated with new data every second (`frame=1`, `delta=3`).

```
float frameSize = 1;
float deltaSize = 3;

// Create transformers to calculate head movement features
AccelerationFeatures accFeatures = new AccelerationFeatures();
AccelerationFeatures gyrFeatures = new AccelerationFeatures();

// Add components to pipeline
pipeline.addTransformer(accFeatures, glassAcc, frameSize, deltaSize);
pipeline.addTransformer(gyrFeatures, glassGyr, frameSize, deltaSize);
```

Since the early fusion approach yielded the highest accuracy, we concatenated the resulting feature vectors and used them as input for the SVM classification model trained in Section 8.2.5. For this component, we used the same frame size but reduced the delta time to zero so that the model is only executed on the newest iteration of feature data.

```
28 // Create SVM model and select model file
29 SVM searchModel = new SVM();
30 searchModel.options.file.set(new FilePath("/model/search.trainer"));
31
32 // Create classifier and select model
33 ClassifierT searchClassifier = new ClassifierT();
34 searchClassifier.setModel(searchModel);
35
36 // Add component to pipeline and use features as input
37 pipeline.addTransformer(searchClassifier, new Provider[] {
38     accFeatures, gyrFeatures
39 }, frameSize, 0);
```

Finally, we added a `ThresholdEventSender` to the pipeline and configured it to trigger an event if the classifier detects visual search behavior with at least 80% confidence (line 42). In this case, an external `SearchHandler` component gets notified and sends a signal to the Google Glass, where the object selection screen is displayed (Figure 8.6).

```
40 // Create event sender and set threshold to 0.8
41 ThresholdEventSender resultSender = new ThresholdEventSender();
42 resultSender.options.thresin.set(new float[] {0.8f});
43
44 // Add component to pipeline and use classification result as input
45 pipeline.addConsumer(resultSender, searchClassifier);
46
47 // Add external event receiver to output event channel
48 EventChannel resultChannel = resultSender.getEventChannelOut();
49 resultChannel.addEventListener(SearchHandler.getInstance());
```

8.3.2 Visual Search Support

Before an item of interest can appear in the selection screen on the Google Glass, it must first be added to the list of tracked objects. This process involves providing a suitable and meaningful name to describe the new item. Additionally, the user is asked to place the object at typical locations and is instructed to look at it from different viewing angles and positions. During this step, a video of the person's field of view containing the item is recorded with the camera of the Google Glass. Afterwards, the video is split into individual frames, which are then used to retrain the last layer of a neural network-based object detection model. For that, the method proposed in cooperation with Seiderer et al. [2018] was applied to perform the training process directly on the mobile device. Upon completion, the updated model was used to analyze the video stream of the camera and identify the tracked items. For that, several components were added to perform

the necessary preprocessing steps. These included converting the video stream into an appropriate encoding format, resizing the image to fit the input dimensions of the object detection model, and normalizing the pixel color values.

```
50 // Create transformer to convert encoding format from NV21 to RGB
51 NV21ToRGBDecoder nv21ToRGBDecoder = new NV21ToRGBDecoder();
52 pipeline.addTransformer(nv21ToRGBDecoder, glassImage);
53
54 // Create transformer to resize image
55 ImageResizer imageResizer = new ImageResizer();
56 imageResizer.options.size.set(224);
57 pipeline.addTransformer(imageResizer, nv21ToRGBDecoder);
58
59 // Create transformer to normalize image pixel values between -1 and 1
60 ImageNormalizer imageNormalizer = new ImageNormalizer();
61 pipeline.addTransformer(imageNormalizer, imageResizer);
```

Subsequently, the processed images were passed to a `Classifier` component to perform the object detection task. As soon as it recognized one of the configured items, an event was sent to the `ImageWriter`, which saved the corresponding video frame and assigned it to the detected object. Using this technique, a tracked item is always associated with an image from the last time it appeared in the person's field of view.

```
62 // Create object detection model and select model file
63 TFLite objModel = new TFLite();
64 objModel.options.file.set(new FilePath("/model/obj_detection.trainer"));
65
66 // Create classifier and select model
67 Classifier imageClassifier = new Classifier();
68 imageClassifier.setModel(objModel);
69 EventChannel imageChannel = imageClassifier.getEventChannelOut();
70
71 // Create image writer triggered by events
72 ImageWriter imageWriter = new ImageWriter();
73 imageWriter.options.triggeredByEvent.set(true);
74
75 // Add components to pipeline
76 pipeline.addConsumer(imageClassifier, imageNormalizer);
77 pipeline.addConsumer(imageWriter, nv21ToRGBDecoder);
78 pipeline.registerEventListener(imageWriter, imageChannel);
```

8.4 Evaluation

The study described in this section was conducted in cooperation with the Geriatrics Research Group at Charité – Universitätsmedizin Berlin as part of the Glassistant project. At the beginning of the project, an external advisory board consisting of ten experts from various fields was established to monitor its ethical, legal, and social implications (ELSI). Apart from periodic feedback and general recommendations, they also reviewed the proposals and approved the designs of all performed studies.

In order to evaluate the final system, we conducted a study with eight participants (50% female) aged between 70 and 81 ($M = 74.3$) years. To ensure that participants could relate to the visual search problem, we defined subjectively perceived memory issues and a minimum age above 65 years as criteria for inclusion. However, candidates with severe affective or cognitive disorders were excluded during recruitment.

8.4.1 Procedure

Since the main goal of this study was to evaluate the proposed system's effectiveness, usability, and potential influence on the mental workload of older adults, we applied a within-subjects counterbalanced design, where each participant was exposed to two conditions: a visual search task with and without the system's assistance. At the begin-

(a) Target object (stopwatch) **(b)** Part of the search area (laboratory)

Figure 8.16: Impressions from the user study.

ning of each session, we briefly explained the study procedure to participants and gave them an introduction on how to use the system. After being familiar with the device, we began with the first visual search task. For that, the stopwatch shown in Figure 8.16a was hidden at one of two predefined locations in a large laboratory (Figure 8.16b), and the subjects were asked to find the object. In both conditions, we instructed them to wear the device to prevent any external influences that might be caused by it (the device was turned off in the *without system* condition). Furthermore, we measured the required time to find the object and told participants to complete the tasks as quickly as possible to induce the time pressure component of a real-world visual search task.

After finding the object in the first condition, participants were instructed to fill out the first part (magnitude of load) of the NASA-TLX (Task Load Index) questionnaire by Hart and Staveland [1988], which is one of the most widely used instruments to measure the subjective workload of different tasks. This process was then repeated for the second condition as well. Upon completing both conditions, participants were asked to fill out the second part (sources of load) of the NASA-TLX questionnaire to rate the relative contributions of the six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Additionally, participants were instructed to rate the usability of the system by filling out the System Usability Scale (SUS) questionnaire [Brooke, 1996]. At the end of each session, we also asked them two questions regarding their rating of the general system, which they could answer on a 5-point Likert scale.

8.4.2 Results

In general, all participants successfully completed both tasks of our experiment in one session with an average duration of about five minutes. However, in two cases (users 4 and 7), the Bluetooth connection between the Google Glass and the smartphone got disrupted due to signal interferences. Although the pipelines were configured to automatically reconnect in such events, the system could not be used until the connection was reestablished a few minutes later. These disturbances left a negative impression of the system on affected users, which also impacted their subjective ratings, as shown in Figure 8.17. While the average SUS score for users without issues was 79.2, which is a relatively good result according to Bangor et al. [2008], both affected individuals rated it with 42.5 and 40.0, respectively. For these reasons, we excluded their ratings from the NASA-TLX results to only compare instances where the system behaved as intended.

Figure 8.18 shows an overview of the subjectively rated task load across the six subscales of the NASA-TLX questionnaire. Among the current set of participants, no sig-

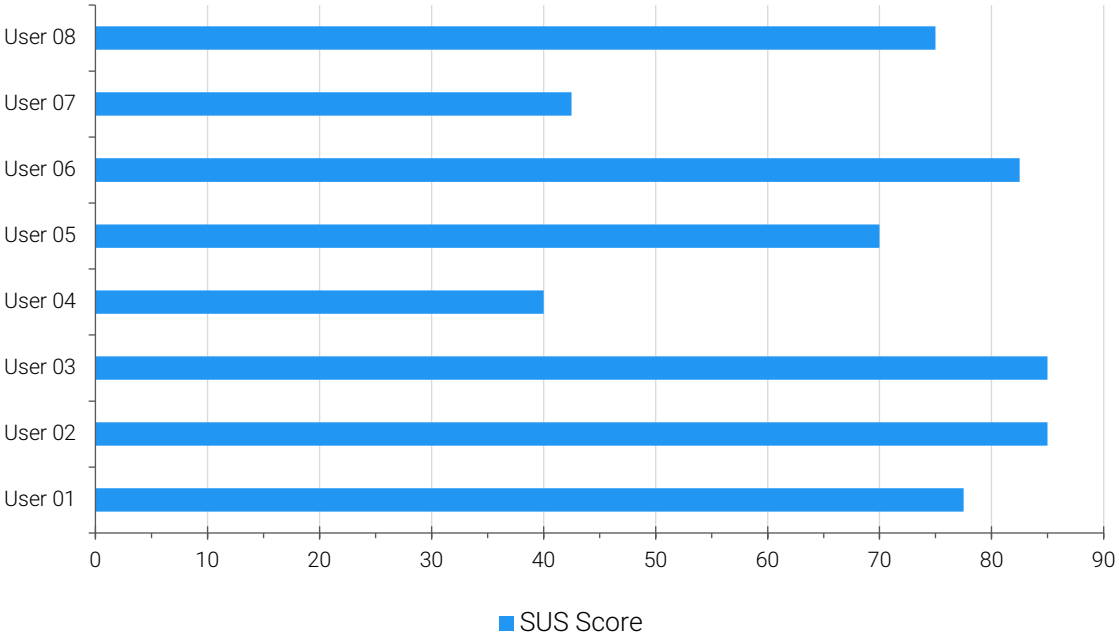


Figure 8.17: Results of the SUS questionnaire.

nificant differences between the two conditions could be determined. However, the task load for the condition without the system was rated higher or equal in five out of six dimensions, which is also reflected in the average task load index of 21.6 (with the system) versus 27.1 (without the system). The only area where the system caused a higher load was the mental demand, although the absolute rating of around 25.0 in this case was still relatively low. Apart from these standardized questionnaire results, the system as a whole was rated on average $4.0/5$, and its general helpfulness even scored $4.8/5$ points. When including the users with Bluetooth connection issues, these values slightly change to 3.8 for the overall rating and 4.3 for the system helpfulness.

8.5 Discussion

Overall, the study yielded valuable insights and positive results. As outlined in the previous section, the system proved useful in supporting the visual search for misplaced objects of older adults with memory impairments. Based on the results of the NASA-TLX questionnaire, we observed that our approach reduced the load experienced during the search task across almost all dimensions. Although the mental demand was higher while using the system, this trade-off seems to be acceptable for our participants according to their ratings. One explanation for this circumstance could be that the load only increased by around eight points, which is relatively small in absolute terms. Ad-

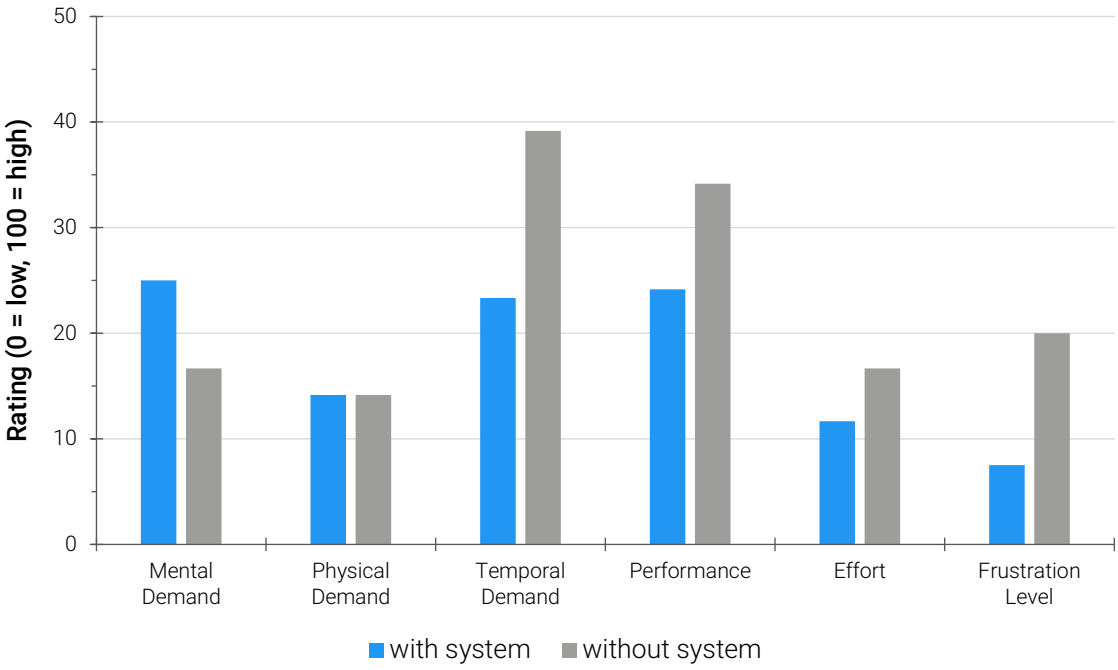


Figure 8.18: Results of the NASA-TLX questionnaire.

ditionally, the higher mental load could be attributed to the unfamiliarity with the novel technology and might be lower after getting accustomed to the system through further training and usage. One area where users benefited the most from our approach was the temporal demand. It was almost 16 points lower when using the system than the baseline, representing a load reduction of more than 40%. The second-largest difference could be observed in the frustration level, which was lowered by 12.5 points to less than half of its initial value, followed by performance with 10 and effort with 5 points, respectively. Despite the relatively small sample size, these results indicate that our approach can effectively reduce the cognitive load experienced while searching by assisting individuals with declining memory.

However, if the system does not perform as intended, even due to uncontrollable circumstances, people's attitudes towards it can change dramatically, as observed with users 4 and 7. To prevent such situations, alternative methods for critical system components should be considered as fallback solutions in future iterations. For instance, in case the Bluetooth connection gets disrupted again, an automatically created Wi-Fi hotspot on the smartphone could be used to maintain communication between devices until regular functionality is restored. Apart from implementation-related improvements, we also observed some hardware limitations that might be addressed with more technologically advanced devices in the future. This especially applies to the very short battery life of

the Google Glass. It depleted after only around one hour and forty minutes in one of our data collection sessions (see Section 8.2.3), resulting in an incomplete and unusable recording. While the primary goal was only to evaluate our approach with a proof of concept, this limitation prevented interested individuals from actually using the system for extended periods of time.

Another related weakness concerns the head-mounted display's limited processing capabilities. To prevent the device from overheating, we had to reduce the sample rates of the accelerometer and gyroscope sensors to 40 Hz and the frame rate of the camera even down to 1 Hz. Consequently, object recognition could only be performed once per second, which might have resulted in cases where an item was not detected because it was not in the camera's field of view for more than one second. Furthermore, we did not use our custom eye-tracking solution in the final system evaluation since a fusion of accelerometer and gyroscope features was sufficient to detect visual search instances. However, despite these restrictions, we were still able to perform all necessary processing steps directly on the mobile devices, which was especially important due to the highly personal nature of the collected sensor data. While cloud-based solutions might have improved certain results, we prioritized protecting people's privacy instead.

8.6 Summary

In this chapter, we demonstrated how the SSJ framework can be used to design and develop a memory augmentation system that assists older adults in situations where they cannot rely on their natural memory. Since forgetting and searching for the location of important items, such as keys or wallets, can be very frustrating and time-consuming, our goal was to automatically detect and appropriately support this process. To achieve that, we first conducted a design probe with ten elderly participants and evaluated different augmentation concepts. Based on their feedback, we implemented an approach that shows users an image of the desired object at its last seen position. This information can trigger episodic memories about previous interactions with the item and helps individuals find the right location through contextual clues within the image. Afterwards, we identified suitable signals to automatically detect the visual search behavior and provide assistance at appropriate moments. For the collection of meaningful training data, we constructed a custom eye- and head-tracking device based on the Google Glass head-mounted display and recorded 30 older adults during various everyday activities, including the visual search for misplaced objects.

We then trained and compared several binary classification models using different combinations of modalities, features, and fusion methods. While utilizing the accelerometer, gyroscope, and eye-tracking data yielded the best results, head movements alone still achieved reasonably high accuracies and only required a relatively simple setup, which is why we relied on them in the final system. To evaluate the effectiveness of our approach, we conducted a study with eight older adults affected by memory issues. Using a within-subjects counterbalanced design, we compared their performance and subjectively perceived workload in a visual search task with and without the system. Although two users experienced Bluetooth-related connectivity issues, all participants were able to successfully complete the task in both conditions. Considering the users where the system worked as intended, we observed a task load reduction in most of the NASA-TLX dimensions, including temporal demand, frustration level, performance, and effort. Despite the small sample size, these findings suggest that our approach can be beneficial for older adults with declining memory by supporting the visual search for misplaced objects and reducing the mental effort required during this process.

Chapter 9

Assisting Emotional Disorder

Depression and related cognitive disorders have a considerable impact on society and the healthcare system as a whole. According to the World Health Organization [2017], more than 4.4% of the global population suffered from depression in 2015, which represents an increase of 18.4% compared to the previous decade. Due to the worldwide impact of the recent COVID-19 pandemic, the prevalence of stress, anxiety, and depression has further increased significantly to around 30% among certain populations [Hawes et al., 2022; Salari et al., 2020]. While the clinical treatment of depression can be an effective, temporary solution, about 21% of patients in Germany are readmitted within a year after discharge [Wiegand et al., 2020]. These relapses can be mainly attributed to a lack of outpatient treatment opportunities in rural and economically underdeveloped regions, resulting in untreated conditions with severe consequences for affected individuals and their families [World Health Organization, 2017]. While some resources for follow-up care exist, such as educational web portals, email counseling, and text message check-ins, most of these options lack a social component, which can be important for successful treatment outcomes [Davis and Hadiks, 1994].

As shown by Lucas et al. [2014], autonomous social agents can positively affect the willingness of patients to disclose sensitive information, which they otherwise often hold back out of fear of being judged negatively by therapists and practitioners [Farber, 2003]. Since these details form the basis for potential treatment decisions, collecting truthful information is essential to ensure appropriate assistance. Although several approaches have utilized social agents to provide interventions for depression and related conditions, most of them relied on text-based interactions with chatbots [Otero-González et al., 2024], which are less natural than conversations with a visual character

Figure 9.1: Individual using the proposed cognitive augmentation system.

and might lead to different experiences. Among the few proposed solutions that actually employed a virtual avatar (e.g., Bresó et al. [2016], Burton et al. [2016], Egede et al. [2021], Philip et al. [2017], and Ring et al. [2016]), the majority did not support natural language input, required powerful hardware for computations, and were not usable on mobile devices. These limitations prevent potential applications from being utilized as an everyday companion that can be consulted at any place and time.

Consequently, we introduce a cognitive augmentation approach, which combines a virtual agent with the processing capabilities of the SSJ framework to create a ubiquitous assistant that can be used to counteract the gap in treatment options (see Figure 9.1). The system described in this chapter was developed as part of the EmmA[1] and UBIDENZ[2] projects, which were both funded by the German Federal Ministry of Research and Education (BMBF) and aimed to provide outpatient treatment for individuals with cognitive disorders. In this regard, a core principle of both projects was to complement gaps in existing care without replacing established elements, such as therapy sessions or clinical visits. The general idea was to create a virtual companion that adapts its behavior to the user's current condition and can be consulted in any situation. To achieve that, we first identified suitable use cases through expert interviews with therapists and patients. Based on insights gained from these discussions, we designed specific interaction sce-

[1] `https://interaktive-technologien.de/projekte/emma`

[2] `https://ubidenz.de`

narios with the social agent to support the outpatient treatment of affected individuals. For the analysis of people's conditions, we then trained a classification model and integrated it with the virtual avatar into a mobile application. Finally, we evaluated the system with 10 participants showing signs of depression and compared it to traditional methods, including paper-based diaries, questionnaires, and mindfulness exercises.

Parts of this chapter are based on the following publications:

Reference Gebhard, P., Schneeberger, T., Dietz, M., André, E., and Bajwa, N. u. H. (2019). Designing a Mobile Social and Vocational Reintegration Assistant for Burn-out Outpatient Treatment. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, pages 13–15. ACM.

Reference Schiller, D., Huber, T., Dietz, M., and André, E. (2020). Relevance-Based Data Masking: A Model-Agnostic Transfer Learning Approach for Facial Expression Recognition. *Frontiers in Computer Science*, 2(6).

9.1 Augmentation Design

Since the primary objective was to support the outpatient treatment of individuals after suffering from depression, burnout, and related cognitive disorders, certain characteristics of the dimensions identified in Section 4.4 were chosen accordingly to achieve this goal (see Figure 9.2). For instance, we shifted the augmentation initiative towards the users since the automatic recognition of suitable conditions would have required a more intrusive sensor setup, which could discourage individuals from using the system. While this meant that assistance could be requested at any time, the augmentation was designed to provide a situational response (i.e., short dialog or conversation) with temporary presence and limited duration. Additionally, we decided that the system and its sensors should be directed primarily at the user to analyze people's conditions during interactions with the social agent. Based on the resulting insights, the virtual avatar should dynamically adapt its behavior and provide empathic responses.

Similar to previous research probes, we followed a participatory design approach to develop the cognitive augmentation system. The foundation for that consisted of several expert interviews with patients and therapists from the fields of cognitive behavior therapy and cognitive psychoanalysis. In contrast to our initial concept, clinical experts opted to introduce the application already during a patient's stay at the clinic. They argued that, on the one hand, this period could be used for patients and therapists to get acquainted with the system. On the other hand, it could improve people's commitment

during outpatient treatment since they are already familiar with the application. This approach could be especially helpful in the early stages after hospitalization, when patients usually still have one weekly therapy session. Between these sessions, individuals could use the system to monitor their behavioral change and practice new skills, which could then be discussed with the therapist.

Figure 9.2: Selected design dimensions for augmenting cognitive disorder.

In later stages, when patients are on their own, they could still rely on the application to provide guidance and assess their current condition. To achieve that, three interaction scenarios were proposed: (1) initial acquaintance, (2) daily conversation, and (3) weekly overview. The initial introduction aims to help individuals get to know and build trust towards the virtual agent, which is crucial for a successful working relationship.

Additionally, general information about the user is gathered through questionnaires and conversations. This includes data about their personality, health, and current goals. The final stage of the introductory phase consists of a walk-through where all features of the application are explained. During daily interactions, the virtual assistant tracks and reminds the people of their goals and assesses personal variables, such as drive, strain, sleep, and well-being. Furthermore, it considers peoples expressed emotions, thoughts, and non-verbal signals to adapt its behavior accordingly.

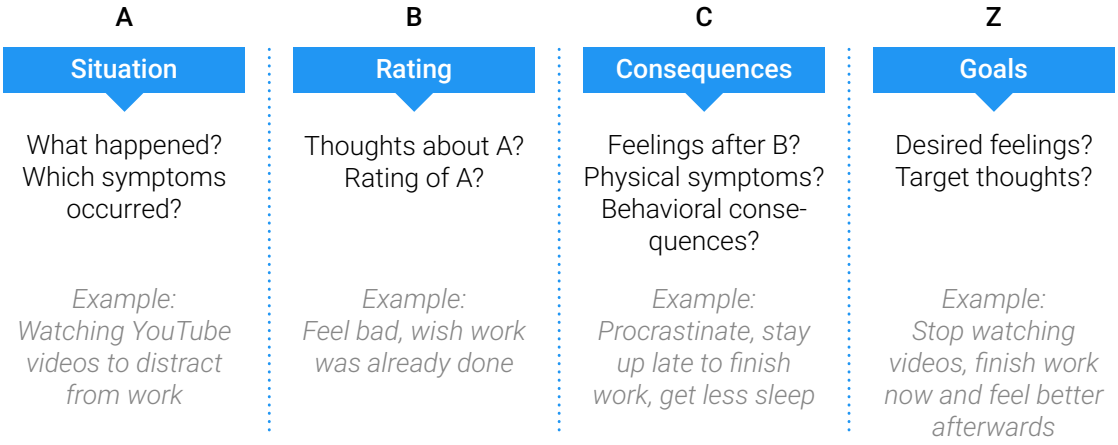


Figure 9.3: Structure of the ABCZ-model.

Regarding the content of these daily interactions, we used the ABCZ-model by Stavemann [2015] as a foundation for the dialogs of the virtual agent. The model identifies behaviors, thoughts, emotions, and their consequences in the current situation and compares them to a person's desired state. In addition to highlighting the differences between both conditions, it encourages users to think about potential solutions and consider behavioral changes to achieve the desired situation. As shown in Figure 9.3, the model generally consists of four stages. The first stage establishes the initial situation through questions regarding what happened and which symptoms occurred. In the second stage, the thoughts and ratings of the current condition are gathered. The third stage then asks about resulting feelings, accompanying physical symptoms, and behavioral consequences. Finally, the fourth stage determines people's desired thoughts and feelings and encourages them to view the initial situation from an outside perspective. This is achieved through questions such as *"Do you believe everyone would assess the situation similarly or are there other possibilities?"*, *"What would others do in this situation?"*, or *"How could you adapt your behavior to feel better?"*. After these questions, the virtual agent offers one out of four optional relaxation exercises, including breathing, mindfulness, and general meditation instructions.

The data acquired from these daily interactions servers as the foundation for the weekly overviews. They provide a history of the person's past emotional states and enable them to quickly identify potential trends based on their mood over the last few days. For instance, a steadily declining emotional state could trigger the decision to seek professional help, which otherwise might not have been the case until it's already too late. Overall, the application aims to provide a ubiquitous companion that can be used for self-assessment and personal reflection during challenging situations. These features are intended to complement existing outpatient treatment opportunities and improve people's mental health in the long term.

9.2 Condition Analysis

In order to detect people's conditions during interactions with the virtual agent, we first examined publicly available depression-related datasets. While we found a few selected options at the time, such as the Multimodal Open Dataset for Mental-disorder Analysis (MODMA) by Cai et al. [2022], the Pittsburgh dataset by Dibeklioglu et al. [2018], and the Distress Analysis Interview Corpus (DAIC) by Gratch et al. [2014], none of them provided raw audio and video data of participants. Instead, one of these modalities was either not included or was already encoded in precomputed feature sets, which could not be reproduced in real-time on mobile devices. Since recording our own data was also not an option due to the COVID-19 pandemic, we had to rely on general emotion-related corpora. As a result, we ended up using the following datasets to train various models for the facial expression-based recognition of people's current emotional state: (1) AffectNet by Mollahosseini et al. [2019], which contains 420,299 images and is one of the largest emotion-related databases, (2) FERPlus by Barsoum et al. [2016], which includes 35,887 images that were each labeled by 10 different annotators, and (3) CMU-MOSEI by Zadeh et al. [2018], which is composed of 23,453 video segments. While all of these datasets include annotations for at least six discrete emotions, AffectNet also contains valence and arousal values.

Due to the large number of samples in these databases and the widespread adoption of neural networks as a de facto standard for classifiers in recent years, we also decided to use them for our condition analysis models. In this regard, the current trend to accomplish higher detection rates is mainly achieved through increasingly complex architectures with billions of parameters. However, alongside the growing complexity, more and more computational resources are required to train and execute these networks within a reasonable amount of time. While the performance of mobile devices has also increased

significantly in recent years, it is still not comparable with that of current servers or even desktop computers. For this reason, using neural networks on mobile devices, such as smartphones or smartwatches, usually involves a trade-off between accurate results and required resources (i.e., processing power, inference duration, and battery life). Consequently, we first evaluated the following established neural network architectures from the field of computer vision regarding their recognition performance and execution time: InceptionV3 [Szegedy et al., 2016], Xception [Chollet, 2017], VGG-Face [Parkhi et al., 2015], and MobileNetV2 [Sandler et al., 2018].

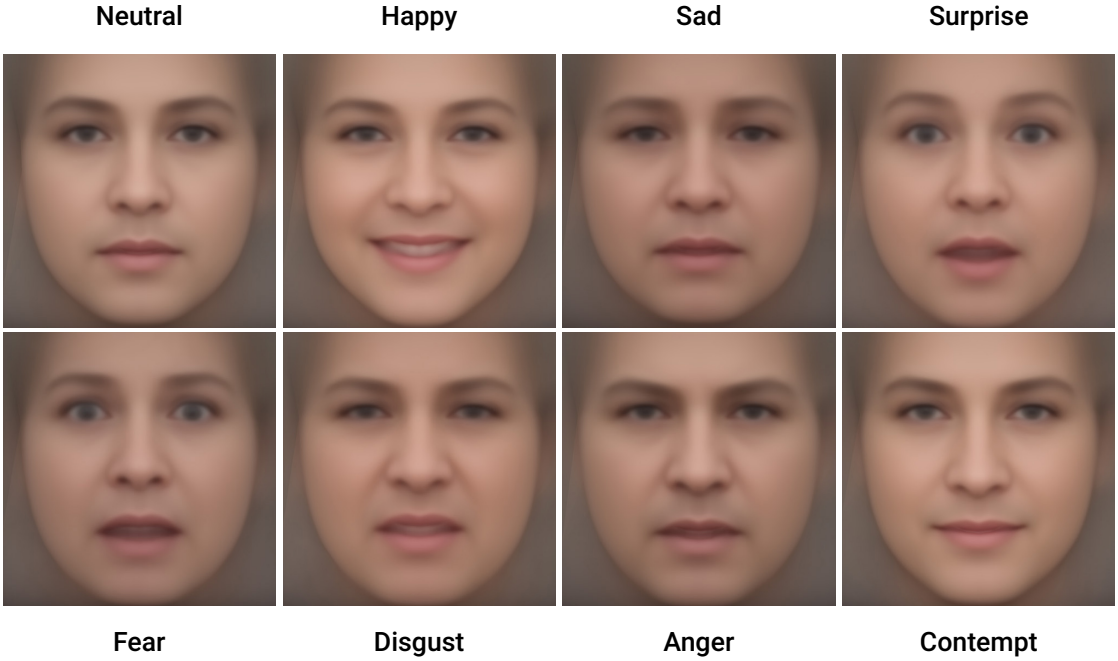


Figure 9.4: Average facial expressions per emotion class.

The goal was to identify suitable model architectures that could run on mobile devices and achieve reasonably high recognition results. Using transfer learning, we replaced the output nodes of the original domains (e.g., object detection) with corresponding classes for our targeted states (emotions). Additionally, we applied the following set of empirically evaluated data augmentation steps across all models to increase their robustness and prevent them from learning position-dependent features: Each image was randomly rotated by up to 25°, shifted by up to 10% of its total dimensions, and zoomed by up to 85% of its original size along both axes. Furthermore, each color channel was shifted within a range of 20% and the overall brightness of the images was adjusted between 50% and 150% of the original values. To counteract potentially imbalanced sample distributions in the datasets, we also applied a weighted loss function as suggested by Mollahosseini et al. [2019], which weights each class according to its

relative proportion in the training set. Overall, the MobileNetV2 architecture by Sandler et al. [2018] yielded the best balance between accurate results and required resources, which is why we used it for all further experiments.

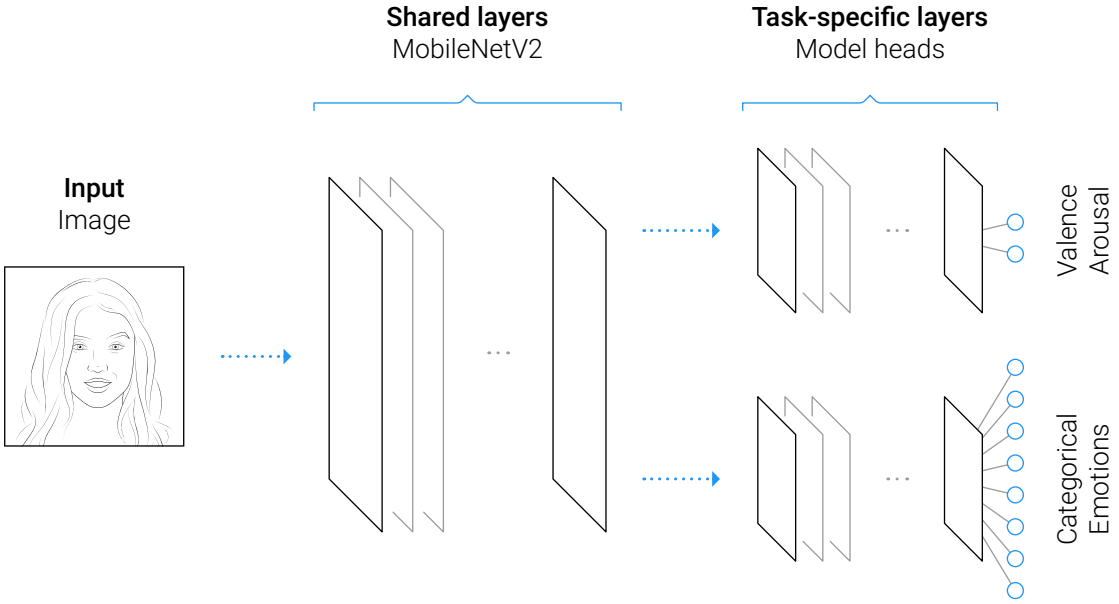


Figure 9.5: Multi-task MobileNetV2 architecture.

We then applied various multi-task learning approaches to stabilize and improve the recognition results. The idea behind this method is to adapt the network architecture in such a way that the model simultaneously calculates other tasks in addition to the main objective based on the same input data (e.g., object recognition and scene classification). Ideally, there are common properties between all tasks, which can mutually enhance and potentially improve the results of each individual goal. As shown in Figure 9.4, we first calculated images of the average facial expression per emotion class for each dataset and used their generation as a secondary task to complement the primary emotion recognition capabilities of the model. The hypothesis behind this procedure was that emotions might be recognized more easily by mapping the input image to the average facial expression of the respective emotion class. Unfortunately, this approach did not produce the desired improvements. As an alternative, we applied a method from the field of explainable artificial intelligence called "deep Taylor decomposition" by Montavon et al. [2017] as a secondary task, which highlights parts of the input image that were relevant for the model's decision. Our rationale behind this procedure was based on the assumption that mapping the images to the corresponding average facial expression might have been too complex, and instead, focusing only on relevant areas could lead to more promising results. However, this approach did also not deliver the ex-

pected improvements, which might be related to the different nature of these secondary tasks compared to the processes involved in recognizing the emotional state.

	Baseline (AlexNet)		MobileNetV2 Model	
Metric	Valence	Arousal	Valence	Arousal
RMSE	0.37	0.41	0.40	0.37
CORR	0.66	0.54	0.60	0.52
SAGR	0.74	0.65	0.73	0.75
CCC	0.60	0.34	0.57	0.44

Table 9.1: Performance comparison between AffectNet baseline and our model.

Consequently, we combined the continuous prediction of valence and arousal values with the assessment of discrete emotion classes due to the similarity of both tasks. Since the required dimensional labels (valence and arousal) were only available in the AffectNet dataset, the other corpora were not included in the training process of this model. An overview of the adapted MobileNetV2 architecture is shown in Figure 9.5. Despite previous setbacks, this approach finally yielded the desired improvements in recognition rates, which are comparable to the results of much larger models. As shown in Table 9.1, the lightweight MobileNetV2 architecture with only 3.4 million parameters was able to match the performance of the AlexNet model [Krizhevsky et al., 2017] with 60 million parameters that was used for the AffectNet baseline. While these results were sufficient for our purposes at the time, newer architectures, such as GhostNet [Han et al., 2020] or EfficientNetV2 [Tan and Le, 2021], have emerged since then and could be used to achieve even better recognition rates in future iterations of the system.

9.3 System Overview

Based on the participatory design results and the recognition model for people's emotional state described in the previous sections, we developed a mobile virtual assistant to support the outpatient treatment of individuals recovering from depression and related cognitive disorders. Similar to a diary, the application was designed so that users can interact with it on a daily basis. This enables people to regularly share their thoughts and feelings with the virtual agent and encourages a continued reflection of their behavior. Additionally, it provides an interactive tool for self-assessment and allows the system to progressively monitor people's condition across several days and weeks. In

this regard, the collected data is mainly used to adapt the behavior of the social agent in real-time according to the user's current circumstances but can also serve as a foundation in follow-up sessions with a therapist. Due to the highly sensitive nature of the analyzed social signals, all necessary processing steps were performed directly on the mobile devices. This requirement was especially important to give people complete control over their personal data and increase their trust towards the system.

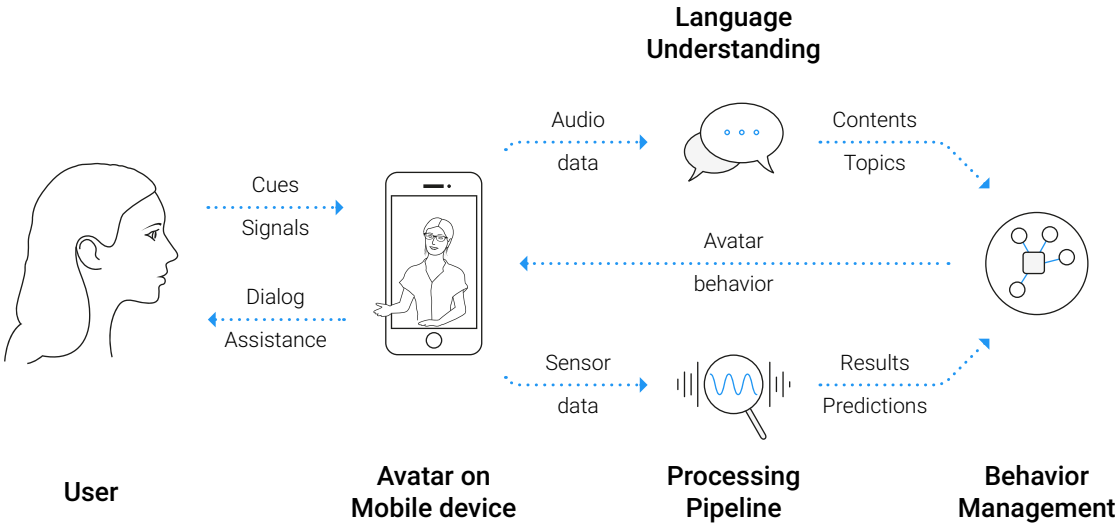


Figure 9.6: Architecture of the cognitive augmentation system.

Figure 9.6 shows an overview of the system architecture. In general, users primarily interact with the WebGL-based virtual avatar, which was developed by the project partner Charamel GmbH[3]. Based on feedback from therapists and patients, the consortium decided to use a fixed appearance for the agent in the initial version of the system. However, it would be possible to provide individual customization options for personalized virtual characters in future iterations to further increase people's bond with the avatar. During interactions, the audio signal is passed to a component from the project partner semvox GmbH[4], which performs on-device natural language understanding and topic identification. The resulting dialog content is then sent to a mobile version of the Visual SceneMaker (VSM) by Gebhard et al. [2012]. It is used to model and control the behavior and responses of the virtual character in real-time. While significant progress has recently been made with generative language models (e.g., GPT-3 [Brown et al., 2020], PaLM [Chowdhery et al., 2023], or LLaMA [Touvron et al., 2023]), we explicitly chose a manually created interaction model based on expert knowledge to prevent the sys-

[3] `https://charamel.com`

[4] `https://semvox.de`

tem from giving wrong advice and inappropriate responses that could be particularly harmful to individuals with cognitive disorders.

In parallel, people's behavior and condition are analyzed using the SSJ framework. Although it would have been possible to recreate the previously mentioned functions with custom pipeline components, we instead decided to showcase alternative methods of integrating the framework with existing external modules. To this end, we first added a `CameraSensor` and a `CameraChannel` to capture the video stream of the front-facing camera containing the user's face and upper body. Since the recognition model (Section 9.2) only required an input size of 224×224 pixels, we set the camera resolution to 640×480 pixels (lines 4-5) and limited the frame rate to 5 Hz (line 9) to reduce power consumption and prevent the devices from overheating.

```
1  // Create camera sensor with resolution of 640x480 pixels
2  CameraSensor cameraSensor = new CameraSensor();
3  cameraSensor.options.cameraType.set(Cons.CameraType.FRONT_CAMERA);
4  cameraSensor.options.width.set(640);
5  cameraSensor.options.height.set(480);
6
7  // Create camera channel with sample rate of 5 Hz
8  CameraChannel cameraChannel = new CameraChannel();
9  cameraChannel.options.sampleRate.set(5);
10
11 // Add components to pipeline
12 pipeline.addSensor(cameraSensor, cameraChannel);
```

Following that, we added several components to perform specific preprocessing steps on each input frame. These included converting the video stream into an appropriate encoding format, extracting the user's facial region, and normalizing the pixel color values. In this regard, resizing the camera images to fit the input dimensions of the emotion recognition model was not necessary because the `FaceCrop` component already provides output images with a resolution of 224×224 pixels by default.

```
13 // Create transformer to convert encoding format from NV21 to RGB
14 NV21ToRGBDecoder rgbDecoder = new NV21ToRGBDecoder();
15 pipeline.addTransformer(rgbDecoder, cameraChannel);
16
17 // Create transformer to extract facial region
18 FaceCrop faceCrop = new FaceCrop();
19 pipeline.addTransformer(faceCrop, rgbDecoder);
20
21 // Create transformer to normalize image pixel values between -1 and 1
22 ImageNormalizer imageNormalizer = new ImageNormalizer();
23 pipeline.addTransformer(imageNormalizer, faceCrop);
```

The processed images were then passed to a `ClassifierT` transformer component to perform the emotion recognition task. It loads the model stored in the Tensorflow Lite[5] format from the file system and provides the normalized pixel values as input. While we initially intended to use categorical predictions for people's emotional state, we quickly noticed that incorrect detection results of similar expressions could severely alter and negatively impact interactions (e.g., when the model detects "fear" instead of "surprise"). To counteract this problem, we used the valence and arousal values from the secondary model head instead. A significant advantage of these dimensional prediction results is that even though the ground truth values might be slightly higher or lower, their general direction is usually still correct (e.g., positive valence and high arousal).

```
24 // Create TensorFlow Lite model and select model file
25 TFLite vaModel = new TFLite();
26 vaModel.options.file.set(new FilePath("/model/valence_arousal.trainer"));
27
28 // Create classifier and select model
29 ClassifierT emotionClassifier = new ClassifierT();
30 emotionClassifier.setModel(vaModel);
31
32 // Add components to pipeline
33 pipeline.addModel(vaModel);
34 pipeline.addTransformer(emotionClassifier, imageNormalizer);
```

Subsequently, the model outputs were converted to XML-based events and sent to the Visual SceneMaker (VSM) module with the `SocketEventWriter` component, which internally uses UDP sockets for communication. On the receiving side, we implemented a small VSM plugin to properly handle these events and assign the transmitted values to corresponding variables within the Visual SceneMaker. This enabled their usage in conditional queries and branching paths of the interaction model.

```
35 // Create transformer to convert float values to XML event
36 FloatsEventSender fesEmotion = new FloatsEventSender();
37 fesEmotion.options.sender.set("face");
38 fesEmotion.options.event.set("emotion");
39 pipeline.addConsumer(fesEmotion, emotionClassifier);
40
41 // Create socket writer to send XML event to VSM
42 SocketEventWriter sewEmotion = new SocketEventWriter();
43 sewEmotion.options.ip.set("127.0.0.1");
44 sewEmotion.options.port.set(5000);
45 sewEmotion.options.sendAsMap.set(true);
46 sewEmotion.options.mapKeys.set("valence,arousal");
```

[5] `https://tensorflow.org/lite`

```
47 pipeline.registerEventListener(sewEmotion, fesEmotion);
```

In addition to predicting a person's emotional state, we extracted their facial landmarks with the BlazeFace model by Bazarevsky et al. [2019]. The resulting coordinates were then used to calculate various features, such as the relative face position or the mouth open/close score [Nilsson et al., 2010]. Similar to the valence and arousal values (lines 35-47), these facial features were sent to the Visual SceneMaker, enabling further options for adapting the virtual character's behavior to a person's current circumstances (e.g., the avatar always looks at the user based on the recognized face position).

```
48 // Create transformer to calculate facial landmarks
49 FaceLandmarks landmarkTransformer = new FaceLandmarks();
50 pipeline.addTransformer(landmarkTransformer, rgbDecoder);
51
52 // Create transformer to calculate landmark features
53 LandmarkFeatures landmarkFeatures = new LandmarkFeatures();
54 pipeline.addTransformer(landmarkFeatures, landmarkTransformer);
```

Apart from examining visual characteristics, we also analyzed people's vocal signals. To this end, a `Microphone` and an `AudioChannel` component were added to the processing pipeline. The audio stream's sample rate was set to 16 kHz, which is still high enough to retain most vocal characteristics but requires much less processing power compared to more common sample rates for music and high-quality audio recordings (i.e., 44.1 kHz or 48 kHz).

```
55 // Create microphone sensor
56 Microphone microphone = new Microphone();
57
58 // Create audio channel with 16 kHz sample rate
59 AudioChannel audio = new AudioChannel();
60 audio.options.sampleRate.set(16000);
61
62 // Add components to pipeline
63 pipeline.addSensor(microphone, audio);
```

We then extracted various features from the captured audio signals, including pitch, energy, intensity, Mel-frequency cepstral coefficients (MFCCs), and eGeMAPS [Eyben et al., 2016]. For that, several components based on the established audio processing libraries openSMILE [Eyben et al., 2010], PRAAT [Boersma, 2001], and TarsosDSP [Six et al., 2014] were used. Similar to the visual features (lines 35-47), they were also sent to the Visual SceneMaker for consideration in the interaction model.

```
64 // Create transformers to calculate audio features
65 Pitch pitch = new Pitch();
66 Intensity intensity = new Intensity();
67 Energy energy = new Energy();
68 OpenSmileFeatures egemaps = new OpenSmileFeatures();
69 OpenSmileFeatures mfcc = new OpenSmileFeatures();
70
71 egemaps.options.configFile.set(new FilePath("/ssj/os_egemaps_23.conf"));
72 mfcc.options.configFile.set(new FilePath("/ssj/os_mfcc_39.conf"));
73
74 // Add components to pipeline
75 pipeline.addTransformer(pitch, audio);
76 pipeline.addTransformer(intensity, audio);
77 pipeline.addTransformer(energy, audio);
78 pipeline.addTransformer(egemaps, audio);
79 pipeline.addTransformer(mfcc, audio);
```

Finally, all recorded signals and calculated feature streams were stored directly on the mobile devices (see Appendix C), allowing users to review the data and potentially enabling the training of personalized models in the future.

9.4 Evaluation

The study described in this section was conducted as part of the EmmA project. Due to its personal nature and potential impact on society, the research and development process was accompanied by an external advisory board that monitored the ethical, legal, and social implications (ELSI) and consisted of experts from each field. They provided recommendations throughout each step of the project and approved the design of the present study. Additionally, approval was obtained from the Ethical Review Board of the Faculty of Mathematics and Computer Science at Saarland University[6].

In order to evaluate the final system, we initially planned to conduct a study with 60 participants split into three conditions: (1) experimental group with the system, (2) active control group with established paper-based questionnaires and exercises, and (3) passive control group without assistance. However, after screening more than 600 people with the Beck Depression Inventory-II (BDI-II) [Beck et al., 1996] and defining a score of 11 or higher (mild depression) as the minimum inclusion criteria, only 10 suitable individuals (40% female) could be identified as potential candidates for our study.

[6] https://erb.cs.uni-saarland.de

9.4.1 Procedure

Due to the low number of participants, the project consortium decided to omit the passive control group and split the users between the remaining groups to at least get an indication of how the system performs compared to established methods. At the beginning of each experimental run, participants received a brief overview of the study details and procedures. In addition to surveying general demographic data, users were also asked to fill out the following standardized questionnaires: Beck Depression Inventory-II (BDI-II) [Beck et al., 1996], Perceived Stress Scale (PSS) [Cohen et al., 1983], and a student version of the Oldenburg Burnout Inventory (OLBI-S) [Reis et al., 2015]. Depending on their associated group, participants either received a tablet (Samsung Galaxy S8+) with the augmentation system installed or a paper-based ABCZ questionnaire and were instructed on how to use them. Once they became familiar with the respective item, they were tasked to utilize it at least once per day over the course of a week. After this period, participants were asked to complete the same questionnaires as in the initial examination for comparison (BDI-II, PSS, and OLBI-S). Additionally, the experimental group received several questions regarding their subjective experience with the application as well as the system's helpfulness and usability.

9.4.2 Results

Overall, the study was conducted without any major problems despite taking place during the COVID-19 pandemic. However, one participant from the control group did not fully complete the experimental trial and was therefore excluded from all further analyses. The average results of the standardized questionnaires from the remaining nine participants before and after the study are summarized in Table 9.2. Among the current set of subjects, no significant differences could be determined between the experimental and the control group. Instead, the scores between both conditions were relatively similar across all questionnaires. These similarities also apply when comparing the general trend before and after the experiment in two out of three cases.

	Before Study			**After Study**		
Condition	BDI-II	PSS	OLBI-S	BDI-II	PSS	OLBI-S
Virtual Agent	12.60	20.00	38.20	16.00	21.40	37.00
Paper Questionnaires	11.75	21.25	39.25	13.25	19.00	38.25

Table 9.2: Results of the BDI-II, PSS, and OLBI-S questionnaires.

While the Beck Depression Inventory scores increased from 12.6 and 11.75 to 16.0 and 13.25 respectively, the Oldenburg Burnout Inventory results slightly decreased by 1.2 and 1.0 points after the study. Other than that, only the Perceived Stress Scale showed a complementary trend between the experimental and the control group, where the former increased from a score of 20.0 to 21.4 and the latter decreased from 21.25 to 19.0. Regarding the subjective experience with the virtual agent, most participants found the application to be helpful and user-friendly, although some people mentioned they would have liked even more control over the interactions. Suggestions included a "pause button" to think about responses and continue the conversations at a later point in time, as well as a "recording feature" to replace or complement previous answers. Apart from that, almost all participants praised the relaxation exercises and expressed the desire to include additional ones for an increased variety and enhanced overall experience.

9.5 Discussion

Although participants' subjective experiences with the application were relatively positive, the objective results between the agent-based application and traditional paper-based methods are less conclusive and vary depending on the selected questionnaire. For instance, the Oldenburg Burnout Inventory showed reduced scores for both groups after the experiment, which could indicate that both methods contributed to minor improvements. Since the decrease was more prominent for users of the augmentation system, this result might also suggest that the application provides a more engaging and interactive way for participants to reflect on their situation and manage potential sources of stress. However, the Beck Depression Inventory results indicate the exact opposite and show increased symptoms of depression in both groups, with a slightly higher rise in the virtual agent condition. Consequently, it could be argued that both approaches might not be suitable to prevent a depression-related decline, even though the traditional methods have been established for this purpose. Finally, the PSS results lead to yet another conclusion and indicate that individuals felt more stressed after using the system for one week. In contrast, participants with the paper-based questionnaires reported slightly lower stress levels, which could imply that the application is not as effective as traditional methods and might even negatively impact people's conditions.

Before considering any of the previous assumptions, it is important to acknowledge the circumstances and limitations that led to these results. Firstly, the sample size was significantly smaller than initially planned due to strict inclusion criteria and reduced participant availability during the COVID-19 pandemic, which limits the generalizability

of the findings. Secondly, the passive control group had to be omitted to compensate for the low number of participants, which means that potential changes can not be definitively attributed to the approaches themselves since no baseline containing the natural changes over time is available for comparison. The absence of a passive control group also makes it difficult to assess the influence of external factors, such as the COVID-19 pandemic, which could have exacerbated participants' mental health challenges during the evaluation. For these reasons, we will conduct a follow-up study in cooperation with the Karl-Jaspers-Clinic[7] for psychiatry and psychotherapy in Oldenburg as part of the UBIDENZ project to address the shortcomings of the present experiment. Their direct contact to current and prior patients with depression facilitates recruiting a larger sample size and enables the possibility to include a passive control group. Additionally, a second study with customers of the UBIDENZ project partner Better@Home Service GmbH[8] is planned to evaluate the system over an extended period of time. For that, the application will be integrated into their tablet-based solution and deployed in people's homes to collect interaction data over several weeks. Although the present study did not produce conclusive evidence for the system's effectiveness, it still provided valuable insights, especially regarding participants' subjective experiences, which can be used to further improve the application in subsequent iterations.

9.6 Summary

This chapter illustrated how the SSJ framework can be integrated with external components to rapidly design and develop a cognitive augmentation system that supports the outpatient treatment of individuals recovering from depression and related cognitive disorders. Similar to previous research probes, we followed a participatory design approach and conducted several expert interviews with patients and therapists from the fields of cognitive behavior therapy and cognitive psychoanalysis. Based on their feedback, we implemented a mobile virtual assistant that enables users to regularly share their thoughts and feelings with a trusted companion and encourages a continued reflection of their behavior. Additionally, it provides an interactive tool for self-assessment and allows the system to monitor people's condition across extended periods. For that, we trained a recognition model on publicly available datasets using various machine-learning methods and techniques. Afterwards, the model was integrated into a real-time processing pipeline that analyzes the captured sensor data during conversations with the

[7] `https://karl-jaspers-klinik.de`

[8] `https://behome.info`

virtual avatar and provides the results to the interaction management component, which adapts the agent's behavior according to people's current circumstances.

To evaluate the effectiveness of our approach, we initially intended to conduct a study with 60 participants, comparing the system with traditional paper-based methods and a passive baseline. Unfortunately, only 10 suitable individuals showing signs of mild depression were willing to participate in the experiment due to strict inclusion criteria and reduced participant availability during the COVID-19 pandemic. While the limited number of records reduced the conclusiveness of the employed objective measures, people's subjective responses were largely positive and yielded valuable insights regarding potential opportunities for further improvements. Consequently, we plan to incorporate their suggestions and conduct multiple follow-up studies to evaluate the system's effectiveness with a larger sample size and over an extended period of time.

IV Conclusion

10 Contributions 229

10.1 Conceptual Analysis
10.2 Technical Implementation
10.3 Empirical Validation

11 Future Work 235

11.1 Long-Term Studies
11.2 Multi-User Applications
11.3 Additional Scenarios
11.4 Ethical Considerations

Bibliography 239

Appendix 289

A Component Options
B Smartphone Pipeline for Visual Search Support
C Smartphone Pipeline for Video and Audio Analysis

Chapter 10

Contributions

The primary goal of this thesis was to support individuals with impairments and disorders of cognitive processes by developing a flexible and easily usable framework that utilizes the capabilities of mobile devices and enables the rapid prototyping and implementation of assistive augmentation approaches. In order to achieve that, we produced the following conceptual, technical, and empirical contributions, which are described in more detail below.

10.1 Conceptual Analysis

The conceptual contributions of this work are grounded in a systematic literature analysis conducted to identify similarities and shared concepts among previous assistive augmentation approaches (see Section 4.2). For that, we first gathered a set of 15,358 publications from relevant HCI venues and followed the PRISMA 2020 guidelines to determine suitable candidates for further comparison. Based on the resulting evaluation, we provided an extensive overview of the most commonly applied sensory, memory, and cognitive augmentation strategies encountered in our review and included concrete examples for each area to demonstrate the specific methods and procedures that should be considered when creating new assistive systems and technologies in the respective field.

Apart from the identified strategies, we also derived five general design dimensions that reflect the primary characteristics of the analyzed approaches and can be used to classify existing systems (see Section 4.4). These dimensions specify the targeted cognitive process (see Section 4.4.1), how the augmentation gets initiated (see Section 4.4.2), when and how long the assistance should be present (see Section 4.4.3), which aspects of

reality are being analyzed (see Section 4.4.4), and whether a system can be adjusted after its initial deployment (see Section 4.4.5). The intention behind these dimensions is to guide designers and developers of future augmentation systems in their decision-making process by informing them about relevant characteristics and associated implications of available solutions. To further demonstrate the feasibility of these dimensions, we used them during the conceptual phase of three research probes and specified the properties of the respective systems for supporting visual impairment (see Section 7.1), memory decline (see Section 8.1), and cognitive disorders (see Section 9.1).

In addition to the strategies and design dimensions, we also identified a common technical structure among the analyzed approaches (see Section 4.2.2). While not every application required all components, most of them at least included a subset to achieve their goals. Consequently, the shared structure can be used to guide the design of future systems and serves as the foundation for our universal framework, which consists of reusable and easily exchangeable components that represent each part of the identified architecture (see Chapter 6). Furthermore, we provided an overview of the underlying concepts and theories from the field of cognitive psychology to improve the understanding of involved mental processes (see Chapter 2). Combined with a selection of suitable non-verbal signals and their associated insights about a person's cognitive state (see Chapter 3), this information can facilitate the creation of more effective solutions.

10.2 Technical Implementation

The primary technical contribution of this thesis is the SSJ framework for building and prototyping assistive augmentation systems that support cognitive processes using mobile signal processing techniques (see Chapter 6). SSJ enables the real-time detection of people's current mental states on mobile devices and facilitates providing ubiquitous assistance based on analyzed sensor data. During the development phase, we ensured that the framework includes all necessary capabilities to replicate, adapt, extend, and innovate the augmentation strategies described in Section 4.3. This was achieved by aligning the framework's architecture with the common technical structure of existing approaches. The resulting modular design enables developers to rearrange, repurpose, reuse, and replace all involved components, which increases flexibility, reduces iteration times, and encourages experimentation with alternative solutions. Besides integrating several sensors and output devices, we also developed a simple interface to extend the framework with additional components. Combined with the implemented synchronization mechanisms and support for all primitive data types, SSJ enables mobile appli-

cations to handle multimodal signals and provide tailored assistance through various communication channels. Moreover, the framework performs all processing steps directly on people's devices to protect their privacy and keep them in control of sensitive information. In addition to that, SSJ can be applied in combination with user-centered design processes (see research probes in Part III) to further increase individuals' trust and acceptance of potential augmentation solutions. Due to its flexible and modular architecture, frequent refinements and revisions based on people's feedback can be easily implemented and evaluated. Overall, the framework supports recording, processing, and classifying signals from various sensors on mobile devices in real-time and can deliver appropriate assistance across multiple modalities.

To demonstrate the practical feasibility of our proposed solution, we designed and developed three assistive augmentation approaches, each targeting a different group of cognitive processes. The resulting systems successfully made otherwise not perceivable information accessible to individuals (see Section 7.2), reminded them about forgotten details (see Section 8.3), and supported specific cognitive conditions (see Section 9.3). Additionally, the framework served as the technical foundation for various nationally funded research projects, including Glassistant, SenseEmotion, EmmA, and Ubidenz. While its primary purpose was the development of assistive augmentation approaches, the framework also functioned as a flexible and reliable solution to record multimodal datasets. For instance, during the Glassistant project, we conducted a field study with 16 older adults (aged 66-81), which involved capturing physiological signals and stress annotations from wearable devices with SSJ over four weeks [Dietz et al., 2019]. The resulting corpus contained more than 2,400 hours of data and yielded valuable insights regarding participants' annotation behavior and physiological reactions.

Apart from its internal usage, the first publicly available[1] version of the framework was released in February 2016. Since then, it has been continuously improved and expanded across more than 1,000 commits and 24 major versions. The framework itself is provided as an Android library that can be integrated into other applications by simply downloading the package and specifying its contents as a dependency. Once this is done, the public programming interface can be used to instantiate existing components, create new ones, or configure and execute pipelines. While these capabilities are sufficient for developers to work with SSJ, our goal was to make the framework accessible to an even bigger audience. Consequently, we developed a user-friendly Android application that enables people without any technical background or programming knowledge

[1] `https://github.com/hcmlab/ssj`

to visually create and execute pipelines with the same functionalities (see Section 6.4). Besides that, the direct integration of end-users into the design and development process further improves their understanding of and increases trust in mobile and assistive technologies. Since October 2016, the application has also been made available through the Google Play Store, from where over 1,000 researchers and interested users have downloaded it across more than 70 countries.

10.3 Empirical Validation

The proposed conceptual and technical solutions have been empirically evaluated in three user studies. To ensure the validity of our results, we conducted each experiment under the most realistic conditions possible and only recruited participants from the intended user groups of the respective approaches. In addition to evaluating the general feasibility of using our proposed solutions to augment different groups of cognitive processes, each research probe also investigated novel aspects within the corresponding fields. For instance, the first study not only examined the possibility of implementing a sonification system with the SSJ framework but also explored whether blind and visually impaired people can control the augmentation with the remaining movement capabilities of their eyes (see Section 7.3). Despite the relatively small sample size of seven visually impaired participants, our experiments still provided valuable results. While the system itself worked as intended, three users were not able to reliably control the augmentation with their eyes due to the nature of their impairments. In these cases, we adjusted the pipelines so they could use their head movements instead. Such modifications are in line with the goals of assistive augmentation, which strives to make technology accessible to as many people as possible, regardless of their conditions, and showcases the ability of the SSJ framework to achieve that. For people who can fully utilize our approach, gaze- and head-movement-based inputs appear to be very promising methods to control the sonification of visual information. Apart from objective measures, the study also yielded various positive subjective responses, including high ratings for the system's usefulness, usage probability, and sound pleasantness.

In the second research probe, we designed and developed a memory augmentation system that automatically detects when users search for misplaced objects and appropriately supports this process by showing them the forgotten location. To evaluate the effectiveness of our approach, we conducted a study involving eight older adults (aged 70-81) with memory impairments (see Section 8.4). During our experiments, we compared their objective performance and subjectively perceived workload in a visual search task

with and without the system. Although two users experienced temporary Bluetooth-related connectivity issues, the proposed solution generally worked as intended and enabled all participants to complete the given task successfully. For users unaffected by these signal interferences, we observed a task load reduction in four out of six NASA-TLX dimensions, including effort, frustration level, performance, and temporal demand. Despite the limited sample size, these findings indicate that our approach can benefit older adults with declining memory by supporting the process of visually searching for misplaced objects and reducing the mental effort required for this task. Moreover, participants rated the system as very positive and helpful in their subjective responses.

Finally, the last research probe involved developing a cognitive augmentation system to support the outpatient treatment of individuals recovering from depression and related cognitive disorders. The approach was evaluated in a one-week field study with 10 participants showing signs of mild depression (see Section 9.4). During this time, people either interacted with the system or with equivalent paper-based questionnaires. From a technical perspective, the proposed solution worked exactly as intended without any incidents or failures. Although the objective measures did not reveal significant differences between both conditions, people's subjective responses leaned more positively towards the system. Especially the interactive relaxation exercises were particularly well received and praised by all participants. Moreover, the collected feedback also yielded valuable insights regarding potential opportunities to further improve the application. Suggestions included the ability to pause interactions and replace or complement previous responses. Overall, the empirical evaluations throughout this thesis demonstrated the effectiveness of our proposed solutions and showcased the technical feasibility of our universal framework for building and prototyping assistive augmentation systems to support various groups of cognitive processes.

Chapter 11

Future Work

Overall, the assistive augmentation of cognitive processes is still a relatively young and emerging research field. Although this thesis has made significant advancements regarding mobile technologies that can be used to build solutions for this purpose, several areas remain for future research and development. In addition to the solid foundation laid by our conceptual, technical, and empirical contributions, further investigations and enhancements can be performed to fully realize the potential of assistive augmentation approaches for supporting cognitive processes. As a result, this chapter outlines potential directions and opportunities for future work to refine, expand, and innovate upon the current findings.

11.1 Long-Term Studies

The studies conducted throughout this thesis primarily focused on evaluating the effectiveness and immediate implications of our proposed solutions. While some results indicated the feasibility of using our approaches for prolonged durations, these hypotheses need to be confirmed in long-term experiments. For instance, the acoustic pleasantness of the sonification system introduced in Chapter 7 was rated very positive, which might lead to the conclusion that the sounds do not negatively affect individuals even after continued exposure. However, such assumptions can only be verified with long-term trials. Additionally, it should be investigated whether there are any unintended interactions or unexpected side effects that only occur after a certain period. This applies to both negative and positive consequences of long-term deployments. For example, it would be interesting to examine whether individuals show learning effects and use the systems differently or adjust their behavior over time. In this regard, it could also be evalu-

ated whether subsequent modifications of augmentation systems after specific periods could improve their effectiveness. Apart from the long-term implications for individuals, another important aspect are the potential consequences for their social sphere. Since impairments and disorders of cognitive processes can become a heavy burden for family members of affected individuals, future work should investigate whether these solutions can relieve the impact and improve their lives as well.

11.2 Multi-User Applications

For the research probes in this thesis, we only considered the data of individual users. The reasoning behind this decision was our intention to provide reliable solutions that solely depend on signals with guaranteed availability. Consequently, we focused on supporting people's personal conditions by analyzing their individual behavior, environment, or a combination of both. While the resulting augmentation approaches achieved their intended goals, some of them could be improved even further in specific situations by incorporating additional information from other users. For instance, the memory augmentation system introduced in Chapter 8 could be complemented with the data from all people living in the same household. Once a potential object of interest appears within their field of view, its position could be automatically updated in a shared directory. If a family member later searches for the respective item, the current location would always be shown, even if that particular person has not seen it there. Another possible extension concerns the cognitive augmentation approach described in Chapter 9. In addition to the virtual agent interactions, conversations with other people could also be considered for the analysis of a person's mental state, which would increase the number of available data points and might improve the resulting assessments. However, this example also highlights potential privacy concerns that can arise when involving data from other people. For instance, some conversational partners might not consent to the analysis of their interactions or might behave differently due to the knowledge of being recorded. These challenges need to be considered and addressed in future research.

11.3 Additional Scenarios

While this thesis demonstrated the effectiveness of our proposed assistive augmentation approaches based on concrete research probes for each primary group of cognitive processes (perception, memory, and higher-order cognition), various other application scenarios could be explored within the respective areas. One potential example of aug-

menting a different perceptual process could involve converting auditory information into visual representations to support people with hearing impairments. More precisely, a pipeline could be developed that maps specific sounds to visual icons and overlays them at the location they originated from within a person's field of view. The necessary sensor data could be acquired with a microphone, and the resulting visualizations could be shown on a wearable head-mounted display. Other examples concern the different processes involved in higher-order cognition. For instance, investigating cognitive augmentation approaches that automatically recognize specific problems and assist individuals in solving them could be very beneficial. Due to the recent advancements with generative language models (e.g., GPT-3 [Brown et al., 2020], PaLM [Chowdhery et al., 2023], or LLaMA [Touvron et al., 2023]), potential solutions could translate challenging situations into prompts and utilize the produced responses to support their users. However, employing such approaches requires additional safety mechanisms to prevent these systems from providing individuals with wrong or even harmful advice.

11.4 Ethical Considerations

As the ubiquitous nature of mobile technologies enables assistive augmentation approaches to become increasingly integrated into people's daily lives, it is essential to address the ethical considerations that arise from their development and permanent deployment. Although we accounted for common ethical challenges (see Section 5.1.3) when designing our proposed conceptual and technical solutions, further aspects still need to be considered. Starting with the implications for individuals, it should be examined how much support can be provided by assistive systems without undermining people's agency and autonomy. While augmentation solutions can be very beneficial for affected users, there is also a risk of becoming overly reliant on them, which could lead to a further decline in their cognitive abilities. Instead of creating such involuntary dependencies, these technologies should empower users to make their own decisions and always stay in control of the utilized assistance. Consequently, future work should explore ways to balance the support provided by augmentation approaches with opportunities for users to engage in cognitive exercises and activities that promote independence and sustained mental health whenever possible. In this regard, investigating the impact of potential system failures and downtimes on individuals could yield valuable insights to better understand the role of augmentation solutions within their daily lives and further improve their personal assistance. Apart from direct consequences for users, indirect effects on others should also be considered. To this end, future work should ex-

plore whether augmentation approaches for cognitive processes affect people's opinion of or behavior towards users of such systems. This research objective also applies to the views of other affected individuals without current access to such solutions and the possible implications for them. Since creating socially acceptable technologies is an essential aspect of assistive augmentation, appropriate measures should be employed in case potential biases are found. Finally, developing guidelines and policies to ensure the ethical use and equal access to assistive augmentation technologies is another important challenge that should be addressed in future work.

Bibliography

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep Learning with Differential Privacy. In *Computer and Communications Security (CCS), Conference Proceedings*, pages 308–318. ACM.

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., and Amedi, A. (2014). EyeMusic: Introducing a Visual Colorful Experience for the Blind Using Auditory Sensory Substitution. *Restorative Neurology and Neuroscience*.

Abdelrahman, Y., Knierim, P., Wozniak, P. W., Henze, N., and Schmidt, A. (2017a). See through the Fire: Evaluating the Augmentation of Visual Perception of Firefighters Using Depth and Thermal Cameras. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '17, page 693–696, New York, NY, USA. Association for Computing Machinery.

Abdelrahman, Y., Schmidt, A., and Knierim, P. (2017b). Snake View: Exploring Thermal Imaging as a Vision Extender in Mountains. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '17, page 1067–1071, New York, NY, USA. Association for Computing Machinery.

Abhang, P. A., Gawali, B. W., and Mehrotra, S. C. (2016). Technological Basics of EEG Recording and Operation of Apparatus. In *Introduction to EEG- and Speech-Based Emotion Recognition*, pages 19–50. Elsevier.

Adib, F., Mao, H., Kabelac, Z., Katabi, D., and Miller, R. C. (2015). Smart Homes that Monitor Breathing and Heart Rate. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 837–846. ACM.

Ahmetovic, D., Sato, D., Oh, U., Ishihara, T., Kitani, K., and Asakawa, C. (2020). ReCog: Supporting Blind People in Recognizing Personal Objects. In *Human Fac-*

tors in Computing Systems (CHI), Conference Proceedings, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Alt, F., Pfleging, B., and Schmidt, A. (2013). Sonify - A Platform for the Sonification of Text Messages. In Boll, S., Maaß, S., and Malaka, R., editors, *Mensch & Computer 2013: Interaktive Vielfalt*, pages 149–158, München. Oldenbourg Verlag.

Alt, F., Shirazi, A. S., Legien, S., Schmidt, A., and Mennenöh, J. (2010). Creating Meaningful Melodies from Text Messages. In Beilharz, K., Bongers, B., Johnston, A., and Ferguson, S., editors, *New Interfaces for Musical Expression (NIME), Conference Proceedings*, pages 63–68.

Altun, K. and Barshan, B. (2010). Human Activity Recognition Using Inertial/Magnetic Sensor Units. In *Human-Behavior Understanding*, pages 38–51. Springer.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Publishing, Washington, DC, 5th edition.

Amini, N., Lim, J. S., Mohammadi, F., Thodos, C., Braun, B., Ghasemzadeh, H., Chun, M. W., and Nouri-Mahdavi, K. (2020). Design and Evaluation of a Wearable Assistive Technology for Hemianopic Stroke Patients. In *International Symposium on Wearable Computers (ISWC), Conference Proceedings*, ISWC '20, page 7–11, New York, NY, USA. Association for Computing Machinery.

An, P., Holstein, K., d'Anjou, B., Eggen, B., and Bakker, S. (2020). The TA Framework: Designing Real-time Teaching Augmentation for K-12 Classrooms. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 1–17. ACM.

Anderson, I. M., Haddad, P. M., and Scott, J. (2012). Bipolar Disorder. *BMJ (Clinical Research Edition)*, 345:e8508.

Anikin, A., Bååth, R., and Persson, T. (2018). Human Non-linguistic Vocal Repertoire: Call Types and Their Meaning. *Journal of Nonverbal Behavior*, 42(1):53–80.

Aoki, P. M., Romaine, M., Szymanski, M. H., Thornton, J. D., Wilson, D., and Woodruff, A. (2003). The Mad Hatter's Cocktail Party: A Social Mobile Audio Space Supporting Multiple Simultaneous Conversations. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '03, page 425–432, New York, NY, USA. Association for Computing Machinery.

Argyle, M. and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press, Oxford, England.

Arts, N. J., Walvoort, S. J., and Kessels, R. P. (2017). Korsakoff's Syndrome: A Critical Review. *Neuropsychiatric Disease and Treatment*, 13:2875–2890.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In *Psychology of Learning and Motivation*, volume 2, pages 89–195. Academic Press.

Baddeley, A. (2000). The Episodic Buffer: A New Component of Working Memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Baddeley, A., Eysenck, M. W., and Anderson, M. C. (2015). *Memory*. Psychology Press, London and New York, 2nd edition.

Baddeley, A. D. and Hitch, G. (1974). Working Memory. In *Psychology of Learning and Motivation*, volume 8, pages 47–89. Elsevier.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2016). OpenFace: An Open Source Facial Behavior Analysis Toolkit. In *Winter Conference on Applications of Computer Vision (WACV), Conference Proceedings*, pages 1–10. IEEE.

Banf, M. and Blanz, V. (2012). A Modular Computer Vision Sonification Model for the Visually Impaired. In *International Conference on Auditory Display (ICAD), Conference Proceedings*, Atlanta, USA.

Banf, M. and Blanz, V. (2013). Sonification of Images for the Visually Impaired using a Multi-Level Approach. In *Augmented Human (AH), Conference Proceedings*, pages 162–169. ACM Press.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6):574–594.

Bannach, D., Amft, O., and Lukowicz, P. (2008). Rapid Prototyping of Activity Recognition Applications. *IEEE Pervasive Computing*, 7(2):22–31.

Banos, O., Villalonga, C., Garcia, R., Saez, A., Damas, M., Holgado-Terriza, J. A., Lee, S., Pomares, H., and Rojas, I. (2015). Design, Implementation and Validation of a Novel Open Framework for Agile Development of Mobile Health Applications. *Biomedical Engineering Online*, 14 Suppl 2(Suppl 2):S6.

Bao, L. and Intille, S. S. (2004). Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing*, volume 3001, pages 1–17. Springer.

Barrett, L. F. and Barrett, D. J. (2001). An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review*, 19(2):175–185.

Barrios, L., Oldrati, P., Santini, S., and Lutterotti, A. (2019). Evaluating the Accuracy of Heart Rate Sensors Based on Photoplethysmography for In-the-Wild Analysis. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 251–261. ACM.

Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 279–283. ACM.

Barz, M., Bhatti, O. S., Lüers, B., Prange, A., and Sonntag, D. (2021). Multisensor-Pipeline: A Lightweight, Flexible, and Extensible Framework for Building Multimodal-Multisensor Interfaces. In *Companion Publication of the International Conference on Multimodal Interaction (ICMI Companion), Conference Proceedings*, pages 13–18. ACM.

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., and André, E. (2013). NovA: Automated Analysis of Nonverbal Signals in Social Interactions. In *Human Behavior Understanding*, volume 8212 of *Lecture Notes in Computer Science*, pages 160–171. Springer International Publishing, Cham.

Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., and André, E. (2020). eXplainable Cooperative Machine Learning with NOVA. *KI - Künstliche Intelligenz*, 34(2):143–164.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., and Grundmann, M. (2020). BlazePose: On-device Real-time Body Pose Tracking.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M. (2019). BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs.

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Pearson, San Antonio, 2nd edition.

Berning, M., Braun, F., Riedel, T., and Beigl, M. (2015). ProximityHat: A Head-Worn System for Subtle Sensory Augmentation with Tactile Stimulation. In *International Symposium on Wearable Computers (ISWC), Conference Proceedings*, ISWC '15, page 31–38, New York, NY, USA. Association for Computing Machinery.

Biasiucci, A., Franceschiello, B., and Murray, M. M. (2019). Electroencephalography. *Current Biology*, 29(3):R80–R85.

Bixler, R. and D'Mello, S. (2014). Toward Fully Automated Person-Independent Detection of Mind Wandering. In *User Modeling, Adaptation, and Personalization*, volume 8538, pages 37–48. Springer.

Bixler, R. and D'Mello, S. (2016). Automatic Gaze-Based User-Independent Detection of Mind Wandering During Computerized Reading. *User Modeling and User-Adapted Interaction*, 26(1):33–68.

Boersma, P. (2001). PRAAT, A System for Doing Phonetics by Computer. *Glot International*, 5(9/10):341–345.

Bohus, D., Andrist, S., Feniello, A., Saw, N., Jalobeanu, M., Sweeney, P., Thompson, A. L., and Horvitz, E. (2021). Platform for Situated Intelligence.

Boldu, R., Dancu, A., Matthies, D. J., Buddhika, T., Siriwardhana, S., and Nanayakkara, S. (2018). FingerReader2.0: Designing and Evaluating a Wearable Finger-Worn Camera to Assist People with Visual Impairments While Shopping. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 2(3).

Boldu, R., Matthies, D. J., Zhang, H., and Nanayakkara, S. (2020). AiSee: An Assistive Wearable Device to Support Visually Impaired Grocery Shoppers. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 4(4).

Bonawitz, K., Kairouz, P., McMahan, B., and Ramage, D. (2021). Federated Learning and Privacy. *Queue*, 19(5):87–114.

Boucsein, W. (2012). *Electrodermal Activity*. Springer, New York and Heidelberg, 2nd edition.

Boulemtafes, A., Derhab, A., and Challal, Y. (2020). A Review of Privacy-Preserving Techniques for Deep Learning. *Neurocomputing*, 384:21–45.

Boyd, L. E., Jiang, X., and Hayes, G. R. (2017). ProCom: Designing and Evaluating a Mobile and Wearable System to Support Proximity Awareness for People with Autism. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '17, page 2865–2877, New York, NY, USA. Association for Computing Machinery.

Braisby, N. and Gellatly, A., editors (2005). *Cognitive Psychology*. Oxford Univ. Press, Oxford, 1st edition.

Brandimonte, M. A., Einstein, G. O., and McDaniel, M. A. (2014). *Prospective Memory*. Psychology Press.

Bresó, A., Martínez-Miranda, J., Botella, C., Baños, R. M., and García-Gómez, J. M. (2016). Usability and Acceptability Assessment of an Empathic Virtual Agent to Prevent Major Depression. *Expert Systems*, 33(4):297–312.

Brigden, R. L. (1933). A Tachistoscopic Study of the Differentiation of Perception. *Psychological Monographs: General and Applied*, 44(1):153–166.

Broadbent, D. E. (1958). *Perception and Communication*. Pergamon Press, Oxford, 3rd reprint edition.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR), Conference Proceedings*.

Brock, M. and Kristensson, P. O. (2013). Supporting Blind Navigation using Depth Sensing and Sonification. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 255–258. ACM.

Brooke, J. (1996). SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–7.

Brown, C. (2006). *Cognitive Psychology*. Sage Course Companions. Sage Publications, Thousand Oaks, Calif and London, 1st edition.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish,

S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Brunette, W., Sodt, R., Chaudhri, R., Goel, M., Falcone, M., van Orden, J., and Borriello, G. (2012). Open Data Kit Sensors: A Sensor Integration Framework for Android at the Application-Level. In *Mobile Systems, Applications, and Services (MobiSys), Conference Proceedings*, pages 351–364. ACM.

Buchs, G., Maidenbaum, S., and Amedi, A. (2015). Augmented Non-Visual Distance Sensing with the EyeCane. In *Augmented Human (AH), Conference Proceedings*, AH '15, page 209–210, New York, NY, USA. Association for Computing Machinery.

Bulling, A., Ward, J. A., Gellersen, H., and Tröster, G. (2011). Eye Movement Analysis for Activity Recognition using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):741–753.

Burns, M. S. (2004). Clinical Management of Agnosia. *Topics in Stroke Rehabilitation*, 11(1):1–9.

Burton, C., Szentagotai Tatar, A., McKinstry, B., Matheson, C., Matu, S., Moldovan, R., Macnab, M., Farrow, E., David, D., Pagliari, C., Serrano Blanco, A., and Wolters, M. (2016). Pilot Randomised Controlled Trial of Help4Mood, an Embodied Virtual Agent-based System to Support Treatment of Depression. *Journal of telemedicine and telecare*, 22(6):348–355.

Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology of Perception in Art*. University of Chicago Press, Chicago, Illinois.

Cai, H., Yuan, Z., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., Li, J., Yang, Z., Li, X., Zhao, Q., Liu, Z., Yao, Z., Yang, M., Peng, H., Zhu, J., Zhang, X., Gao, G., Zheng, F., Li, R., Guo, Z., Ma, R., Yang, J., Zhang, L., Hu, X., Li, Y., and Hu, B. (2022). A Multi-Modal Open Dataset for Mental-Disorder Analysis. *Scientific Data*, 9(1):178.

Can, Y. S., Mahesh, B., and André, E. (2023). Approaches, Applications, and Challenges in Physiological Emotion Recognition—A Tutorial Overview. *Proceedings of the IEEE*, 111(10):1287–1313.

Canali, S., Schiaffonati, V., and Aliverti, A. (2022). Challenges and Recommendations for Wearable Devices in Digital Health: Data Quality, Interoperability, Health Equity, Fairness. *PLOS digital health*, 1(10):e0000104.

Cañigueral, R. and Hamilton, A. F. d. C. (2019). The Role of Eye Gaze During Natural Social Interactions in Typical and Autistic People. *Frontiers in Psychology*, 10:560.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Carbonell, K. M. and Lotto, A. J. (2014). Speech is not Special... Again. *Frontiers in Psychology*, 5:427.

Carcedo, M. G., Chua, S. H., Perrault, S., Wozniak, P., Joshi, R., Obaid, M., Fjeld, M., and Zhao, S. (2016). HaptiColor: Interpolating Color Information as Haptic Feedback to Assist the Colorblind. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '16, page 3572–3583, New York, NY, USA. Association for Computing Machinery.

Cardone, G., Cirri, A., Corradi, A., Foschini, L., and Maio, D. (2013). MSF: An Efficient Mobile Phone Sensing Framework. *International Journal of Distributed Sensor Networks*, 9(3):538937.

Carlson, D. and Schrader, A. (2012). Dynamix: An Open Plug-and-Play Context Framework for Android. In *Internet of Things (IoT), Conference Proceedings*, pages 151–158. IEEE.

Carriço, L., de Sá, M., Duarte, L., and Antunes, T. (2012). Therapy: Location-Aware Assessment and Tasks. In *Augmented Human (AH), Conference Proceedings*, AH '12, New York, NY, USA. Association for Computing Machinery.

Carter, S. and Mankoff, J. (2005). When Participants Do the Capturing: The Role of Media in Diary Studies. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 899–908. ACM.

Carton, A. and Dunne, L. E. (2013). Tactile Distance Feedback for Firefighters: Design and Preliminary Evaluation of a Sensory Augmentation Glove. In *Augmented Human (AH), Conference Proceedings*, AH '13, page 58–64, New York, NY, USA. Association for Computing Machinery.

Castelhano, M. S., Mack, M. L., and Henderson, J. M. (2009). Viewing Task Influences Eye Movement Control During Active Scene Perception. *Journal of Vision*, 9(3):6.1–15.

Chabanne, H., de Wargny, A., Milgram, J., Morel, C., and Prouff, E. (2017). Privacy-Preserving Classification on Deep Neural Network. *Cryptology ePrint Archive*.

Chang, Y.-J., Chen, C.-N., Chou, L.-D., and Wang, T.-Y. (2008). A Novel Indoor Wayfinding System Based on Passive RFID for Individuals with Cognitive Impairments. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 108–111.

Chang, Y.-J., Paruthi, G., and Newman, M. W. (2015). A Field Study Comparing Approaches to Collecting Annotated Activity Data in Real-world Settings. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 671–682. ACM.

Chartrand, T. L. and Bargh, J. A. (1999). The Chameleon Effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

Chen, B., Zi, B., Wang, Z., Qin, L., and Liao, W.-H. (2019). Knee Exoskeletons for Gait Rehabilitation and Human Performance Augmentation: A State-of-the-art. *Mechanism and Machine Theory*, 134:499–511.

Chen, K., Huang, Y., Chen, Y., Zhong, H., Lin, L., Wang, L., and Wu, K. (2022). LiSee: A Headphone That Provides All-Day Assistance for Blind and Low-Vision Users to Reach Surrounding Objects. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 6(3).

Chen, Y. and Jones, G. J. F. (2010). Augmenting Human Memory Using Personal Lifelogs. In *Augmented Human (AH), Conference Proceedings*, AH '10, New York, NY, USA. Association for Computing Machinery.

Chen, Y., Qin, X., Wang, J., Yu, C., and Gao, W. (2020). FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. *IEEE Intelligent Systems*, 35(4):83–93.

Chen, Y.-P., Yang, J.-Y., Liou, S.-N., Lee, G.-Y., and Wang, J.-S. (2008). Online Classifier Construction Algorithm for Human Activity Detection Using a Tri-axial Accelerometer. *Applied Mathematics and Computation*, 205(2):849–860.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 1800–1807. IEEE.

Chomsky, N. (1959). A Review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Christin, D., Reinhardt, A., Kanhere, S. S., and Hollick, M. (2011). A Survey on Privacy in Mobile Participatory Sensing Applications. *Journal of Systems and Software*, 84(11):1928–1946.

Ciliberto, M., Morales, F. J. O., Gjoreski, H., Roggen, D., Mekki, S., and Valentin, S. (2017). High Reliability Android Application for Multidevice Multimodal Mobile Data Acquisition and Annotation. In *Embedded Network Sensor Systems (SenSys), Conference Proceedings*, pages 1–2. ACM.

Coco, M. I. and Keller, F. (2014). Classification of Visual and Linguistic Tasks Using Eye-Movement Features. *Journal of Vision*, 14(3):11.1–18.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, 24(4):385.

Costa, M., Dinsbach, W., Manstead, A. S. R., and Bitti, P. E. R. (2001). Social Presence, Embarrassment, and Nonverbal Behavior. *Journal of Nonverbal Behavior*, 25(4):225–240.

Costandi, M. (2016). *Neuroplasticity*. The MIT Press Essential Knowledge Series. The MIT Press, Cambridge, MA.

Covaco, S., Henriques, J., Mengucci, M., Correia, N., and Medeirous, F. (2013). Color Sonification for the Visually Impaired. In *Procedia Technology*, pages 1048–1057, Amsterdam, Netherlands. Elsevier.

Cowie, R., Sawey, M., Doherty, C., Jaimovich, J., Fyans, C., and Stapleton, P. (2013). GTrace: General Trace Program Compatible with EmotionML. In *Affective Computing and Intelligent Interaction (ACII), Conference Proceedings*, pages 709–710. IEEE.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., and Anderson, N. D. (1996). The Effects of Divided Attention on Encoding and Retrieval Processes in Human Memory. *Journal of Experimental Psychology. General*, 125(2):159–180.

Craik, F. I. and Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Wiley.

Damasio, A. R. (1992). Aphasia. *The New England Journal of Medicine*, 326(8):531–539.

Damian, I. (2017). *Social Augmentation Using Behavioural Feedback Loops*. Dissertation, Universität Augsburg, Augsburg.

Damian, I., Dietz, M., and André, E. (2018). The SSJ Framework: Augmenting Social Interactions Using Mobile Signal Processing and Live Feedback. *Frontiers in ICT*, 5.

Damian, I., Dietz, M., Gaibler, F., and André, E. (2016). Social Signal Processing for Dummies. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 394–395. ACM.

Damian, I., Tan, C. S., Baur, T., Schöning, J., Luyten, K., and André, E. (2015). Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 565–574. ACM.

Danry, V., Pataranutaporn, P., Mao, Y., and Maes, P. (2020). Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant. In *Augmented Humans (AHs), Conference Proceedings*, AHs '20, New York, NY, USA. Association for Computing Machinery.

Darwin, C. (1872). *Expression of the Emotions in Man and Animals*. John Murray, London.

Darwin, C. J., Turvey, M. T., and Crowder, R. G. (1972). An Auditory Analogue of the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage. *Cognitive Psychology*, 3(2):255–267.

Davis, M. and Hadiks, D. (1994). Nonverbal Aspects of Therapist Attunement. *Journal of Clinical Psychology*, 50(3):393–405.

Dawson, M. E., Schell, A. M., Filion, D. L., and Berntson, G. G. (2007). The Electrodermal System. In *Handbook of Psychophysiology*, pages 157–181. Cambridge University Press, Cambridge.

de Araujo, M. (2017). Editing the Genome of Human Beings: CRISPR-Cas9 and the Ethics of Genetic Enhancement. *Journal of Evolution and Technology*, 27(1):24–42.

de Boeck, M. and Vaes, K. (2021). Structuring Human Augmentation Within Product Design. *Proceedings of the Design Society*, 1:2731–2740.

de Siati, R. D., Rosenzweig, F., Gersdorff, G., Gregoire, A., Rombaux, P., and Deggouj, N. (2020). Auditory Neuropathy Spectrum Disorders: From Diagnosis to Treatment: Literature Review and Case Reports. *Journal of Clinical Medicine*, 9(4).

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, 93(3):283–321.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 248–255. IEEE.

Deroy, O. and Auvray, M. (2012). Reading the World through the Skin and Ears: A New Perspective on Sensory Substitution. *Frontiers in Psychology*, 3:457.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some Theoretical Considerations. *Psychological Review*, 70:80–90.

Development, Concepts and Doctrine Centre (2021). Human Augmentation - The Dawn of a New Paradigm: A Strategic Implications Project. Technical report, UK Ministry of Defence.

Dibeklioglu, H., Hammal, Z., and Cohn, J. F. (2018). Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 22(2):525–536.

Dietz, M., Aslan, I., Schiller, D., Flutura, S., Steinert, A., Klebbe, R., and André, E. (2019). Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 149–158. ACM.

Dietz, M., Elgarf, M., Damian, I., and André, E. (2016a). Exploring Eye-Tracking-Driven Sonification for the Visually Impaired. In *Augmented Human (AH), Conference Proceedings*, pages 1–8. ACM.

Dietz, M., Schork, D., and André, E. (2016b). Exploring Eye-Tracking-Based Detection of Visual Search for Elderly People. In *Intelligent Environments (IE), Conference Proceedings*, pages 151–154. IEEE.

Dietz, M., Schork, D., Damian, I., Steinert, A., Haesner, M., and André, E. (2017). Automatic Detection of Visual Search for the Elderly using Eye and Head Tracking Data. *KI - Künstliche Intelligenz*, 31(4):339–348.

Dingler, T., Agroudy, P. E., Rzayev, R., Lischke, L., Machulla, T., and Schmidt, A. (2021). Memory Augmentation Through Lifelogging: Opportunities and Challenges. In *Technology-Augmented Perception and Cognition*, pages 47–69. Springer International Publishing, Cham.

Dodge, M. and Kitchin, R. (2007). 'Outlines of a World Coming into Existence': Pervasive Computing and the Ethics of Forgetting. *Environment and Planning B: Planning and Design*, 34(3):431–445.

Drummond, J. and Litman, D. (2010). In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning. In *Intelligent Tutoring Systems*, volume 6095 of *Lecture Notes in Computer Science*, pages 306–308. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *FNT in Theoretical Computer Science (Foundations and Trends in Theoretical Computer Science)*, 9(3-4):211–407.

Egede, J. O., Price, D., Krishnan, D. B., Jaiswal, S., Elliott, N., Morriss, R., Trigo, M. J. G., Nixon, N., Liddle, P., Greenhalgh, C., and Valstar, M. (2021). Design

and Evaluation of Virtual Human Mediated Tasks for Assessment of Depression and Anxiety. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, pages 52–59. ACM.

Eghtebas, C., Kiss, F., Koelle, M., and Woźniak, P. (2021). Advantage and Misuse of Vision Augmentation – Exploring User Perceptions and Attitudes Using a Zoom Prototype. In *Augmented Humans (AHs), Conference Proceedings*, AHs '21, page 77–85, New York, NY, USA. Association for Computing Machinery.

Einstein, G. O. and McDaniel, M. A. (1990). Normal Aging and Prospective Memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 16(4):717–726.

Ekman, P. (1984). Expression and the Nature of Emotion. In *Approaches to Emotion*, pages 319–344. Psychology Press, New York.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ekman, P. (1999a). Basic Emotions. In *Handbook of Cognition and Emotion*, pages 45–60. Wiley.

Ekman, P. (1999b). Facial Expressions. In *Handbook of Cognition and Emotion*, pages 301–320. Wiley.

Ekman, P. (2009). Lie Catching and Microexpressions. In *The Philosophy of Deception*, pages 118–136. Oxford University Press.

Ekman, P. and Friesen, W. V. (1969a). Nonverbal Leakage and Clues to Deception. *Psychiatry*, 32(1):88–106.

Ekman, P. and Friesen, W. V. (1969b). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1(1):49–98.

Ekman, P. and Friesen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto.

Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework. Summary report, Stanford Research Institute, Menlo Park California.

Epshtein, B., Ofek, E., and Wexler, Y. (2010). Detecting Text in Natural Scenes with Stroke Width Transform. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 2963–2970. IEEE.

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, 49(8):709–724.

Ericcson (2024). Ericsson Mobility Report. Technical report, Ericcson.

Eriksson, M. and Papanikotopoulos, N. P. (1997). Eye-Tracking for Detection of Driver Fatigue. In *Intelligent Transportation Systems (ITSC), Conference Proceedings*, pages 314–319. IEEE.

Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., and Shah, S. (2011). AutoSense: Unobtrusively Wearable Sensor Suite for Inferring the Onset, Causality, and Consequences of Stress in the Field. In *Embedded Networked Sensor Systems (SenSys), Conference Proceedings*, pages 274–287. ACM.

Evans, J. S. B. T. and Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 8(3):223–241.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and 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.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). OpenSmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *ACM Multimedia (MM), Conference Proceedings*, pages 1459–1462. ACM.

Eysenck, M. W. and Keane, M. T. (2020). *Cognitive Psychology: A Student's Handbook*. A Psychology Press book. Psychology Press Taylor & Francis Group, London and New York, 8th edition.

Fan, K., Huber, J., Nanayakkara, S., and Inami, M. (2014). SpiderVision: Extending the Human Field of View for Augmented Awareness. In *Augmented Human (AH), Conference Proceedings*, AH '14, New York, NY, USA. Association for Computing Machinery.

Farber, B. A. (2003). Self-Disclosure in Psychotherapy Practice and Supervision: An Introduction. *Journal of Clinical Psychology*, 59(5):525–528.

Fardoun, H. M., González, L. C., and Mashat, A. S. (2013). Rehabilitation Low Vision Algorithm: For People with Central or Multiple Losses of Vision. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, PervasiveHealth '13, page 339–343, Brussels, BEL. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Feiz, S., Billah, S. M., Ashok, V., Shilkrot, R., and Ramakrishnan, I. (2019). Towards Enabling Blind People to Independently Write on Printed Forms. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Ferreira, D., Kostakos, V., and Dey, A. K. (2015). AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT*, 2.

Findlay, J. M. and Gilchrist, I. D. (1998). Eye Guidance and Visual Search. In *Eye Guidance in Reading and Scene Perception*, pages 295–312. Elsevier, Amsterdam and New York.

Flatla, D. R., Andrade, A. R., Teviotdale, R. D., Knowles, D. L., and Stewart, C. (2015). ColourID: Improving Colour Identification for People with Impaired Colour Vision. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '15, page 3543–3552, New York, NY, USA. Association for Computing Machinery.

Fletcher, R. R., Tam, S., Omojola, O., Redemske, R., Fedor, S., and Moshoka, J. M. (2011). Mobile Application and Wearable Sensors for use in Cognitive Behavioral Therapy for Drug Addiction and PTSD. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 202–203.

Flobak, E., Guribye, F., Jensen, D. A., and Lundervold, A. J. (2017). Designing Data-Driven Interventions for Mental Health Care. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, PervasiveHealth '17, page 423–426, New York, NY, USA. Association for Computing Machinery.

Flores, G. and Manduchi, R. (2018). Easy Return: An App for Indoor Backtracking Assistance. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Flutura, S., Wagner, J., Lingenfelser, F., Seiderer, A., and André, E. (2016). MobileSSI: Asynchronous Fusion for Social Signal Interpretation in the Wild. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 266–273. ACM.

Frackowiak, R. S., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T., and Penny, W. D. (2004). *Human Brain Function*. Elsevier Professional, s.l., 2nd edition.

Freeman, E., Wilson, G., Brewster, S., Baud-Bovy, G., Magnusson, C., and Caltenco, H. (2017). Audible Beacons and Wearables in Schools: Helping Young Visually Impaired Children Play and Move Independently. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '17, page 4146–4157, New York, NY, USA. Association for Computing Machinery.

Fridlund, A. J., Schwartz, G. E., and Fowler, S. C. (1984). Pattern Recognition of Self-Reported Emotional State from Multiple-Site Facial EMG Activity During Affective Imagery. *Psychophysiology*, 21(6):622–637.

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze Cueing of Attention: Visual attention, Social Cognition, and Individual Differences. *Psychological Bulletin*, 133(4):694–724.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. A. (2007). MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones. In *Mobile Systems, Applications and Services (MobiSys), Conference Proceedings*, pages 57–70. ACM.

Funke, J. (2010). Complex Problem Solving: A Case for Complex Cognition? *Cognitive Processing*, 11(2):133–142.

Féré, C. (1888). Note on Changes in Electrical Resistance Under the Effect of Sensory Stimulation and Emotion. *Comptes rendus des seances de la societé de biologie*, 5:217–219.

Gao, Z., Cui, X., Wan, W., Zheng, W., and Gu, Z. (2021). ECSMP: A Dataset on Emotion, Cognition, Sleep, and Multi-Model Physiological Signals. *Data in brief*, 39:107660.

Garrett, M. F. (1975). The Analysis of Sentence Production. In *Psychology of Learning and Motivation*, volume 9, pages 133–177. Academic Press.

Gebhard, P., Mehlmann, G., and Kipp, M. (2012). Visual SceneMaker—A Tool for Authoring Interactive Virtual Characters. *Journal on Multimodal User Interfaces*, 6(1-2):3–11.

Gebhard, P., Schneeberger, T., Dietz, M., André, E., and Bajwa, N. u. H. (2019). Designing a Mobile Social and Vocational Reintegration Assistant for Burn-out Outpatient Treatment. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, pages 13–15. ACM.

Gidlöf, K., Wallin, A., Dewhurst, R., and Holmqvist, K. (2013). Using Eye Tracking to Trace a Cognitive Process: Gaze Behaviour During Decision Making in a Natural Environment. *Journal of Eye Movement Research*, 6(1).

Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., and Wernsing, J. (2016). CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 201–210, New York, New York, USA. PMLR.

Gillian, N. and Paradiso, J. A. (2014). The Gesture Recognition Toolkit. *Journal of Machine Learning Research*, 15(101):3483–3487.

Ginelli, D., Micucci, D., Mobilio, M., and Napoletano, P. (2018). UniMiB AAL: An Android Sensor Data Acquisition and Labeling Suite. *Applied Sciences*, 8(8):1265.

Giubilini, A. and Sanyal, S. (2015). The Ethics of Human Enhancement. *Philosophy Compass*, 10(4):233–243.

Glaser, B. G. (1965). The Constant Comparative Method of Qualitative Analysis. *Social Problems*, 12(4):436–445.

Glenn, T. and Monteith, S. (2014). Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections. *Current psychiatry reports*, 16(11):494.

Goldinger, S. D. and Papesh, M. H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science*, 21(2):90–95.

Goldstein, E. B. (2010). *Sensation and Perception*. Wadsworth Cengage Learning, Belmont, Calif., 8th edition.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Interspeech, Conference Proceedings*, pages 571–575. ISCA.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts and London, England.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Goodman, S., Kirchner, S., Guttman, R., Jain, D., Froehlich, J., and Findlater, L. (2020). Evaluating Smartwatch-Based Sound Feedback for Deaf and Hard-of-Hearing Users Across Contexts. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Gouveia, R. and Karapanos, E. (2013). Footprint Tracker: Supporting Diary Studies with Lifelogging. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '13, page 2921–2930, New York, NY, USA. Association for Computing Machinery.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., Devault, D., Marsella, S., Traum, D., Rizzo, A. S., and Morency, L.-P. (2014). The Distress Analysis Interview Corpus of Human and Computer Interviews. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Language Resources and Evaluation (LREC), Conference Proceedings*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gravenhorst, F., Muaremi, A., Bardram, J., Grünerbl, A., Mayora, O., Wurzer, G., Frost, M., Osmani, V., Arnrich, B., Lukowicz, P., and Tröster, G. (2015). Mobile Phones as Medical Devices in Mental Disorder Treatment: An Overview. *Personal and Ubiquitous Computing*, 19(2):335–353.

Greene, M. R., Liu, T., and Wolfe, J. M. (2012). Reconsidering Yarbus: A Failure to Predict Observers' Task from Eye Movement Patterns. *Vision Research*, 62:1–8.

Groome, D. (2014). *An Introduction to Cognitive Psychology: Processes and Disorders*. Psychology Press, Hove, East Sussex, 3rd edition.

GSMA (2023). The State of Mobile Internet Connectivity 2024. Technical report, GSM Association.

Gu, F., Chung, M.-H., Chignell, M., Valaee, S., Zhou, B., and Liu, X. (2022). A Survey on Deep Learning for Human Activity Recognition. *ACM Computing Surveys*, 54(8):1–34.

Guan, Y. and Plötz, T. (2017). Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 1(2):1–28.

Guerreiro, J. a., Ahmetovic, D., Sato, D., Kitani, K., and Asakawa, C. (2019). Airport Accessibility and Navigation Assistance for People with Visual Impairments. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Guilbourd, R., Yogev, N., and Rojas, R. (2012). Stereo Camera Based Wearable Reading Device. In *Augmented Human (AH), Conference Proceedings*, AH '12, New York, NY, USA. Association for Computing Machinery.

Guy, R. and Truong, K. (2012). CrossingGuard: Exploring Information Content in Navigation Aids for Visually Impaired Pedestrians. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '12, page 405–414, New York, NY, USA. Association for Computing Machinery.

Habermas, J. (2003). *The Future of Human Nature*. Polity Press.

Haggard, E. A. and Isaacs, K. S. (1966). Micromomentary Facial Expressions as Indicators of Ego Mechanisms in Psychotherapy. In *Methods of Research in Psychotherapy*, pages 154–165. Springer US, Boston, MA.

Hamilton, M. A., Beug, A. P., Hamilton, H. J., and Norton, W. J. (2021). Augmented Reality Technology for People Living with Dementia and their Care Partners. In *International Conference on Virtual and Augmented Reality Simulations (ICVARS), Conference Proceedings*, pages 21–30. Association for Computing Machinery.

Han, J., Zhang, Z., Mascolo, C., Andre, E., Tao, J., Zhao, Z., and Schuller, B. W. (2021). Deep Learning for Mobile Mental Health: Challenges and Recent Advances. *IEEE Signal Processing Magazine*, 38(6):96–105.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). GhostNet: More Features From Cheap Operations. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 1577–1586. IEEE.

Hanai, Y., Nishimura, J., and Kuroda, T. (2009). Haar-Like Filtering for Human Activity Recognition Using 3D Accelerometer. In *13th Digital Signal Processing Workshop*, pages 675–678. IEEE.

Harley, T. A. (2014). *The Psychology of Language: From Data to Theory*. Psychology Press, London and New York, 4th edition.

Harnad, S. R., editor (1990). *Categorical Perception: The Groundwork of Cognition*. Cambridge Univ. Press, Cambridge, 1st edition.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. Elsevier.

Harvey, M., Langheinrich, M., and Ward, G. (2016). Remembering through Lifelogging: A Survey of Human Memory Augmentation. *Pervasive and Mobile Computing*, 27:14–26.

Hasegawa, S., Ishijima, S., Kato, F., Mitake, H., and Sato, M. (2012). Realtime Sonification of the Center of Gravity for Skiing. In *Augmented Human (AH), Conference Proceedings*, AH '12, New York, NY, USA. Association for Computing Machinery.

Hawes, M. T., Szenczy, A. K., Klein, D. N., Hajcak, G., and Nelson, B. D. (2022). Increases in Depression and Anxiety Symptoms in Adolescents and Young Adults During the COVID-19 Pandemic. *Psychological Medicine*, 52(14):3222–3230.

He, Z. and Jin, L. (2009). Activity Recognition from Acceleration Data based on Discrete Consine Transform and SVM. In *Systems, Man and Cybernetics (SMC), Conference Proceedings*, pages 5041–5044. IEEE.

Heimerl, A., Weitz, K., Baur, T., and Andre, E. (2022). Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing*, 13(3):1155–1167.

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., and Olejarczyk, J. (2013). Predicting Cognitive State from Eye Movements. *PloS one*, 8(5):e64937.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427.

Hills, P. J. (2016). *Cognitive Psychology for Dummies*. John Wiley & Sons Incorporated, Newark, NJ.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., and Wood, K. (2006). SenseCam: A Retrospective Memory Aid. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, volume 4206, pages 177–193. Springer.

Hogrel, J.-Y. (2005). Clinical Applications of Surface Electromyography in Neuromuscular Disorders. *Neurophysiologie Clinique/Clinical Neurophysiology*, 35(2-3):59–71.

Hoisko, J. (1999). Using Wearable Computer as an Audiovisual Memory Prosthesis. In Gellersen, H.-W., editor, *Handheld and Ubiquitous Computing*, pages 317–318, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hossain, S. M., Hnat, T., Saleheen, N., Nasrin, N. J., Noor, J., Ho, B.-J., Condie, T., Srivastava, M., and Kumar, S. (2017). mCerebrum: A Mobile Sensing Software Platform for Development and Validation of Digital Biomarkers and Interventions. In *Embedded Network Sensor Systems (SenSys), Conference Proceedings*, volume 2017. ACM.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q. (2019). Searching for MobileNetV3. In *International Conference on Computer Vision (ICCV), Conference Proceedings*, pages 1314–1324. IEEE.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

Howe, E., Suh, J., Bin Morshed, M., McDuff, D., Rowan, K., Hernandez, J., Abdin, M. I., Ramos, G., Tran, T., and Czerwinski, M. P. (2022). Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Huber, J., Shilkrot, R., Maes, P., and Nanayakkara, S. (2018). *Assistive Augmentation*. Springer Singapore, Singapore.

Hutt, S., Krasich, K., R. Brockmole, J., and K. D'Mello, S. (2021). Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Huynh, T. and Schiele, B. (2005). Analyzing Features for Activity Recognition. In Bailly, G. and Crowley, J. L., editors, *Smart Objects and Ambient Intelligence (sOc-EUSAI), Conference Proceedings*, page 159.

Ibraheem, N. A. and Khan, R. Z. (2012). Survey on Various Gesture Recognition Technologies and Techniques. *International Journal of Computer Applications*, 50(7):38–44.

Itoh, Y. and Klinker, G. (2015). Vision Enhancement: Defocus Correction via Optical See-through Head-Mounted Displays. In *Augmented Human (AH), Conference Proceedings*, AH '15, page 1–8, New York, NY, USA. Association for Computing Machinery.

Itoh, Y., Orlosky, J., Kiyokawa, K., and Klinker, G. (2016). Laplacian Vision: Augmenting Motion Prediction via Optical See-Through Head-Mounted Displays. In *Augmented Human (AH), Conference Proceedings*, AH '16, New York, NY, USA. Association for Computing Machinery.

Iwamura, M., Kunze, K., Kato, Y., Utsumi, Y., and Kise, K. (2014). Haven't We Met before? A Realistic Memory Assistance System to Remind You of the Person in Front of You. In *Augmented Human (AH), Conference Proceedings*, AH '14, New York, NY, USA. Association for Computing Machinery.

Jacob, R. J. and Karn, K. S. (2003). Eye Tracking in Human-Computer Interaction and Usability Research. In *The Mind's Eye*, pages 573–605. Elsevier.

Jain, D., Chiu, B., Goodman, S., Schmandt, C., Findlater, L., and Froehlich, J. E. (2020). Field Study of a Tactile Sound Awareness Device for Deaf Users. In *International Symposium on Wearable Computers (ISWC), Conference Proceedings*, ISWC '20, page 55–57, New York, NY, USA. Association for Computing Machinery.

Jain, D., Findlater, L., Gilkeson, J., Holland, B., Duraiswami, R., Zotkin, D., Vogler, C., and Froehlich, J. E. (2015). Head-Mounted Display Visualizations to Support

Sound Awareness for the Deaf and Hard of Hearing. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 241–250. ACM.

Jain, D., Huynh Anh Nguyen, K., M. Goodman, S., Grossman-Kahn, R., Ngo, H., Kusupati, A., Du, R., Olwal, A., Findlater, L., and E. Froehlich, J. (2022). ProtoSound: A Personalized and Scalable Sound Recognition System for Deaf and Hard-of-Hearing Users. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '22, New York, NY, USA. Association for Computing Machinery.

James, W. (1890). *The Principles of Psychology, Vol I.* Henry Holt and Co, New York.

Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). Physiological Signals Based Human Emotion Recognition: A Review. In *Colloquium on Signal Processing and its Applications (CSPA), Conference Proceedings*, pages 410–415. IEEE.

Juengst, E. and Moseley, D. (2019). Human Enhancement. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition.

Just, M. A. and Carpenter, P. A. (1976). Eye Fixations and Cognitive Processes. *Cognitive Psychology*, 8(4):441–480.

Kächele, M., Werner, P., Al-Hamadi, A., Palm, G., Walter, S., and Schwenker, F. (2015). Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity. In Schwenker, F., Roli, F., and Kittler, J., editors, *Multiple Classifier Systems (MCS), Conference Proceedings*, pages 220–230. Springer.

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall Series in Experimental Psychology. Prentice-Hall Inc, Englewood Cliffs, New Jersey.

Kahneman, D., Tursky, B., Shapiro, D., and Crider, A. (1969). Pupillary, Heart Rate, and Skin Resistance Changes During a Mental Task. *Journal of Experimental Psychology*, 79(1):164–167.

Kalnikaite, V., Sellen, A., Whittaker, S., and Kirk, D. (2010). Now Let Me See Where i Was: Understanding How Lifelogs Mediate Memory. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '10, page 2045–2054, New York, NY, USA. Association for Computing Machinery.

Kaniusas, E. (2012a). Fundamentals of Biosignals. In *Biomedical Signals and Sensors I*, Biological and Medical Physics, Biomedical Engineering, pages 1–26. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kaniusas, E. (2012b). Physiological Phenomena and Biosignals. In *Biomedical Signals and Sensors I*, Biological and Medical Physics, Biomedical Engineering, pages 183–282. Springer Berlin Heidelberg, Berlin, Heidelberg.

Karpathy, A. and Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 1725–1732. IEEE.

Kasahara, S., Ando, M., Suganuma, K., and Rekimoto, J. (2016). Parallel Eyes: Exploring Human Capability and Behaviors with Paralleled First Person View Sharing. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '16, page 1561–1572, New York, NY, USA. Association for Computing Machinery.

Kassner, M., Patera, W., and Bulling, A. (2014). Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 1151–1160, New York, NY, USA. ACM.

Kayukawa, S., Higuchi, K., Guerreiro, J. a., Morishima, S., Sato, Y., Kitani, K., and Asakawa, C. (2019). BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Kayukawa, S., Ishihara, T., Takagi, H., Morishima, S., and Asakawa, C. (2020). Guiding Blind Pedestrians in Public Spaces by Understanding Walking Behavior of Nearby Pedestrians. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 4(3).

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*, volume 7 of *Studies in Interactional Sociolinguistics*. Cambridge Univ. Press, Cambridge.

Kenyon, G. N. and Sen, K. C. (2015). The Perception Process. In *The Perception of Quality*, pages 41–50. Springer London, London.

Kern, D., Marshall, P., and Schmidt, A. (2010). Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '10, page 2093–2102, New York, NY, USA. Association for Computing Machinery.

Khan, A. M., Lee, Y.-K., and Lee, S. Y. (2010a). Accelerometer's Position Free Human Activity Recognition using a Hierarchical Recognition Model. In *E-Health Networking, Applications and Services (Healthcom), Conference Proceedings*, pages 296–301. IEEE.

Khan, A. M., Lee, Y.-K., Lee, S. Y., and Kim, T.-S. (2010b). A Triaxial Accelerometer-based Physical-activity Recognition via Augmented-signal Features and a Hierarchical Recognizer. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1166–1172.

Khan, M., Fernandes, G., and Maes, P. (2021). PAL: Wearable and Personalized Habit-Support Interventions in Egocentric Visual and Physiological Contexts. In *Augmented Humans (AHs), Conference Proceedings*, AHs '21, page 265–267, New York, NY, USA. Association for Computing Machinery.

Kianpisheh, M., Li, F. M., and Truong, K. N. (2019). Face Recognition Assistant for People with Visual Impairments. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 3(3).

Kipp, M. (2014). Anvil: The Video Annotation Research Tool. *Handbook of Corpus Phonology*, pages 420–436.

Kiss, F. (2020). *Reshaping Ubiquitous Interaction through Sensory Augmentation*. Dissertation, Universität Stuttgart, Stuttgart.

Kiss, F., Woźniak, P. W., Scheerer, F., Dominiak, J., Romanowski, A., and Schmidt, A. (2019). Clairbuoyance: Improving Directional Perception for Swimmers. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal Communication in Human Interaction*. Wadsworth Cengage Learning, Boston, MA and Australia and Brazil and Japan, 8th edition.

Knauer, U. and Seiffert, U. (2013). A Comparison of Late Fusion Methods for Object Detection. In *International Conference on Image Processing (ICIP), Conference Proceedings*, pages 3297–3301.

Knauff, M. and Wolf, A. G. (2010). Complex Cognition: The Science of Human Reasoning, Problem-Solving, and Decision-Making. *Cognitive Processing*, 11(2):99–102.

Knierim, P., Kosch, T., LaBorwit, G., and Schmidt, A. (2020). Altering the Speed of Reality? Exploring Visual Slow-Motion to Amplify Human Perception Using Augmented Reality. In *Augmented Humans (AHs), Conference Proceedings*, AHs '20, New York, NY, USA. Association for Computing Machinery.

Köhler, W. (1925). *The Mentality of Apes*. Harcourt, Brace, Oxford, England.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 60(6):84–90.

Kukkonen, J., Lagerspetz, E., Nurmi, P., and Andersson, M. (2009). BeTelGeuse: A Platform for Gathering and Processing Situational Data. *IEEE Pervasive Computing*, 8(2):49–56.

Kuribayashi, M., Kayukawa, S., Takagi, H., Asakawa, C., and Morishima, S. (2021). LineChaser: A Smartphone-Based Navigation System for Blind People to Stand in Lines. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Kurze, M. and Roselius, A. (2011). Smart Glasses Linking Real Live and Social Network's Contacts by Face Recognition. In *Augmented Human (AH), Conference Proceedings*, AH '11, New York, NY, USA. Association for Computing Machinery.

Lam, D., Rao, S. K., Ratra, V., Liu, Y., Mitchell, P., King, J., Tassignon, M.-J., Jonas, J., Pang, C. P., and Chang, D. F. (2015). Cataract. *Nature Reviews. Disease Primers*, 1:15014.

Lane, N., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. (2010). A Survey of Mobile Phone Sensing. *IEEE Communications Magazine*, 48(9):140–150.

Langlotz, T., Sutton, J., Zollmann, S., Itoh, Y., and Regenbrecht, H. (2018). ChromaGlasses: Computational Glasses for Compensating Colour Blindness. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Lara, O. D. and Labrador, M. A. (2013). A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209.

Larson, R. and Csikszentmihalyi, M. (1983). The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science*.

Lazarus, R. S. (1966). *Psychological Stress and the Coping Process*. McGraw-Hill, New York, NY, US.

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press, New York, NY, US.

Le Phong, T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. (2018). Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 105–114. IEEE.

Lee, H., Upright, C., Eliuk, S., and Kobsa, A. (2016). Personalized Object Recognition for Augmenting Human Memory. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '16, page 1054–1061, New York, NY, USA. Association for Computing Machinery.

Lee, K., Sato, D., Asakawa, S., Kacorri, H., and Asakawa, C. (2020). Pedestrian Detection with Wearable Cameras for the Blind: A Two-Way Perspective. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Lee, M. L. and Dey, A. K. (2008). Lifelogging Memory Appliance for People with Episodic Memory Impairment. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '08, page 44–53, New York, NY, USA. Association for Computing Machinery.

Leelasawassuk, T., Damen, D., and Mayol-Cuevas, W. (2017). Automated Capture and Delivery of Assistive Task Guidance with an Eyewear Computer: The GlaciAR System. In *Augmented Human (AH), Conference Proceedings*, AH '17, New York, NY, USA. Association for Computing Machinery.

Li, F. M., Chen, D. L., Fan, M., and Truong, K. N. (2019). FMT: A Wearable Camera-Based Object Tracking Memory Aid for Older Adults. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 3(3).

Li, Y., Shi, D., Ding, B., and Liu, D. (2014). Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors. In *Mining Intelligence and Knowledge Exploration*, volume 8891 of *Lecture Notes in Computer Science*, pages 99–107. Springer International Publishing, Cham.

Liang, F., Kevin, S., Baldauf, H., Kunze, K., and Pai, Y. S. (2020). OmniView: An Exploratory Study of 360 Degree Vision Using Dynamic Distortion Based on Direction of Interest. In *Augmented Humans (AHs), Conference Proceedings*, AHs '20, New York, NY, USA. Association for Computing Machinery.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the Speech Code. *Psychological Review*, 74(6):431–461.

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1):4–11.

Lilija, K., Pohl, H., Boring, S., and Hornbæk, K. (2019). Augmented Reality Views for Occluded Interaction. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Ling, J., Catling, J., and Upton, D. (2011). *Cognitive Psychology*. Psychology Express. Pearson Education Limited.

Lingenfelser, F., Wagner, J., and André, E. (2011). A Systematic Discussion of Fusion Techniques for Multi-modal Affect Recognition Tasks. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 19–26. ACM.

Lingenfelser, F., Wagner, J., Vogt, T., Kim, J., and André, E. (2010). Age and Gender Classification from Speech using Decision Level Fusion and Ensemble Based Techniques. In *Interspeech, Conference Proceedings*, pages 2798–2801.

Logie, R. H. (1995). *Visuo-Spatial Working Memory*. Essays in Cognitive Psychology. Lawrence Erlbaum, Hove.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 3431–3440. IEEE.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. (2012). StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 351–360. ACM.

Lu, H., Pan, W., Lane, N. D., Choudhury, T., and Campbell, A. T. (2009). SoundSense: Scalable Sound Sensing for People-centric Applications on Mobile Phones. In *Mobile Systems, Applications, and Services (MobiSys), Conference Proceedings*, pages 165–178. ACM.

Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., and Campbell, A. T. (2010). The Jigsaw Continuous Sensing Engine for Mobile Phone Applications. In *Embedded Networked Sensor Systems (SenSys), Conference Proceedings*, pages 71–84. ACM.

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's Only a Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior*, 37:94–100.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., and Lee, J. (2019). MediaPipe: A Framework for Perceiving and Processing Reality. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*. IEEE.

Luzhnica, G. and Veas, E. (2018). Skin Reading Meets Speech Recognition and Object Recognition for Sensory Substitution. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '18, page 146–149, New York, NY, USA. Association for Computing Machinery.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Computer Vision – ECCV 2018*, volume 11218 of *Lecture Notes in Computer Science*, pages 122–138. Springer International Publishing, Cham.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, 101(4):676–703.

Macpherson, F. (2011). Taxonomising the Senses. *Philosophical Studies*, 153(1):123–142.

Maier, N. R. F. (1931). Reasoning in Humans. II. The Solution of a Problem and its Appearance in Consciousness. *Journal of Comparative Psychology*, 12(2):181–194.

Malim, T. (1994). *Cognitive Processes: Attention, Perception, Memory, Thinking and Language*. Introductory Psychology. Macmillan, London.

Manduchi, R. and Coughlan, J. M. (2014). The Last Meter: Blind Visual Guidance to a Target. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '14, page 3113–3122, New York, NY, USA. Association for Computing Machinery.

Marieb, E. N. and Hoehn, K. (2019). *Human Anatomy & Physiology*. Pearson, Harlow, 11th edition.

Markova, V., Ganchev, T., and Kalinkov, K. (2019). CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In *Biomedical Innovations and Applications (BIA), Conference Proceedings*, pages 1–4. IEEE.

Mateevitsi, V., Haggadone, B., Leigh, J., Kunzer, B., and Kenyon, R. V. (2013). Sensing the Environment through SpiderSense. In *Augmented Human (AH), Conference Proceedings*, AH '13, page 51–57, New York, NY, USA. Association for Computing Machinery.

Matthews, T., Carter, S., Pai, C., Fong, J., and Mankoff, J. (2006). Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Deaf. In Dourish, P. and Friday, A., editors, *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 159–176, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mayer, R. E. (2013). Problem Solving. In *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

McBride, D. M. and Cutting, J. C. (2019). *Cognitive Psychology: Theory, Process, and Methodology*. SAGE Publications Inc, Thousand Oaks California, 2nd edition.

McCambridge, J., Witton, J., and Elbourne, D. R. (2014). Systematic Review of the Hawthorne Effect: New Concepts are Needed to Study Research Participation Effects. *Journal of clinical epidemiology*, 67(3):267–277.

McClelland, J. L. and Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, 18(1):1–86.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago, IL, US.

Mehlman, M. J. and Botkin, J. R. (1998). *Access to the Genome: The Challenge to Equality*. Georgetown University Press.

Mehrabian, A. (1995). Framework for a Comprehensive Description and Measurement of Emotional states. *Genetic, Social, and General Psychology Monographs*, 121(3):339–361.

Mehrabian, A. (1996). Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament. *Current Psychology*, 14(4):261–292.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing information. *Psychological Review*, 63(2):81–97.

Mills, M., Hollingworth, A., van der Stigchel, S., Hoffman, L., and Dodd, M. D. (2011). Examining the Influence of Task Set on Eye Movements and Fixations. *Journal of Vision*, 11(8):17.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., and Campbell, A. T. (2008). Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In *Embedded Network Sensor Systems (SenSys), Conference Proceedings*, pages 337–350. ACM.

Min Htike, H., H. Margrain, T., Lai, Y.-K., and Eslambolchilar, P. (2021). Augmented Reality Glasses as an Orientation and Mobility Aid for People with Low Vision: A Feasibility Study of Experiences and Requirements. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Mirman, D., Holt, L. L., and McClelland, J. L. (2004). Categorization and Discrimination of Nonspeech Sounds: Differences between Steady-State and Rapidly-Changing Acoustic Cues. *The Journal of the Acoustical Society of America*, 116(2):1198–1207.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222.

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60.

Mumuni, A. and Mumuni, F. (2022). Data Augmentation: A Comprehensive Survey of Modern Approaches. *Array*, 16:100258.

Nanayakkara, S., Shilkrot, R., Yeo, K. P., and Maes, P. (2013). EyeRing: A Finger-Worn Input Device for Seamless Interactions with Our Surroundings. In *Augmented Human (AH), Conference Proceedings*, AH '13, page 13–20, New York, NY, USA. Association for Computing Machinery.

Nanayakkara, S., Taylor, E., Wyse, L., and Ong, S. H. (2009). An Enhanced Musical Experience for the Deaf: Design and Evaluation of a Music Display and a Haptic Chair. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '09, page 337–346, New York, NY, USA. Association for Computing Machinery.

Navon, D. and Gopher, D. (1979). On the Economy of the Human-Processing System. *Psychological Review*, 86(3):214–255.

Neisser, U. (1967). *Cognitive Psychology*. Century Psychology Series. Prentice Hall, Englewood Cliffs, N.J.

Neumann, E. and Blanton, R. (1970). The Early History of Electrodermal Research. *Psychophysiology*, 6(4):453–475.

Newell, A. and Simon, H. A. (1961). GPS, A Program that Simulates Human Thought. *Lernende Automaten*.

Newell, A. and Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition.

Niforatos, E. (2018). *The Role of Context in Human Memory Augmentation*. Dissertation, Università della Svizzera italiana.

Niforatos, E., Cinel, C., Mack, C. C., Langheinrich, M., and Ward, G. (2017). Can Less be More?: Contrasting Limited, Unlimited, and Automatic Picture Capture for Augmenting Memory Recall. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 1(2):1–22.

Niforatos, E., Laporte, M., Bexheti, A., and Langheinrich, M. (2018). Augmenting Memory Recall in Work Meetings. In *Augmented Human (AH), Conference Proceedings*, pages 1–7. ACM.

Nilsson, M., Gertsovich, I., and Bartunek, J. S. (2010). Mouth Open or Closed Decision for Frontal Face Images with Given Eye Locations. In *Biometrics: Theory, Applications and Systems (BTAS), Conference Proceedings*, pages 1–6. IEEE.

Nirjon, S., Dickerson, R. F., Asare, P., Li, Q., Hong, D., Stankovic, J. A., Hu, P., Shen, G., and Jiang, X. (2013). Auditeur: A Mobile-Cloud Service Platform for Acoustic Event Detection on Smartphones. In *Mobile Systems, Applications, and Services (MobiSys), Conference Proceedings*, pages 403–416. ACM.

Norooz, L., Mauriello, M. L., Jorgensen, A., McNally, B., and Froehlich, J. E. (2015). BodyVis: A New Approach to Body Learning Through Wearable Sensing and Visualization. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '15, page 1025–1034, New York, NY, USA. Association for Computing Machinery.

Olwal, A., Balke, K., Votintcev, D., Starner, T., Conn, P., Chinh, B., and Corda, B. (2020). Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *User Interface Software and Technology (UIST), Conference Proceedings*, pages 1108–1120. Association for Computing Machinery.

Osmani, V., Zhang, D., and Balasubramaniam, S. (2009). Human Activity Recognition Supporting Context-appropriate Reminders for Elderly. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 1–4.

Otero-González, I., Pacheco-Lorenzo, M. R., Fernández-Iglesias, M. J., and Anido-Rifón, L. E. (2024). Conversational Agents for Depression Screening: A Systematic Review. *International journal of medical informatics*, 181:105272.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Ouwerkerk, M., Pasveer, F., and Langereis, G. (2008). Unobtrusive Sensing of Psychophysiological Parameters. In *Probing Experience*, volume 8 of *Philips Research*, pages 163–193. Springer Netherlands, Dordrecht.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The PRISMA 2020 Statement:

An Updated Guideline for Reporting Systematic Reviews. *BMJ (Clinical Research Edition)*, 372:n71.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Panëels, S. A., Olmos, A., Blum, J. R., and Cooperstock, J. R. (2013). Listen to It Yourself! Evaluating Usability of What's around Me? For the Blind. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '13, page 2107–2116, New York, NY, USA. Association for Computing Machinery.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., and Talwar, K. (2016). Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data.

Parkhi, O., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. In *British Machine Vision Conference (BMVC), Conference Proceedings*. British Machine Vision Association.

Partala, T. and Surakka, V. (2003). Pupil Size Variation as an Indication of Affective Processing. *International Journal of Human-Computer Studies*, 59(1-2):185–198.

Partridge, L. D. and Partridge, L. D. (2003). Muscle Activity. In *Nervous System Actions and Interactions*, pages 289–309. Springer US, Boston, MA.

Pavlov, P. I. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, London.

Pease, A. and Pease, B. (2008). *The Definitive Book of Body Language*. Bantam Books, New York, NY, 1st edition.

Peng, H., Hu, B., Liu, Q., Dong, Q., Zhao, Q., and Moore, P. (2011). User-centered Depression Prevention: An EEG Approach to Pervasive Healthcare. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 325–330.

Peng, Y.-H., Hsi, M.-W., Taele, P., Lin, T.-Y., Lai, P.-E., Hsu, L., Chen, T.-c., Wu, T.-Y., Chen, Y.-A., Tang, H.-H., and Chen, M. Y. (2018). SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '18, page 1–10, New York, NY, USA. Association for Computing Machinery.

Pentland, A. (2008). *Honest Signals: How They Shape Our World*. The MIT Press.

Pérez Fornos, A., van de Berg, R., Sommerhalder, J., and Guinand, N. (2019). Designing Artificial Senses: Steps from Physiology to Clinical Implementation. *Swiss Medical Weekly*, 149:w20061.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild Cognitive Impairment: Clinical Characterization and Outcome. *Archives of Neurology*, 56(3):303–308.

Petry, B., Illandara, T., Elvitigala, D. S., and Nanayakkara, S. (2018). Supporting Rhythm Activities of Deaf Children Using Music-Sensory-Substitution Systems. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '18, page 1–10, New York, NY, USA. Association for Computing Machinery.

Pfleging, B., Alt, F., and Schmidt, A. (2012). Meaningful Melodies: Personal Sonification of Text Messages for Mobile Devices. In *MobileHCI, Conference Proceedings*, pages 189–192. ACM.

Philip, P., Micoulaud-Franchi, J.-A., Sagaspe, P., de Sevin, E., Olive, J., Bioulac, S., and Sauteraud, A. (2017). Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports*, 7:42656.

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

Pielot, M., Dingler, T., Pedro, J. S., and Oliver, N. (2015). When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '15, page 825–836, New York, NY, USA. Association for Computing Machinery.

Pina, L., Rowan, K., Roseway, A., Johns, P., Hayes, G. R., and Czerwinski, M. (2014). In Situ Cues for ADHD Parenting Strategies Using Mobile Technology. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, PervasiveHealth '14, page 17–24, Brussels, BEL. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., and Wittmers, L. E. (2011). Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment. In *Information Processing in Sensor Networks (IPSN), Conference Proceedings*, pages 97–108.

Poggi, I. and Francesca, D. (2010). Cognitive Modelling of Human Social Signals. In *Social Signal Processing Workshop (SSPW), Conference Proceedings*, pages 21–26. ACM.

Power, M. J. and Dalgleish, T. (2016). *Cognition and Emotion: From Order to Disorder*. Psychology Press, London, 3rd edition.

Poyatos, F. (1984). The Multichannel Reality of Discourse: Language-Paralanguage-Kinesics and the Totality of Communicative Systems. *Language Sciences*, 6(2):307–337.

Quintana, D. S., Guastella, A. J., Outhred, T., Hickie, I. B., and Kemp, A. H. (2012). Heart Rate Variability is Associated with Emotion Recognition: Direct Evidence for a Relationship between the Autonomic Nervous System and Social Cognition. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, 86(2):168–172.

Rachuri, K. K., Mascolo, C., Musolesi, M., and Rentfrow, P. J. (2011). SociableSense: Exploring the Trade-offs of Adaptive Sampling and Computation Offloading for Social Sensing. In *Mobile Computing and Networking (MobiCom), Conference Proceedings*, pages 73–84. ACM.

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. (2010). EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 281–290. ACM.

Radüntz, T. (2018). Signal Quality Evaluation of Emerging EEG Devices. *Frontiers in Physiology*, 9:98.

Ragni, M. and Stolzenburg, F. (2015). Higher-Level Cognition and Computation: A Survey. *KI - Künstliche Intelligenz*, 29(3):247–253.

Rahman, S. A., Merck, C., Huang, Y., and Kleinberg, S. (2015). Unintrusive Eating Recognition Using Google Glass. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, pages 108–111. ICST.

Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., and Farooq, A. (2019). Human Augmentation: Past, Present and Future. *International Journal of Human-Computer Studies*, 131:131–143.

Ravi, N., Dandekar, N., Mysore, P., and Littman, M. L. (2005). Activity Recognition from Accelerometer Data. In *Innovative Applications of Artificial Intelligence (IAAI), Conference Proceedings*, pages 1541–1546. AAAI Press.

Reeves, L. M. and Schmorrow, D. D. (2007). Augmented Cognition Foundations and Future Directions—Enabling "Anyone, Anytime, Anywhere" Applications. In *Coping with Diversity*, volume 4554 of *Lecture Notes in Computer Science*, pages 263–272. Springer, Berlin.

Reis, D., Xanthopoulou, D., and Tsaousis, I. (2015). Measuring Job and Academic Burnout with the Oldenburg Burnout Inventory (OLBI): Factorial Invariance Across Samples and Countries. *Burnout Research*, 2(1):8–18.

Richmond, V. P. and MacCroskey, J. C. (1995). *Nonverbal Behavior in Interpersonal Relations*. Allyn and Bacon, Bonston, 3rd edition.

Ring, L., Bickmore, T., and Pedrelli, P. (2016). An Affectively Aware Virtual Therapist for Depression Counseling. In *Human Factors in Computing Systems (CHI), Workshop on Computing and Mental Health*, pages 01951–12.

Robbins, T. W. (2005). Chemistry of the Mind: Neurochemical Modulation of Prefrontal Cortical Function. *Journal of Comparative Neurology*, 493(1):140–146.

Roediger, H. L. and Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, 17(3):249–255.

Roseman, I. J. and Smith, C. A. (2001). Appraisal Theory Overview, Assumptions, Varieties, Controversies. In *Appraisal Processes in Emotion*, pages 3–19. Oxford University Press, New York, NY.

Rubin, J., Eldardiry, H., Abreu, R., Ahern, S., Du, H., Pattekar, A., and Bobrow, D. G. (2015). Towards a Mobile and Wearable System for Predicting Panic Attacks. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '15, page 529–533, New York, NY, USA. Association for Computing Machinery.

Rudovic, O., Tobis, N., Kaltwang, S., Schuller, B., Rueckert, D., Cohn, J. F., and Picard, R. W. (2021). Personalized Federated Deep Learning for Pain Estimation From Face Images.

Ruf, T., Ernst, A., and Küblbeck, C. (2011). Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE). In *Microelectronic Systems*, pages 243–252. Springer Berlin Heidelberg, Berlin, Heidelberg.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Sabour, R. M., Benezeth, Y., de Oliveira, P., Chappé, J., and Yang, F. (2023). UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress. *IEEE Transactions on Affective Computing*, 14(1):622–636.

Sadasivan, S., Greenstein, J. S., Gramopadhye, A. K., and Duchowski, A. T. (2005). Use of Eye Movements As Feedforward Training for a Synthetic Aircraft Inspection Task. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 141–149. ACM.

Saganowski, S., Komoszyńska, J., Behnke, M., Perz, B., Kunc, D., Klich, B., Kaczmarek, Ł. D., and Kazienko, P. (2022). Emognition Dataset: Emotion Recognition with Self-Reports, Facial Expressions, and Physiology using Wearables. *Scientific Data*, 9(1):158.

Saganowski, S., Perz, B., Polak, A. G., and Kazienko, P. (2023). Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review. *IEEE Transactions on Affective Computing*, 14(3):1876–1897.

Sakurada, M. and Yairi, T. (2014). Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Machine Learning for Sensory Data Analysis (MLSDA), Conference Proceedings*, pages 4–11. ACM.

Salari, N., Hosseinian-Far, A., Jalali, R., Vaisi-Raygani, A., Rasoulpoor, S., Mohammadi, M., Rasoulpoor, S., and Khaledi-Paveh, B. (2020). Prevalence of Stress, Anxiety, Depression Among the General Population During the COVID-19 Pandemic: A Systematic Review and Meta-Analysis. *Globalization and Health*, 16(1):57.

Salthouse, T. A. and Babcock, R. L. (1991). Decomposing Adult Age Differences in Working Memory. *Developmental Psychology*, 27(5):763.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 4510–4520. IEEE.

Savulescu, J. and Bostrom, N., editors (2009). *Human Enhancement*. Oxford University Press, Oxford, 1st edition.

Scheflen, A. E. (1964). The Significance of Posture in Communication Systems. *Psychiatry*, 27:316–331.

Scherer, K. (2003). Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, 40(1-2):227–256.

Scherer, M. J., Hart, T., Kirsch, N., and Schulthesis, M. (2005). Assistive Technologies for Cognitive Disabilities. *Critical Reviews in Physical and Rehabilitation Medicine*, 17(3):195–215.

Schiller, D., Huber, T., Dietz, M., and André, E. (2020). Relevance-Based Data Masking: A Model-Agnostic Transfer Learning Approach for Facial Expression Recognition. *Frontiers in Computer Science*, 2(6).

Schmidt, A. (2017). Technologies to Amplify the Mind. *Computer*, 50(10):102–106.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and van Laerhoven, K. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 400–408. ACM.

Schmidt, P., Reiss, A., Dürichen, R., and van Laerhoven, K. (2019). Wearable-Based Affect Recognition-A Review. *Sensors (Basel, Switzerland)*, 19(19).

Schmorrow, D. and Kruse, A. (2004). Augmented Cognition. In *Berkshire Encyclopedia of Human-Computer Interaction*, pages 54–59. Berkshire Publ. Group, Great Barrington, Mass.

Schmorrow, D., Stanney, K. M., Wilson, G., and Young, P. (2006). Augmented Cognition in Human-System Interaction. In *Handbook of Human Factors and Ergonomics*, pages 1364–1383. Wiley, Hoboken, NJ.

Schneegass, S. (2016). *Enriching Mobile Interaction with Garment-Based Wearable Computing Devices*. Dissertation, Universität Stuttgart, Stuttgart.

Seiderer, A., Dietz, M., Aslan, I., and André, E. (2018). Enabling Privacy with Transfer Learning for Image Classification DNNs on Mobile Devices. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs), Conference Proceedings*, pages 25–30. ACM.

Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C., and Wood, K. (2007). Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using Sensecam. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '07, page 81–90, New York, NY, USA. Association for Computing Machinery.

Shaffer, F. and Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, 5:258.

Shaffer, F., McCraty, R., and Zerr, C. L. (2014). A Healthy Heart is not a Metronome: an Integrative Review of the Heart's Anatomy and Heart Rate variability. *Frontiers in Psychology*, 5:1040.

Shangguan, L., Yang, Z., Zhou, Z., Zheng, X., Wu, C., and Liu, Y. (2014). CrossNavi: Enabling Real-Time Crossroad Navigation for the Blind with Commodity Phones. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '14, page 787–798, New York, NY, USA. Association for Computing Machinery.

Shannon, C. E. (1949). Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21.

Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J. G., Vhaduri, S., Preston, K., and Kumar, S. (2015). Visualization of Time-Series Sensor Data to Inform the Design of Just-in-Time Adaptive Stress Interventions. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '15, page 505–516, New York, NY, USA. Association for Computing Machinery.

Sherwood, L. (2015). *Human Physiology: From Cells to Systems*. Cengage Learning, Boston, MA, USA, 9th edition.

Shilkrot, R., Huber, J., Meng Ee, W., Maes, P., and Nanayakkara, S. C. (2015). FingerReader: A Wearable Device to Explore Printed Text on the Go. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '15, page 2363–2372, New York, NY, USA. Association for Computing Machinery.

Shrout, P. E. and Fiske, D. W. (1981). Nonverbal Behaviors and Social Evaluation. *Journal of Personality*, 49(2):115–128.

Shui, X., Zhang, M., Li, Z., Hu, X., Wang, F., and Zhang, D. (2021). A Dataset of Daily Ambulatory Psychological and Physiological Recording for Emotion Research. *Scientific Data*, 8(1):161.

Sicong, L., Zimu, Z., Junzhao, D., Longfei, S., Han, J., and Wang, X. (2017). UbiEar: Bringing Location-Independent Sound Awareness to the Hard-of-Hearing People with Smartphones. *Interactive, Mobile, Wearable and Ubiquitous Technologies, Journal Proceedings*, 1(2).

Sinex, J. E. (1999). Pulse Oximetry: Principles and Limitations. *The American journal of emergency medicine*, 17(1):59–67.

Six, J., Cornelis, O., and Leman, M. (2014). TarsosDSP, a Real-Time Audio Processing Framework in Java. In *Audio Engineering Society (AES), Conference Proceedings*. Audio Engineering Society.

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, Oxford, England.

Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts, East Norwalk.

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., de Raedt, W., Cornelis, J., Janssens, O., van Hoecke, S., Claes, S., van Diest, I., and van Hoof, C. (2018). Large-Scale Wearable Data Reveal Digital Phenotypes for Daily-Life Stress Detection. *NPJ digital medicine*, 1:67.

Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. (2005). Early Versus Late Fusion in Semantic Video Analysis. In *ACM Multimedia (MM), Conference Proceedings*, pages 399–402. ACM.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Lapata, M. and Ng, H. T., editors, *Empirical Methods in Natural Language Processing (EMNLP), Conference Proceedings*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Sohn, T., Li, K. A., Lee, G., Smith, I., Scott, J., and Griswold, W. G. (2005). Place-Its: A Study of Location-Based Reminders on Mobile Phones. In Beigl, M., Intille, S., Rekimoto, J., and Tokuda, H., editors, *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 232–250, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sperling, G. (1960). The Information Available in Brief Visual Presentations. *Psychological Monographs: General and Applied*, 74(11):1–29.

Spina, G., Roberts, F., Weppner, J., Lukowicz, P., and Amft, O. (2013). CRNTC+: A Smartphone-based Sensor Processing Framework for Prototyping Personal Healthcare Applications. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*. IEEE.

Squire, L. R. (1992). Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory. *Journal of Cognitive Neuroscience*, 4(3):232–243.

Sridhar, S., Khamaj, A., and Asthana, M. K. (2023). Cognitive Neuroscience Perspective on Memory: Overview and Summary. *Frontiers in Human Neuroscience*, 17:1217093.

Stanney, K., Hale, K., and Jones, D. (2009a). Augmented Cognition Design Approaches for Treating Mild Traumatic Brain Injuries. In *Foundations of Augmented Cognition (FAC), Conference Proceedings*, volume 5638 of *Lecture notes in computer science Lecture notes in artificial intelligence*, pages 800–809. Springer, Berlin and Heidelberg.

Stanney, K. M., Schmorrow, D. D., Johnston, M., Fuchs, S., Jones, D., Hale, K. S., Ahmad, A., and Young, P. (2009b). Augmented Cognition: An Overview. *Reviews of Human Factors and Ergonomics*, 5(1):195–224.

Stanovich, K. E. (1999). *Who Is Rational?* Psychology Press.

Stavemann, H. H. (2015). *Sokratische Gesprächsführung in Therapie und Beratung: Eine Anleitung für Psychotherapeuten, Berater und Seelsorger*. Beltz, Weinheim and Basel, 3rd edition.

Stefanov, K., Huang, B., Li, Z., and Soleymani, M. (2020). OpenSense: A Platform for Multimodal Data Acquisition and Behavior Perception. In *International Conference on Multimodal Interaction (ICMI), Conference Proceedings*, pages 660–664. ACM.

Sternberg, R. J. (2019). *The Psychology of Human Thought: An Introduction*. Heidelberg University Publishing (heiUP), Heidelberg.

Sternberg, R. J. and Sternberg, K. (2012). *Cognitive Psychology*. Wadsworth, Belmont, Calif., 6th edition.

Suthana, N., Haneef, Z., Stern, J., Mukamel, R., Behnke, E., Knowlton, B., and Fried, I. (2012). Memory Enhancement and Deep-Brain Stimulation of the Entorhinal Area. *New England Journal of Medicine*, 366(6):502–510. PMID: 22316444.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Svanberg, J. and Evans, J. J. (2013). Neuropsychological Rehabilitation in Alcohol-related Brain Damage: A Systematic Review. *Alcohol and Alcoholism (Oxford, Oxfordshire)*, 48(6):704–711.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 2818–2826. IEEE.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. (2019). MnasNet: Platform-Aware Neural Architecture Search for Mobile. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 2815–2823. IEEE.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. (2021). EfficientNetV2: Smaller Models and Faster Training. In *International Conference on Machine Learning (ICML), Conference Proceedings*, pages 10096–10106. PMLR.

Tang, D., Qin, B., and Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Empirical Methods in Natural Language Processing (EMNLP), Conference Proceedings*, pages 1422–1432. Association for Computational Linguistics.

Tang, T. J. J. and Li, W. H. (2014). An Assistive EyeWear Prototype That Interactively Converts 3D Object Locations into Spatial Audio. In *International Symposium on Wearable Computers (ISWC), Conference Proceedings*, ISWC '14, page 119–126, New York, NY, USA. Association for Computing Machinery.

Tanuwidjaja, E., Huynh, D., Koa, K., Nguyen, C., Shao, C., Torbett, P., Emmenegger, C., and Weibel, N. (2014). Chroma: A Wearable Augmented-Reality Solution for Color Blindness. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '14, page 799–810, New York, NY, USA. Association for Computing Machinery.

Thorndike, E. L. (1898). Animal Intelligence: An Experimental Study of the Associative Processes in Animals. *Psychological Monographs: General and Applied*, 2(4):i–109.

Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., and Pantic, M. (2021). Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions. *Nature Machine Intelligence*, 3(1):42–50.

Tomasi, C. and Manduchi, R. (1998). Bilateral Filtering for Gray and Color Images. In *International Conference on Computer Vision (ICCV), Conference Proceedings*, pages 839–846. Narosa Publishing House.

Torralba, A. and Efros, A. A. (2011). Unbiased Look at Dataset Bias. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 1521–1528. IEEE.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features in Object Search. *Psychological Review*, 113(4):766–786.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models.

Treisman, A. M. (1960). Contextual Cues in Selective Listening. *Quarterly Journal of Experimental Psychology*, 12(4):242–248.

Treisman, A. M. (1964). Verbal Cues, Language, and Meaning in Selective Attention. *The American Journal of Psychology*, 77(2):206.

Treisman, A. M. and Davies, A. (1973). Divided Attention to Ear and Eye. *Attention and Performance IV*, pages 101–117.

Tulving, E. (1972). Episodic and Semantic Memory. In Tulving, E. and Donaldson, W., editors, *Organization of Memory*, pages 381–403. Academic Press, New York.

Twardon, L., Koesling, H., Finke, A., and Ritter, H. (2013). Gaze-Contingent Audio-Visual Substitution for the Blind and Visually Impaired. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*. IEEE.

Uppal, S., Bajaj, Y., Rustom, I., and Coatesworth, A. P. (2009). Otosclerosis 1: The Aetiopathogenesis of Otosclerosis. *International Journal of Clinical Practice*, 63(10):1526–1530.

Utsumi, Y., Kato, Y., Kunze, K., Iwamura, M., and Kise, K. (2013). Who Are You? A Wearable Face Recognition System to Support Human Memory. In *Augmented Human (AH), Conference Proceedings*, AH '13, page 150–153, New York, NY, USA. Association for Computing Machinery.

Vadiraja, P., Dengel, A., and Ishimaru, S. (2021). Text Summary Augmentation for Intelligent Reading Assistant. In *Augmented Humans (AHs), Conference Proceedings*, AHs '21, page 319–321, New York, NY, USA. Association for Computing Machinery.

van Berkel, N., Ferreira, D., and Kostakos, V. (2018). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, 50(6):1–40.

Van der Stigchel, S., Meeter, M., and Theeuwes, J. (2006). Eye Movement Trajectories and What They Tell Us. *Neuroscience and Biobehavioral Reviews*, 30(5):666–679.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Verghese, P. (2001). Visual Search and Attention: A Signal Detection Theory Approach. *Neuron*, 31(4):523–535.

Vigouroux, R. (1879). Sur le role de la resistance electrique des tissues dans l'electrodiagnostic. *Comptes Rendus Societe de Biologie*, 31:336–339.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. In *International Conference on Machine learning (ICML), Conference Proceedings*, pages 1096–1103. ACM Press.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing*, 27(12):1743–1759.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator.

Vogt, T., André, E., and Bee, N. (2008). EmoVoice — A Framework for Online Recognition of Emotions from Voice. In *Perception in Multimodal Dialogue Systems*, volume 5078 of *Lecture Notes in Computer Science*, pages 188–199. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wagner, J., Kim, J., and Andre, E. (2005). From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In *International Conference on Multimedia and Expo (ICME), Conference Proceedings*, pages 940–943. IEEE.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., and André, E. (2013). The Social Signal Interpretation (SSI) Framework. In *ACM Multimedia (MM), Conference Proceedings*, pages 831–834. ACM.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. (2023). Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Wang, X., Zhao, X., Prakash, V., Shi, W., and Gnawali, O. (2013). Computerized-Eyewear Based Face Recognition System for Improving Social Lives of Prosopagnosics. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Conference Proceedings*, PervasiveHealth '13, page 77–80, Brussels, BEL. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Wang, Y., Lin, J., Annavaram, M., Jacobson, Q. A., Hong, J., Krishnamachari, B., and Sadeh, N. (2009). A Framework of Energy Efficient Mobile Sensing for Automatic User State Recognition. In *Mobile Systems, Applications, and Services (MobiSys), Conference Proceedings*, pages 179–192. ACM.

Wason, P. C. and Evans, J. (1974). Dual Processes in Reasoning? *Cognition*, 3(2):141–154.

Watanabe, H. and Terada, T. (2020). Manipulatable Auditory Perception in Wearable Computing. In *Augmented Humans (AHs), Conference Proceedings*, AHs '20, New York, NY, USA. Association for Computing Machinery.

Webb, A. R. (2002). *Statistical Pattern Recognition*. John Wiley & Sons.

Weiler, D. T., Villajuan, S. O., Edkins, L., Cleary, S., and Saleem, J. J. (2017). Wearable Heart Rate Monitor Technology Accuracy in Research: A Comparative Study Between PPG and ECG Technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):1292–1296.

Westin, A. F. (1967). *Privacy and Freedom*. Atheneum, New York.

Weyerer, S. and Bickel, H. (2007). *Epidemiologie psychischer Erkrankungen im höheren Lebensalter*. Kohlhammer.

Wheeler, L. and Reis, H. T. (1991). Self–Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality*, 59(3):339–354.

Wickens, C. D. (1980). The Structure of Attentional Resources. In *Attention and Performance VIII*, pages 239–257. Psychology Press.

Wickens, C. D. (1984). Processing Resources in Attention. In *Varieties of Attention*, pages 63–101. Academic Press.

Wickens, C. D. (2002). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177.

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, 50(3):449–455.

Wiegand, H. F., Saam, J., Marschall, U., Chmitorz, A., Kriston, L., Berger, M., Lieb, K., and Hölzel, L. P. (2020). Challenges in the Transition from In-Patient to Out-Patient Treatment in Depression. *Deutsches Ärzteblatt International*, 117(27-28):472–479.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In *Language Resources and Evaluation (LREC), Conference Proceedings*, pages 1556–1559.

World Health Organization (2017). Depression and Other Common Mental Disorders. Technical Report WHO/MSD/MER/2017.2, World Health Organization.

World Health Organization (2019). World Report on Vision. Technical report, World Health Organization, Geneva.

Woźniak, P., Knaving, K., Obaid, M., Carcedo, M. G., Ünlüer, A., and Fjeld, M. (2015). ChromaGlove: A Wearable Haptic Feedback Device for Colour Recognition. In *Augmented Human (AH), Conference Proceedings*, AH '15, page 219–220, New York, NY, USA. Association for Computing Machinery.

Wu, B., Keutzer, K., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., and Jia, Y. (2019). FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 10726–10734. IEEE.

Xia, C. and Maes, P. (2013). The Design of Artifacts for Augmenting Intellect. In *Augmented Human (AH), Conference Proceedings*, pages 154–161. ACM Press.

Yadav, M., Sakib, M. N., Nirjhar, E. H., Feng, K., Behzadan, A. H., and Chaspari, T. (2022). Exploring Individual Differences of Public Speaking Anxiety in Real-Life and Virtual Presentations. *IEEE Transactions on Affective Computing*, 13(3):1168–1182.

Yamano, S., Hamajo, T., Takahashi, S., and Higuchi, K. (2012). EyeSound: Single-Modal Mobile Navigation Using Directionally Annotated Music. In *Augmented Human (AH), Conference Proceedings*, AH '12, New York, NY, USA. Association for Computing Machinery.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer US, Boston, MA.

Yatani, K., Banovic, N., and Truong, K. (2012). SpaceSense: Representing Geographical Information to Visually Impaired People Using Spatial Tactile Feedback. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '12, page 415–424, New York, NY, USA. Association for Computing Machinery.

Yatani, K. and Truong, K. N. (2012). BodyScope: A Wearable Acoustic Sensor for Activity Recognition. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 341–350. ACM.

Yoshida, T., Kitani, K. M., Koike, H., Belongie, S., and Schlei, K. (2011). Edgesonic: Image Feature Sonification for the Visually Impaired. In *Augmented Human (AH), Conference Proceedings*, pages 1–4. ACM.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research*.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Association for Computational Linguistics (ACL), Conference Proceedings*, pages 2236–2246. Association for Computational Linguistics.

Zhang, Q., Yang, L. T., and Chen, Z. (2016). Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning. *IEEE Transactions on Computers*, 65(5):1351–1362.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Computer Vision and Pattern Recognition (CVPR), Conference Proceedings*, pages 6848–6856. IEEE.

Zhao, Y., Szpiro, S., Knighten, J., and Azenkot, S. (2016). CueSee: Exploring Visual Cues for People with Low Vision to Facilitate a Visual Search Task. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, UbiComp '16, page 73–84, New York, NY, USA. Association for Computing Machinery.

Zhao, Y., Wu, S., Reynolds, L., and Azenkot, S. (2018). A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision (ICCV), Conference Proceedings*, pages 2242–2251. IEEE.

Appendix

A Component Options

In order to define publicly accessible options, we create an inner class that extends the abstract `OptionList` class within the respective component (line 8). Afterwards, we define the configurable parameters as member variables with the generic `Option` type. It expects a display name, default value, primitive base class, and description, as shown in lines 11-15. The reason why we encapsulate each parameter within an `Option` instance is that it still allows developers to quickly access all available settings (by typing the component's name followed by a dot, which brings up all member variables in most IDEs) while also maintaining a uniform structure required for automatically processing each option. For that, we call the `addOptions()` method from the private constructor of the inner class (line 20), which uses Java reflection to read the properties and values of all defined parameters (see Section 6.4.1). Finally, we instantiate the option list (line 25) and pass it to the overwritten `getOptions()` component method (line 31).

```
5  [...]
6
7  // Create inner class that extends the abstract OptionList class
8  public class Options extends OptionList
9  {
10     // Define component options
11     public final Option<String> firstOption = new Option<>(
12     "Option name", "default value", String.class, "Description");
13
14     public final Option<Integer> secondOption = new Option<>(
15     "Option name", 42, Integer.class, "Description");
16
17     private Options()
18     {
19         // Adds options automatically with Java reflection
20         addOptions();
```

```
21      }
22  }
23
24  // Create instance of inner class as a field
25  public final Options options = new Options();
26
27  @Override
28  public OptionList getOptions()
29  {
30      // Return option list reference from field
31      return options;
32  }
33
34  [...]
```

B Smartphone Pipeline for Visual Search Support

This part handles the Bluetooth connection to Google Glass:

```
// Create bluetooth reader as server to read data from Google Glass
BluetoothReader btReader = new BluetoothReader();
btReader.options.connectionType.set(BluetoothConnection.Type.SERVER);
btReader.options.connectionName.set("ssj_stream");

// Create channel to read acceleration from Google Glass
BluetoothChannel glassAcc = new BluetoothChannel();
glassAcc.options.channel_id.set(0);
glassAcc.options.dim.set(3);
glassAcc.options.type.set(Cons.Type.FLOAT);
glassAcc.options.sr.set(40);
glassAcc.options.num.set(40);

// Create channel to read gyroscope from Google Glass
BluetoothChannel glassGyr = new BluetoothChannel();
glassGyr.options.channel_id.set(1);
glassGyr.options.dim.set(3);
glassGyr.options.type.set(Cons.Type.FLOAT);
glassGyr.options.sr.set(40);
glassGyr.options.num.set(40);

// Create channel to read camera image from Google Glass
BluetoothChannel glassImage = new BluetoothChannel();
glassImage.options.channel_id.set(2);
glassImage.options.dim.set((int) (320*240*1.5));
glassImage.options.type.set(Cons.Type.IMAGE);
glassImage.options.sr.set(1);
glassImage.options.num.set(1);
glassImage.options.bytes.set(1);
glassImage.options.imageWidth.set(320);
glassImage.options.imageHeight.set(240);

// Add components to pipeline
pipeline.addSensor(btReader, glassAcc);
pipeline.addSensor(btReader, glassGyr);
pipeline.addSensor(btReader, glassImage);
```

This part handles visual search detection:

```
float frameSize = 1;
float deltaSize = 3;

// Create transformers to calculate head movement features
AccelerationFeatures accFeatures = new AccelerationFeatures();
```

```
AccelerationFeatures gyrFeatures = new AccelerationFeatures();

// Add components to pipeline
pipeline.addTransformer(accFeatures, glassAcc, frameSize, deltaSize);
pipeline.addTransformer(gyrFeatures, glassGyr, frameSize, deltaSize);

// Create search model and classifier
SVM searchModel = new SVM();
searchModel.options.file.set(new FilePath("/model/search.trainer"));

ClassifierT searchClassifier = new ClassifierT();
searchClassifier.setModel(searchModel);

// Add component to pipeline and use features as input
pipeline.addTransformer(searchClassifier, new Provider[] {
    accFeatures, gyrFeatures
}, frameSize, 0);

// Create event sender and set threshold to 0.8
ThresholdEventSender resultSender = new ThresholdEventSender();
resultSender.options.thresin.set(new float[] {0.8f});

// Add component to pipeline and use classification result as input
pipeline.addConsumer(resultSender, searchClassifier);

// Add external event receiver to output event channel
EventChannel resultChannel = resultSender.getEventChannelOut();
resultChannel.addEventListener(SearchHandler.getInstance());
```

This part handles target object detection:

```
// Create transformer to convert encoding format from NV21 to RGB
NV21ToRGBDecoder nv21ToRGBDecoder = new NV21ToRGBDecoder();
pipeline.addTransformer(nv21ToRGBDecoder, glassImage);

// Create transformer to resize image
ImageResizer imageResizer = new ImageResizer();
imageResizer.options.size.set(224);
pipeline.addTransformer(imageResizer, nv21ToRGBDecoder);

// Create transformer to normalize image pixel values between -1 and 1
ImageNormalizer imageNormalizer = new ImageNormalizer();
pipeline.addTransformer(imageNormalizer, imageResizer);

// Create object detection model and classifier
TFLite objModel = new TFLite();
objModel.options.file.set(new FilePath("/model/obj_detection.trainer"));

```

```
87 Classifier imageClassifier = new Classifier();
88 imageClassifier.setModel(objModel);
89 EventChannel imageChannel = imageClassifier.getEventChannelOut();
90
91 // Create image writer triggered by events
92 ImageWriter imageWriter = new ImageWriter();
93 imageWriter.options.triggeredByEvent.set(true);
94
95 // Add components to pipeline
96 pipeline.addConsumer(imageClassifier, imageNormalizer);
97 pipeline.addConsumer(imageWriter, nv21ToRGBDecoder);
98 pipeline.registerEventListener(imageWriter, imageChannel);
```

C Smartphone Pipeline for Video and Audio Analysis

This part handles visual features:

```
// Create camera sensor with resolution of 640x480 pixels
CameraSensor cameraSensor = new CameraSensor();
cameraSensor.options.cameraType.set(Cons.CameraType.FRONT_CAMERA);
cameraSensor.options.width.set(640);
cameraSensor.options.height.set(480);

// Create camera channel with sample rate of 5 Hz
CameraChannel cameraChannel = new CameraChannel();
cameraChannel.options.sampleRate.set(5);

// Add components to pipeline
pipeline.addSensor(cameraSensor, cameraChannel);

// Create transformer to convert encoding format from NV21 to RGB
NV21ToRGBDecoder rgbDecoder = new NV21ToRGBDecoder();
pipeline.addTransformer(rgbDecoder, cameraChannel);

// Create transformer to extract facial region
FaceCrop faceCrop = new FaceCrop();
pipeline.addTransformer(faceCrop, rgbDecoder);

// Create transformer to normalize image pixel values between -1 and 1
ImageNormalizer imageNormalizer = new ImageNormalizer();
pipeline.addTransformer(imageNormalizer, faceCrop);

// Create TensorFlow Lite model and select model file
TFLite vaModel = new TFLite();
vaModel.options.file.set(new FilePath("/model/valence_arousal.trainer"));

// Create classifier and select model
ClassifierT emotionClassifier = new ClassifierT();
emotionClassifier.setModel(vaModel);

// Add components to pipeline
pipeline.addModel(vaModel);
pipeline.addTransformer(emotionClassifier, imageNormalizer);

// Create transformer to convert float values to XML event
FloatsEventSender fesEmotion = new FloatsEventSender();
fesEmotion.options.sender.set("face");
fesEmotion.options.event.set("emotion");
pipeline.addConsumer(fesEmotion, emotionClassifier);

// Create socket writer to send XML event to VSM
```

```
45 SocketEventWriter sewEmotion = new SocketEventWriter();
46 sewEmotion.options.ip.set("127.0.0.1");
47 sewEmotion.options.port.set(5000);
48 sewEmotion.options.sendAsMap.set(true);
49 sewEmotion.options.mapKeys.set("valence,arousal");
50 pipeline.registerEventListener(sewEmotion, fesEmotion);
51
52 // Create socket writer to send XML event to VSM
53 SocketEventWriter sewFace = new SocketEventWriter();
54 sewFace.options.ip.set("127.0.0.1");
55 sewFace.options.port.set(5000);
56 sewFace.options.sendAsMap.set(true);
57 sewFace.options.mapKeys.set("faceX,faceY");
58 pipeline.registerEventListener(sewFace, faceCrop);
59
60 // Create transformer to calculate facial landmarks
61 FaceLandmarks landmarkTransformer = new FaceLandmarks();
62 pipeline.addTransformer(landmarkTransformer, rgbDecoder);
63
64 // Create transformer to calculate landmark features
65 LandmarkFeatures landmarkFeatures = new LandmarkFeatures();
66 pipeline.addTransformer(landmarkFeatures, landmarkTransformer);
67
68 // Create transformer to convert float values to XML event
69 FloatsEventSender fesMocs = new FloatsEventSender();
70 fesMocs.options.sender.set("face");
71 fesMocs.options.event.set("mouth");
72 pipeline.addConsumer(fesMocs, landmarkFeatures);
73
74 // Create socket writer to send XML event to VSM
75 SocketEventWriter sewMocs = new SocketEventWriter();
76 sewMocs.options.ip.set(Constants.VSM_IP); // Receiver IP
77 sewMocs.options.port.set(5000);
78 sewMocs.options.sendAsMap.set(true);
79 sewMocs.options.mapKeys.set("mouthOpen");
80 pipeline.registerEventListener(sewMocs, fesMocs);
```

This part handles audio features:

```
81 // Create microphone sensor
82 Microphone microphone = new Microphone();
83
84 // Create audio channel with 16 kHz sample rate
85 AudioChannel audio = new AudioChannel();
86 audio.options.sampleRate.set(16000);
87
88 // Add components to pipeline
89 pipeline.addSensor(microphone, audio);
```

```
90
91  // Create transformers to calculate audio features
92  Pitch pitch = new Pitch();
93  Intensity intensity = new Intensity();
94  Energy energy = new Energy();
95  OpenSmileFeatures egemaps = new OpenSmileFeatures();
96  OpenSmileFeatures mfcc = new OpenSmileFeatures();
97
98  egemaps.options.configFile.set(new FilePath("/ssj/os_egemaps_23.conf"));
99  mfcc.options.configFile.set(new FilePath("/ssj/os_mfcc_39.conf"));
100
101 // Add components to pipeline
102 pipeline.addTransformer(pitch, audio);
103 pipeline.addTransformer(intensity, audio);
104 pipeline.addTransformer(energy, audio);
105 pipeline.addTransformer(egemaps, audio);
106 pipeline.addTransformer(mfcc, audio);
```

This part handles storing all data to the file system:

```
107 // Write streams to file
108 WavWriter wavWriter = new WavWriter();
109 wavWriter.options.filePath.set("/sdcard/data/");
110 wavWriter.options.fileName.set("audio.wav");
111 pipeline.addConsumer(wavWriter, audio);
112
113 FFMPEGWriter cameraWriter = new FFMPEGWriter();
114 cameraWriter.options.filePath.set("/sdcard/data/");
115 cameraWriter.options.fileName.set("video.mp4");
116 cameraWriter.options.bitRate.set(3000);
117 pipeline.addConsumer(cameraWriter, rgbDecoder);
118
119 FileWriter landmarkWriter = new FileWriter();
120 landmarkWriter.options.filePath.set("/sdcard/data/");
121 landmarkWriter.options.fileName.set("facial_landmarks_binary");
122 landmarkWriter.options.type.set(Cons.FileType.BINARY);
123 pipeline.addConsumer(landmarkWriter, landmarkTransformer, 1);
124
125 FileWriter emotionWriter = new FileWriter();
126 emotionWriter.options.filePath.set("/sdcard/data/");
127 emotionWriter.options.fileName.set("valence_arousal");
128 emotionWriter.options.type.set(Cons.FileType.ASCII);
129 pipeline.addConsumer(emotionWriter, emotionClassifier, 1);
130
131 FileWriter egemapsWriter = new FileWriter();
132 egemapsWriter.options.filePath.set("/sdcard/data/");
133 egemapsWriter.options.fileName.set("audio_egemaps_binary");
134 egemapsWriter.options.type.set(Cons.FileType.BINARY);
```

```
135 pipeline.addConsumer(egemapsWriter, egemaps, 1);
136
137 FileWriter mfccWriter = new FileWriter();
138 mfccWriter.options.filePath.set("/sdcard/data/");
139 mfccWriter.options.fileName.set("audio_mfcc_binary");
140 mfccWriter.options.type.set(Cons.FileType.BINARY);
141 pipeline.addConsumer(mfccWriter, mfcc, 1);
142
143 FileWriter audioFeatureWriter = new FileWriter();
144 audioFeatureWriter.options.filePath.set("/sdcard/data/");
145 audioFeatureWriter.options.fileName.set("audio_features");
146 audioFeatureWriter.options.type.set(Cons.FileType.ASCII);
147 pipeline.addConsumer(audioFeatureWriter, new Provider[] {
148     pitch, intensity, energy
149 }, 1);
```

The LaTeX source of this thesis is based on the *Legrand Orange Book* template by Mathias Legrand and Vel, which can be downloaded from `https://latextemplates.com`.