Presentation Adaptation for Multimodal Interface Systems:

Three Essays on the Effectiveness of User-Centric Content and Modality Adaptation

Inaugural Dissertation

to Obtain the Academic Degree of a
Doctor in Business Administration at the

University of Mannheim

submitted by

Melanie Heck, M.Sc.

from Stuttgart

Dean: Joachim Lutz
Referent: Prof. Dr. Christian Becker
Correferent: Prof. Dr. Armin Heinzl

Day of oral examination: March 24, 2023

Abstract

The use of devices is becoming increasingly ubiquitous and the contexts of their users more and more dynamic. This often leads to situations where one communication channel is rather impractical. Text-based communication is particularly inconvenient when the hands are already occupied with another task. Audio messages induce privacy risks and may disturb other people if used in public spaces. Multimodal interfaces thus offer users the flexibility to choose between multiple interaction modalities. While the choice of a suitable input modality lies in the hands of the users, they may also require output in a different modality depending on their situation. To adapt the output of a system to a particular context, rules are needed that specify how information should be presented given the users' situation and state. Therefore, this thesis tests three adaptation rules that – based on observations from cognitive science – have the potential to improve the interaction with an application by adapting the presented content or its modality.

Following *modality alignment*, the output (audio versus visual) of a smart home display is matched with the user's input (spoken versus manual) to the system. Experimental evaluations reveal that preferences for an input modality are initially too unstable to infer a clear preference for either interaction modality. Thus, the data shows no clear relation between the users' modality choice for the first interaction and their attitude towards output in different modalities.

To apply *multimodal redundancy*, information is displayed in multiple modalities. An application of the rule in a video conference reveals that captions can significantly reduce confusion. However, the effect is limited to confusion resulting from language barriers, whereas contradictory auditory reports leave the participants in a state of confusion independent of whether captions are available or not. We therefore suggest to activate captions only when the facial expression of a user – captured by action units, expressions of positive or negative affect, and a reduced blink rate – implies that the captions effectively improve comprehension.

Content filtering in movies puts the character into the spotlight that – according to the distribution of their gaze to elements in the previous scene – the users prefer. If preferences are predicted with machine learning classifiers, this has the potential to significantly improve the user' involvement compared to scenes of elements that the user does not prefer. Focused attention is additionally higher compared to scenes in which multiple characters take a lead role.

Contents

List of Figures

List of Tables

Abbreviations

$\boldsymbol{R_F}$	functional requirement
$\boldsymbol{R_{NF}}$	non-functional requirement
AOI	area of interest
AR	augmented reality
AU	action unit
CNN	convolutional neural network
EAR	Eye Aspect Ratio
FACS	Facial Action Coding System
FER	facial expression recognition
GUI	graphical user interface
HCI	human-computer interaction
IoT	Internet of Things
ISO	International Organization for Standardization
k-NN	k-nearest neighbor
LSTM	long short-term memory
MIT	Massachusetts Institute of Technology
NN	neural network
STT	speech-to-text
SVM	support vector machine
TLX	Task Load Index
TTS	text-to-speech
W3C	World Wide Web Consortium

1. Introduction

Communication with other humans and the environment has become a naturally multimodal affair in the course of human evolution (Cherry, 1966; Tomasello, 2010). Humans perceive the world through sight, hearing, touch, smell, and taste and employ their senses to communicate intentions or manipulate the environment (Partan, 2013). User interfaces have traditionally focused on manual input and the visual presentation of information on a display (Karray et al., 2008). While such interfaces are suitable for most conventional applications, they lack the flexibility to meet the demands of increasingly ubiquitous technology and smart spaces. In such environments, the context – and thus the utility of a communication channel – may change in an instant (Jeng, 2009). A change of context can, for instance, occur when the user moves to a different location (Want et al., 1992). It can also be triggered by sudden changes in the external environment, such as the arrival of other people (Schilit et al., 1994). Multimodal interfaces address the new requirements that emerge in such dynamic contexts by letting the users choose the most convenient communication modality – for both system input and output – given their personal preferences and current situation (Oviatt, 2003b; Turk, 2014; Yusupov & Ronzhin, 2010).

The history of multimodal interfaces spans several decades. What began as a purely academic interest at the Massachusetts Institute of Technology (MIT) with the 'Put-That-There' system (Bolt, 1980) – an interface that allows users to manipulate objects through spoken commands when pointing at them – soon started to penetrate the consumer market. Commercial appliances include the iPhone (Apple, 2022b) and Apple Watch (Apple, 2022a), fitness trackers (Fitbit, 2022), the Amazon Echo Show smart display (Amazon, 2022), and in-vehicle infotainment and entertainment systems (Infineon, 2022; STMicroelectronics, 2022). These developments – most notably smartphones and smartwatches – have started to turn Mark Weiser's vision of pervasive computing, where interfaces 'disappear' into the background (Weiser, 1991), into a not too distant reality.

Smart devices now accompany and serve users wherever they go and can provide a seemingly unlimited amount of information (Sezer et al., 2018). Through an ever increasing diversity of sensors including cameras, microphones, and touch screens, users can interact with devices in the way that feels most natural to them, without consciously being aware of using technology (Saha & Mukherjee, 2003).

The consequence, however, is a constant influx of information and incessant stimulation of all sensory receptors of the human anatomical form. It is often difficult for the user to distinguish relevant pieces of information from others that can safely be dismissed (Roetzel, 2019). This, in turn, led to a call for adaptive systems that take it upon themselves to decide what information is relevant and how to best present it to the user (Reeves et al., 2004). The call has been answered by a new stream of research which we will henceforth refer to as 'adaptive multimodal interfaces' (Jameson, 2007; Kong et al., 2011; Langley, 1999; Maybury, 1994). Such systems select an interaction modality (Duarte & Carriço, 2006) or change the content and its structure (Firmenich et al., 2019) in response to the (dynamic) state of the users, their task, and the current situation. Adapting the interface to the user in such a way carries the promise of more efficient, effective, and natural interactions (Maybury & Wahlster, 1999).

1.1. Problem definition

Defining adaptation strategies for multimodal interfaces is no trivial undertaking. Characteristics of the device and its available resources (Boll et al., 1999; Elting et al., 2002; Prabhakaran, 2000) or of the task (Arens et al., 1991; Engen et al., 2014; Kerpedjiev et al., 1997; Lee et al., 2001) that affect the usability of a modality have been researched extensively. The challenge in application contexts that are becoming more and more dynamic and seek to cater the needs of increasingly diverse users is the impracticability of a one-fits-all approach (Sebe, 2009). People of all social backgrounds and cultures have access to technology (OECD, 2022), special solutions for people with physical or cognitive impairments are readily available (Raja, 2016), senior citizens embrace technology (Anderson & Perrin, 2017), and children are entrusted with it at a very young age (Erikson Institute, 2016). The two key issues are therefore to determine (1) what content is relevant

to a user – considering both individual interests and relevance for achieving an overarching goal – and (2) through which modality they would like to access it.

Of course, the notion of what information is relevant highly depends on the user's personality, goals, current mood, and situation (Shokeen & Rana, 2018). Human interests are as diverse as their personalities (Ackerman & Heggestad, 1997), and while one person may appreciate being recommended horror movies on Netflix, another person could feel unsettled by the trailer being shown on the front page of their account; Reading a romance novel may be the perfect pastime for a clerk after a long day in the office, whereas the same person might consider a morning ride on the tram the perfect opportunity to leaf through the pages of a textbook on investment strategies.

With regard to the second issue, the disparate physical and intellectual abilities of the users regulate how they interact with their devices. For example, vision impairments make it difficult or even impossible to understand graphical information, whereas auditory communication channels are inaccessible to deaf users (Kimura, 2018). People with cognitive impairments benefit from simplified information. Others have no such constraints in general, but may find themselves in situations where they cannot or do not wish to use a certain mode of communication. Examples are noisy places where spoken communication is difficult or, on the other extreme, quiet spaces such as a library or the rest area in a train where one would be reprimanded for talking (Cowan et al., 2017). Manual or visual communication, on the other hand, is impractical when the hands or eyes are needed to perform another task, for instance while driving (Zue & Glass, 2000). Besides, how users prefer to interact with a device is highly individual (Oviatt, 1999). Some users categorically prefer reading the captions of muted videos on social media (Patel, 2016), whereas others even install special software to exclusively listen to the audio track and save bandwidth (TechViral, 2022).

With the availability of big data and data mining techniques, online service providers are able to address individual preferences and personalize their services (Matz & Netzer, 2017). For instance, Netflix, Instagram, and Amazon personalize feeds and recommendations based on content that the user has interacted with in the past – be it by looking at it, writing comments and reviews, or submitting a rating (Anshari et al., 2019). What these applications ignore is the potential to implicitly and unobtrusively detect the user's dynamically changing state – a

potential that is inherently provided by the various input channels of multimodal interfaces. Making use of the full range of available sensors that collect data from the user enables adaptive behavior in any application and context, irrespective of whether the user currently provides active input and, if so, through which modality. Such a behavior requires rules that specify how the multimodal interface adapts its interaction modality and/or content. To serve the user, they need to be aware of what communication modalities are available given the system and external context and whether information can be filtered without compromising the overarching goal (Duarte & Carriço, 2006; Maybury & Wahlster, 1999).

Against this backdrop, the objective of this thesis is to define and evaluate user-centric adaptation rules whose potential to increase usability receives empirical support from cognitive psychology. We implement different adaptation rules and systematically study their prospects and limitations. The design process is driven by the ambition to produce adaptation rules that can be applied to a broad range of application scenarios and are equally beneficial to all demographic subgroups. Thus, while the rules are tested on specific use cases, moderation analyses appraise their utility for different contexts and target groups. In the following, we describe in more detail the functional requirements ($\boldsymbol{R_F}$) and non-functional requirements ($\boldsymbol{R_{NF}}$) based on which the adaptation rules are defined and evaluated.

1.2. Requirements analysis

The goal of this thesis is to design and evaluate adaptation rules that personalize multimodal interfaces in a way that best supports individual users in their current context. The requirements are driven by the constraints that are set by the research problem discussed in the previous section. We identify five functional requirements ($\boldsymbol{R_F1}$ - $\boldsymbol{R_F5}$) and three non-functional requirements ($\boldsymbol{R_{NF}1}$ - $\boldsymbol{R_{NF}3}$).

1.2.1. Functional requirements

The adaptation rules shall meet five functional requirements. They constrain the configuration space of possible changes that are applied to the interface and specify demands on the adaptation conditions.

***R_F1* – Adaptive presentation.** The interface shall determine autonomously whether an adaptation is needed. If the system is not in its target state – defined by the application designer – a change of the presentation modality or content is initiated. Thus, the system adapts without any explicit action from the user, contrasting it to adaptable systems, where the users themselves can tailor the configurations of a system to their personal preferences (Oppermann, 1994).

***R_F2* – User-centricity.** Adaptation can be conditioned on the state of the system (e.g., available hardware resources and task), the physical environment (e.g., time and location), or the user (Schmidt et al., 1999). While adaptive applications deployed in the real world should always consider the interplay between all three context dimensions, this thesis focuses on adaptation strategies that respond to a model of the user.

***R_F3* – Dynamic states.** User models typically encompass both static characteristics like demographics or personality traits that remain the same throughout the interaction, and dynamic state variables such as emotions or cognitive load that may change from one instant to another (Firmenich et al., 2019). Since context changes can happen at any moment, the interface shall be aware of and respond to the dynamic state of the user. This enables it to cater to the immediate needs and preferences of the user, instead of assuming that dominant interaction patterns and preferences remain constant irrespective of the current context.

***R_F4* – Non-invasiveness.** Data from which inferences are made about the user shall be collected in the background through non-invasive methods. In contrast to explicit data collection, such implicit techniques extract information from the users' natural interactions with the application, without interrupting their current task through prompts for feedback (Gauch et al., 2007). Aside from interaction logs, a variety of data from multiple sensors including cameras, microphones, and physiological sensors is available in multimodal systems (Sebe, 2009; Turk, 2014).

***R_F5* – Ubiquitous deployment.** The state monitoring and execution of the adaptation shall be feasible with consumer-grade devices. In the conception of this thesis, the term 'consumer-grade device' subsumes any personal computer or mobile device equipped with a common camera and graphical user interface (GUI) for visual communication, a mouse or touch display for tactile, and a microphone and audio player for auditory interaction. This excludes adaptation

strategies that manipulate the external environment, e.g., by communicating with other appliances or using external actuators. It also excludes strategies that require input data from physiological or ambient sensors, or any other external hardware components. Eye tracking data – while typically collected with specialized hardware due to their superior performance – can in theory be extracted from images captured with a normal camera (Papoutsaki et al., 2016; Zhang et al., 2019), and is therefore included as a possible input source.

1.2.2. Non-functional requirements

Three non-functional requirements specify general objectives that the adaptation shall meet. The metrics defined in these requirements form the basis for evaluating the effectiveness of the adaptation rules.

$\boldsymbol{R_{NF}1}$ **– Usability.** The adaptation shall improve the user's interaction experience with the multimodal interface. The International Organization for Standardization (ISO) defines usability in human-computer interaction as a three-dimensional construct of efficiency, effectiveness, and satisfaction (*Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts*, 2018). The accuracy and completeness of the task determine effectiveness, whereas efficiency is defined by the time, human effort, and financial or material resources that are consumed in its execution (Tchankue et al., 2011). The latent variables that comprise satisfaction can be operationalized by means of subjective ratings of usability and engagement (Maat & Pantic, 2007; Peng et al., 2018).

$\boldsymbol{R_{NF}2}$ **– Universality.** The adaptation rule shall benefit all users – independent of their abilities, cognitive patterns or preferences, and demographics. An adaptation can be beneficial to users of a certain subgroup, while being futile or even detrimental to others. For example, the adaptive enterprise system Gaze-X has been perceived as useful by novice users, but provides no noticeable benefit for experts (Maat & Pantic, 2007). In contrast, the claim of a universal adaptation rule is that it benefits a broad variety of users.

$\boldsymbol{R_{NF}3}$ **– Robustness.** Depending on the user state dimension upon which the adaptation is based, the required amount of data can range from singular to prolonged time-series data (Sebe, 2009). Rapid learning algorithms such as one-shot learning that are specifically tuned to scarce input situations can learn

the user's state with a minimal amount of training samples (Vinyals et al., 2016). However, optimized algorithms have only been developed for a small number of very specific problems and – even if available – are futile when the user only passively consumes content and does not actively provide any interaction data (Wang et al., 2020). This can, for instance, be the case in movie streaming or reading. We therefore stipulate that, in order to provide robustness, it shall be possible to apply the adaptation rule even if interaction data is sparse.

1.3. Design science research methodology

The creation of artifacts in disciplines related to systems engineering typically follows an iterative process (Simon, 1988). This thesis adopts the design science research methodology by Peffers et al. (2007) which describes the de facto standard process for the development of software artifacts. The definition of 'artifact' encompasses theoretic constructs, models, or methods, as well as concrete instantiations of a resource or its properties. This thesis seeks to define and validate adaptation rules with the intention to enhance the experience of users interacting with a multimodal interface. Designing an executable software product is not the ultimate goal. Nevertheless, in order to evaluate the adaptation rules, they are integrated into a multimodal interface and tested in the context of prototypical applications. Figure 1.1 depicts the six steps of the process.

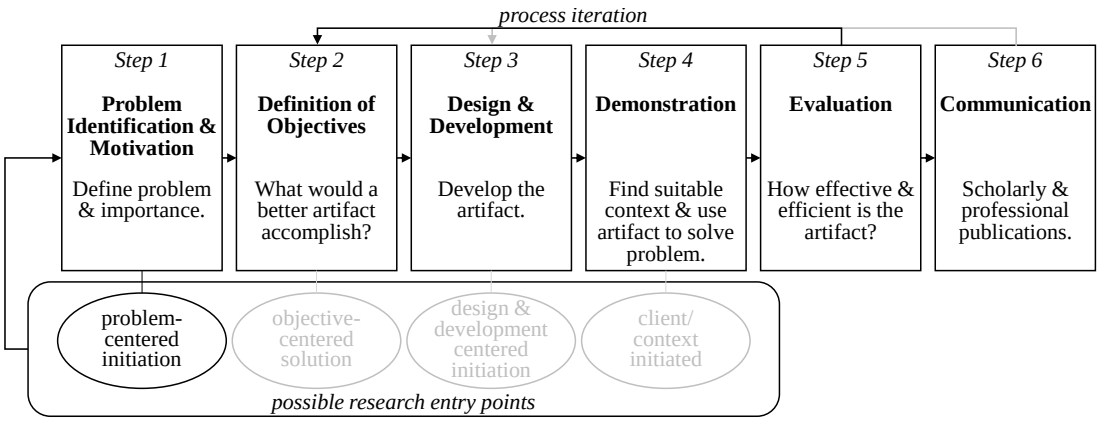


FIGURE 1.1: Design science research methodology. Adapted from Peffers et al. (2007). The development of adaptation rules follows an iterative process, starting with (1) *problem identification & motivation*, and culminating in (6) *communication* of the findings. For each adaptation rule, a new iteration of (2) *definition of objectives*, (3) *design & development*, (4) *demonstration*, and (5) *evaluation* is initiated. Alternative process sequences that were not applied in this thesis are marked in gray.

Following a problem-centered initiation, we identified a lack of empirically validated adaptation rules that specify how multimodal interfaces should adapt their communication modality and/or content to the user's current state in order to improve the interaction for them. The output of the first step of the process, in which we defined and motivated the research problem, is presented in Chapter 1.1. It explains the relevance of interfaces that adapt the presentation of information to the dynamic needs of their users, and identifies a lack of rules to enable such adaptive behavior. In the second step, the objectives of the artifact were defined in the form of requirements (Chapter 1.2). Five functional requirements narrow down the configuration space of possible interface adaptations and specify demands on the adaptation conditions. Three additional non-functional requirements define the desired outcomes of the adaptation. After an adaptation rule was designed (*step 3*), we integrated it into an application to demonstrate its use in a suitable context (*step 4*). Using an experimental research design, it was then evaluated against the metrics defined in the non-functional requirements (*step 5*).

In the first iteration of the process (Chapter 3), we defined an adaptation rule for modality alignment and integrated it into a smart home display:

ADAPTATION RULE 1: Modality alignment

```
if preferredInput = speech then
    call activateAudioOutput()
    call disableTextOutput()
else
    call disableAudioOutput()
    call activateTextOutput()
end if
```

The evaluation revealed that the rule violates the robustness requirement ($\boldsymbol{R_{NF}3}$), as the accurate inference of modality preferences is conditional on prolonged interactions. Thus, a second iteration of the process, starting with a refinement of the objectives, was initiated (Chapter 4). In order to overcome the shortcomings of the first adaptation rule, a stronger emphasis was placed on robustness. The functional requirements for dynamic user-centric adaptation ($\boldsymbol{R_F1}$ - $\boldsymbol{R_F3}$), in return, were relaxed. Consequently, we implemented a non-adaptive parallel

activation of two presentation modalities to mitigate confusion and tested it in a virtual meeting application. It can be applied without having any data about the user, and is therefore robust to situations with no or little user interaction. The following rule for multimodal redundancy was defined:

ADAPTATION RULE 2-1: Multimodal redundancy

```
call activateAudioOutput()
call activateTextOutput()
```

The experimental observations revealed that the benefit of bimodal presentation – more specifically, a spoken discourse with subtitles – is conditional on its perceived usefulness, as well as on the origin of the confusion. As a result, we propose to augment the rule with an adaptive component that deactivates the secondary visual modality when multimodal redundancy leads to no improvement in performance:

ADAPTATION RULE 2-2: Selective multimodal redundancy

```
call activateAudioOutput()

while audioActive do
    if confusionAtStart then
        call activateCaptions()
    end if

    while captionsActive do
        if call checkConfusion() >= confusionAtStart then
            call disableCaptions()
        end if
    end while
end while
```

Since this violates the universality requirement ($\boldsymbol{R_{NF}2}$), a third iteration of the process was started in which the objectives of robustness and universality were prioritized (Chapter 5). The adaptation rule for content filtering achieves robustness by using gaze data to infer user states, which are continuously collected

in the background with no user interactions required. Similar to Rule 1, an adaptation condition based on personal preferences pledges universality:

ADAPTATION RULE 3: Content filtering

```
preferredObject = object_1

for all objects in scene:
    if dwell(Object_i) > dwell(preferredObject) then
        preferredObject = object_i
    end if
end for

call getNextScene(preferredObject)
```

The insights gained throughout the process – including the benefits of each adaptation rule and their limitations – were discussed in multiple workshops and seminars (cf. Table 1.1). The results from the first and third process iteration were presented at distinct holdings of the *Annual Conference on Intelligent User Interfaces (IUI)* and condensed versions of the research essays were published in the conference proceedings. The results of the second process iteration are part of a larger project and are currently in submission.

1.4. Contributions

The three essays presented in this thesis address the research problem by implementing and evaluating the effectiveness of the three adaptation rules defined in the previous section. Each adaptation responds to the user's state which, in turn, is inferred from implicit input. Table 1.1 provides an overview of the essays.

TABLE 1.1: Overview of the three research essays included in the thesis. The confusion detected in facial gestures (in *italic*) is only used as adaptation source in the modified (selective) multimodal redundancy rule of Essay 2 and was not empirically tested.

	Essay 1	**Essay 2**	**Essay 3**
Title	Does using voice authentication in multimodal systems correlate with increased speech interaction during non-critical routine tasks?	Evaluating the potential of caption activation to mitigate confusion inferred from facial gestures in virtual meetings	EyeDirect: A gaze contingent system for personalized video display
Research questions	**RQ1:** Do users perform authentication and routine tasks with the same modality? **RQ2:** Does presentation in a modality that is compatible with the user's input improve satisfaction?	**RQ1:** Can auto-generated captions improve comprehension? **RQ2:** Do facial gestures reveal confusion in virtual meetings?	**RQ1:** How can preferences for video elements be extracted from gaze? **RQ2:** Does personalizing videos increase engagement?
Context	Smart home display	Virtual meeting	Movie streaming
Adaptation rule	Modality alignment	*(Selective)* multimodal redundancy	Content filtering
Adaptation source (user state)	Input modality chosen by the user	*Confusion detected in facial gestures*	Preferences identified with eye tracking
Adaptation target	**Presentation**: Audio versus text output	**Presentation**: Activation of captions	**Content**: Protagonizing preferred elements
Method	Laboratory experiment	Laboratory experiment	Laboratory experiment
Analytic strategy	Statistical analysis of quantitative data	Statistical analysis of quantitative data	Adaptation simulations & statistical analysis of quantitative data
PUBLICATION STATUS (AS OF JANUARY 2023):			
Co-authors	Shon, Becker	Jeong, Becker	Edinger, Bünemann, Becker
Presented	**2022**: IUI **2022**: ACOCA workshop at Deakin University (Prof. Zaslavsky)	**2022**: ACOCA workshop at Deakin University (Prof. Zaslavsky)	**2021**: IUI **2021**: Research talk at SMU (Prof. Misra) **2022**: ACOCA workshop at Deakin University (Prof. Zaslavsky)
Published	Accepted at IUI (CORE Ranking A)	*in submission*	Accepted at IUI (CORE Ranking A)

The methodology for answering the research questions is based on laboratory experiments with primarily quantitative data analysis methods. In each experiment, one of the three adaptation rules is integrated into a different application, either from a commercial provider (Essay 2), or implemented specifically for the purpose of the experiment (Essay 1 and Essay 3). In Essay 1, we implement a smart home display through which ambient settings can be manipulated. Essay 2 is situated in the context of a virtual meeting. For the evaluation in Essay 3, we design and implement a framework for adaptive movie display, called *eyeDirect*.

1.4.1. Overview of Essay 1

Empirical evidence has firmly established that individuals differ strongly with regard to their preferences for interaction modalities (Oviatt et al., 2004). Examples of applications that recognize this variability include messenger application which may take either keyboard or voice input and present incoming responses as text or audio messages – whichever best suits the user's context and preference. At the same time, a person's interaction patterns are persistent and can typically be detected from the first inputs (Oviatt et al., 2003, 2005; Xiao et al., 2002). Authentication is often the first point of interaction with an application. The users' login behavior can thus be used to immediately adapt the communication modality to their preferences. Yet, given the sensitive nature of authentication, this interaction may not be representative of the user's inclination to use speech input in non-critical routine tasks. In a first step, Essay 1 therefore explores whether the interactions during authentication differ from non-critical routine tasks in a smart home application:

> **RQ1.1** Do users perform authentication and routine tasks with the same modality?

The results of an experimental study with 41 participants show that the authentication behavior does not correlate with the observed interaction behavior during non-critical tasks, nor with the perceived usability of speech input.

In a second step, Essay 1 explores the effectiveness of aligning the presentation with the modality in which the user provides input (Rule 1). Given the observed dependence of modality preferences on the task, the effect is assessed for two

different scenarios: Depending on the scenario, the presentation is compatible with the input modality that the user chooses to complete (1) authentication, or (2) routine tasks.

RQ1.2 Does presentation in a modality that is compatible with the user's input improve satisfaction?

The evaluations reveal that aligning the presentation modality to user input from short interactions has no effect on their attitude towards audio presentation, independent of whether the adaptation responds to input from authentication or non-critical tasks.

Essay 1 thus demonstrates that predicting modality preferences requires data from prolonged and extensive interactions, which violates the non-functional requirement for robustness ($\boldsymbol{R_{NF}3}$). Therefore, if the user's dominant modality can not reliably be inferred from the available data, fusion and fission techniques (cf. Section 2.1) should be applied in order to prevent user frustration when they can no longer communicate with the system in the modality of their choice.

1.4.2. Overview of Essay 2

In many situations, explicit input is not provided or its quantity is not sufficient to reliably infer the user's dominant modality. For instance, a user passively watching a video or attending a virtual meeting with a large number of participants is unlikely to interact much. Essay 2 therefore explores whether, under such circumstances, it can be beneficial to present information through multiple modalities simultaneously. A non-adaptive strategy is chosen based on empirical evidence suggesting that static interfaces can be preferred over their adaptive counterpart if perceived as more comprehensible and predictable (Shneiderman, 1997). Essay 2 investigates the effectiveness of the strategy in the context of a virtual meeting.

During the COVID-19 pandemic, virtual meetings have become an integral part of collaboration in industry, academia, and other parts of society and are now likely to persist and complement our daily routines. Thus, it is important that meeting participants understand discussed topics as smoothly as in physical encounters.

However, attendants often experience confusion, but are hesitant to signal their situation out of timidity or politeness. Essay 2 thus applies multimodal redundancy (Rule 2) to investigate whether captioning auditory output is a suitable tool for mitigating confusion:

RQ2.1 Can auto-generated captions improve comprehension?

The results of a study with 45 Google Meet users reveal that multimodal presentation can be beneficial, but with some reservations that violate the non-functional requirement for universality ($\boldsymbol{R_{NF}2}$). Captions can help overcome non-understanding due to poor audio quality or language deficiencies – but not confusion resulting from contradictory or incongruent information – as long as they are perceived as useful.

To mitigate negative side effects such as occlusion of important visual information when captions are not strictly needed, Essay 2 proposes activating them dynamically in a proactive way only when a user effectively experiences confusion. Assuming that the non-understanding issues were not a transient state, but rather caused by persisting auditory or language induced issues, the adaptation seeks to improve the comprehension of the following spoken discourse. To determine instances that require captioning, it tests whether the subliminal cues from facial gestures – specifically blinks, facial action units, and the expression of emotions – can be used to detect confusion:

RQ2.2 Do facial gestures reveal confusion in virtual meetings?

A quantitative analysis confirms that confusion during purely auditory presentation activates six action units – specific facial regions that are defined by the fundamental movements of a muscle group (Ekman & Friesen, 1978). With captioning, it additionally leads to less blinks and expressions of neutral emotion.

1.4.3. Overview of Essay 3

The first two essays reveal important limitations of modality adaptation. Its effectiveness is contingent on situations in which user data can be readily collected and the presentation of information in multiple modalities is perceived as useful,

rather than a distraction. Essay 3 therefore passively monitors the users' gaze to create a model of their current state that is robust to low-input situations. Using the dynamic state information, it investigates whether adaptations that target the content itself are more universally beneficial to different target groups.

The experimental investigation in Essay 3 is conducted in the context of movie streaming. The constant distractions from ubiquitous technologies have turned television into a side event rather than a deliberate pastime. Movie directors thus struggle to find new ways to sustain the attention of their audience. Interactive movies usually require the viewer to actively decide how the plot progresses, creating an experience more akin to video games than film (Netflix, 2021a, 2021b).

Essay 3 therefore evaluates whether tailoring the content of a video to individual preferences can create higher engagement. Adaptive videos are created with *eyeDirect*, a system that analyses gaze data and personalizes the plot of a video without the viewer's active intervention. User preferences are inferred from their gaze distribution to different elements in a scene. The subsequent scene then dynamically zooms in on the object or person of the user's predicted preference.

In a laboratory experiment with 175 participants, the adaptation is evaluated with regard to two research questions:

> **RQ3.1** How can preferences for video elements be extracted from gaze?
>
> **RQ3.2** Does personalizing videos increase engagement?

Essay 3 identifies multiple gaze features that can effectively identify preferences, and demonstrates that personalized videos have a positive effect on focused attention and involvement, but not on novelty perception.

The findings from the experimental investigation in Essay 3 have important implications for the use of content adaptation in movie streaming. Adaptation can increase involvement and attention to videos by putting into focus the viewer's preferred elements.

1.5. Structure of the thesis

This chapter motivated the research problem that the thesis addresses and defined the requirements for user-centric adaptation rules in multimodal systems. Further, it outlined the design process leading to the development of three adaptation rules and provided an overview of the research essays in which they were communicated. Chapter 2 lays the conceptual foundations for the thesis, starting with a general framework for adaptation in multimodal systems. Strategies for user-centric adaptation of multimodal interfaces in related literature are discussed, and theories on which the proposed adaptation rules are grounded are introduced. Chapter 3 and Chapter 4 present two research essays that investigate strategies for modality adaptation. The research essay presented in Chapter 5 discusses the effectiveness of content adaptation. Chapter 6 concludes the thesis with a summary of the theoretical contributions and practical implications for the design of adaptation strategies in multimodal systems. Building on this discussion, the chapter reflects upon limitations and future research directions.

2. Theoretical foundations & related literature

Adaptive multimodal interfaces have emerged as an interdisciplinary field in which adaptivity adds an additional layer of complexity to the challenges of multimodal interfaces. This chapter lays the theoretical foundations for these systems, starting with a brief definition of multimodal interfaces and the introduction of a general framework for their adaptive realization. Theories from cognitive science are introduced that lay the conceptual foundations for the formulation of promising user-centric adaptation rules. Finally, concrete implementations are discussed by means of related literature that presents solutions for a variety of domains.

2.1. Definition of multimodal interfaces

Multimodal systems receive and respond to input through more than one modality which is typically associated with one of the human senses for visual, touch, or auditory communication (Turk, 2014). Following the definition proposed by Sebe (2009), this thesis considers an interface as multimodal only if multiple communication channels are available for input and/or output. Thus, an interface that accepts speech and touch input, or can deliver the output as either an auditory or a written text message is considered multimodal. In contrast, an interface that extracts gaze data from a commercial eye tracker and a common camera, or announces messages through a voice assistant and acoustic warning signals does not classify as multimodal.

The first multimodal system considered as such was developed in 1980 at MIT (Bolt, 1980). The 'Put-That-There' interface enables users to identify virtual objects on a display by pointing at them, and then manipulate the selected object through speech commands. Over the course of the years, an abundance of architectures for designing multimodal interfaces have been developed (Dumas et al., 2009; Turk, 2014), most of them following the reference framework for multimodal interaction of the World Wide Web Consortium (W3C, 2003). Common to all architectures is

the integration of components for synchronizing input and output from multiple modalities. Figure 2.1 depicts the general architectural design that, in compliance with the W3C framework, typically underlies a multimodal interface (Dumas et al., 2009). Input from various sensors is aggregated and interpreted by a *fusion* component (Lalanne et al., 2009; Sebe, 2009). A *fission* component determines the most appropriate output channels with regard to the context and prepares synchronized content in multiple modalities (Boll et al., 1999; Dalal et al., 1997; Dumas et al., 2009).

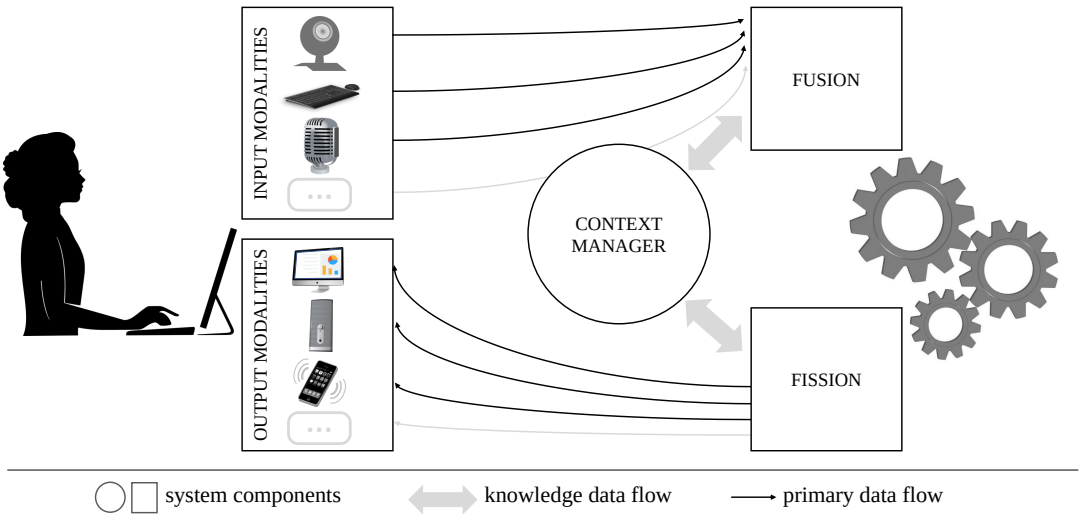


FIGURE 2.1: Architecture of multimodal interfaces. Adapted from Dumas et al. (2009). Multimodal input is integrated by a *fusion* component. A *fission* component manages multimodal output by creating and synchronizing content in multiple modalities that are suitable for the *context*.

As the number – and thus also the diversity of computer users and the variability of their physical surroundings – increases, the call for making multimodal interfaces context-aware has grown louder (Kim et al., 2021). In a set of guidelines, Reeves et al. (2004) specify that modality integration shall take into account user preferences and abilities, and present output in the most appropriate format for their current context. Such behavior requires a system with self-adaptive capabilities.

Adaptive multimodal interfaces (Jameson, 2007; Kong et al., 2011; Langley, 1999) have been introduced under a variety of alternative terms, including 'intelligent multimedia interfaces' (Maybury, 1994) and 'perceptual interfaces' (Turk, 2014). While different names are used to describe the systems, the common idea is to use the context, including the current state of the user, to represent information in a way that best supports the execution of a given task.

Research in the domain of adaptive multimedia has primarily focused on optimizing the technical quality of media presentations given the hardware capabilities of the device and the requirements of the task (Boll et al., 1999; Prabhakaran, 2000). In contrast, this thesis adopts the user-centric perspective of human-computer interaction (HCI). Thus, adaptation is driven by characteristics of the users, including their preferences, abilities and cognitive state (Reeves et al., 2004). In the following, we discuss how such user-centric adaptive functionalities can be integrated into multimodal interfaces. A generic framework for self-adaptive systems is introduced and the relevant design elements for its application to multi-modal interfaces are defined.

2.2. Adaptation in multimodal interfaces

Weyns et al. (2012) defined self-adaptation of a system as the "*capability to adapt itself to internal dynamics and dynamics in the environment in order to achieve certain goals.*" Brun et al. (2009) further specified that "*the systems [must] decide autonomously (i.e., without or with minimal interference) how to adapt or organize to accommodate changes in their contexts and environments.*" This behavior is typically realized through a feedback loop that reacts to changes in the controlled process (Krupitzer et al., 2015). Most self-adaptive systems are based on the MAPE-K cycle introduced by Kephart and Chess (2003). It assigns the responsibilities of the adaptation engine to four central components: **M**onitor, **A**nalyzer, **P**lanner, and **E**xecutor (MAPE). An additional shared **K**nowledge base (K) stores information persistently over multiple feedback loops. Figure 2.2 illustrates how the MAPE-K cycle can be applied to realize adaptivity in multimodal interfaces.

The *monitor* collects raw data from the managed resource – comprising all software and hardware components of the adaptable system – and processes it by applying validation, filtering, and clustering techniques. In multimodal interfaces, data is typically collected through sensors such as the device camera, a microphone, or the physiological sensors of wearable devices (Baig & Kavakli, 2019; Karpov & Yusupov, 2018; Turk, 2014). User input from the keyboard, mouse, or touch display is another rich source of information (Kim et al., 2021).

The *analyzer* reasons about the data and decides whether an adaptation is needed to achieve the objective of the system. In the HCI domain, this typically entails

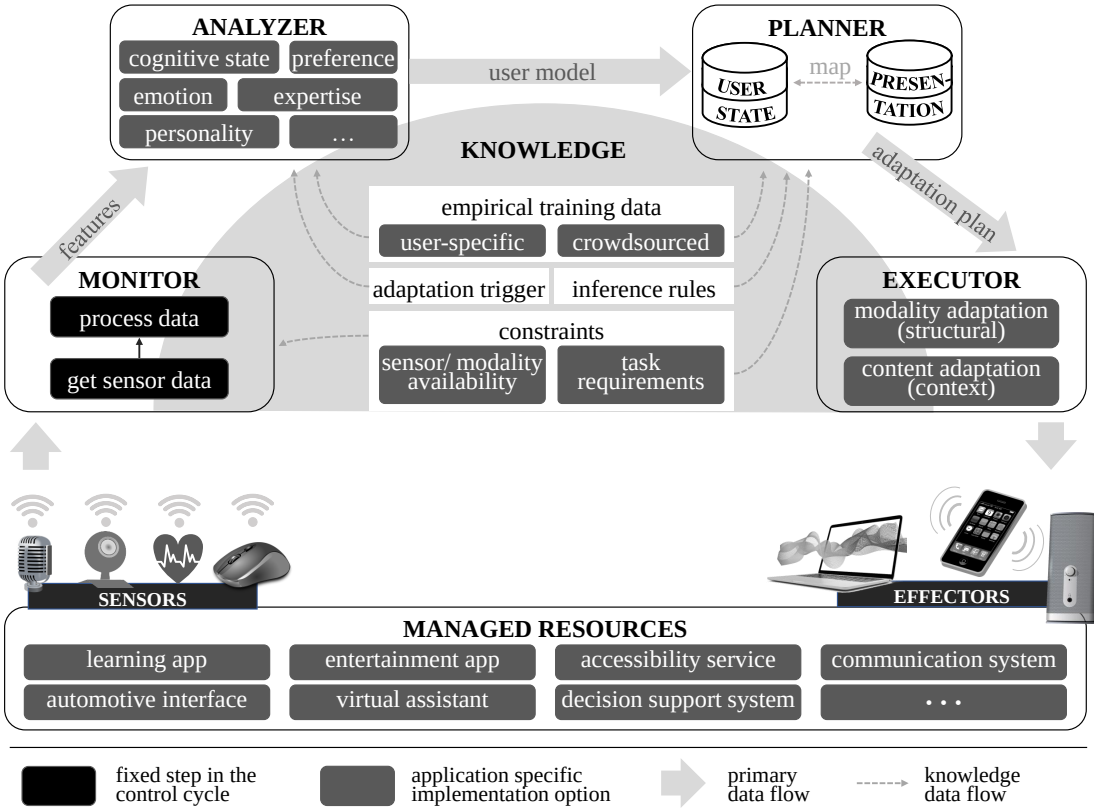


FIGURE 2.2: MAPE-K cycle for adaptation in multimodal interfaces. Adapted from Kephart and Chess (2003). Adaptation responsibilities are divided among an independent monitoring, analysis, planning, and execution component. All four components are connected to a shared knowledge base.

the construction of a user model. Depending on the objectives, the model may, for instance, represent the user's cognitive state (Debie et al., 2021; Mutlu-Bayraktar et al., 2019), preferences (Gal & Simonson, 2021; Hakim & Levy, 2019), emotions (D'Mello & Kory, 2015; Dzedzickis et al., 2020), expertise (Kumari et al., 2017), or personality (Taib et al., 2020). If the user model violates the target state, an adaptation is triggered. For example, cognitive overload leads to a performance decrease (Paas et al., 2004). The expected reaction of an adaptive system would therefore be the initiation of measures that return the user to a healthy state. In addition, adaptation can be triggered by the environment, time, or system and task states (Feigh et al., 2012). Two examples for task state triggers will be discussed in Essay 1 and Essay 3. In Essay 1, adaptation is initiated after successful login. In Essay 3, decisive scene cuts in a movie trigger the adaptation.

By drawing on inference rules, the *planner* determines the configuration space of possible changes that can be applied to accomplish the system's objectives given the state of the user model. If multiple alternative actions are possible, the action that results in the most desirable state is chosen. According to the taxonomy by Feigh et al. (2012), adaptation can take place on four different levels: (1) function allocation (i.e., determining the person or system to perform a task), (2) task scheduling and prioritization, (3) interaction form and modality, or (4) quality and quantity of content. The first two levels target system-level adaptations that may lead to an activity being delayed, delegated to another entity, or not executed at all. In contrast, task-level interface adaptation targets either the modality (level 3) or the content itself (level 4). In the previous example of a user experiencing cognitive overload, one possible configuration could be to present the content in an additional modality, which can reduce extraneous load on the working memory (Gellevij et al., 2002; Mousavi et al., 1995). Essay 2 discusses how effective such a strategy is for reducing confusion. An alternative configuration that targets the reduction of intrinsic load (Paas et al., 2003) by manipulating the amount of simultaneously presented information is discussed in Essay 3.

The *executor* orchestrates the adaptation by sending instructions to the effectors of the managed resource. In case the adaptation plan foresees a modality adaptation, a structural change of the system is necessary. This entails activating or deactivating presentation modalities as specified by the planner. The possibilities for presentation range from visual and auditory to tactile channels, with some interfaces even experimenting on olfactory output (Sarter, 2006). The effects of modality adaptation on the user are studied in Essay 1 and Essay 2. In contrast, content adaptation requires a change of the context, where context refers to the activity that the user performs within a multimodal system (Bakkes et al., 2012; Brusilovsky, 2012). Essay 3 applies content adaptation by tailoring the plot of a video to the viewer's personal preferences.

An additional shared *knowledge base* stores information persistently, thus making it available to the MAPE components in future feedback loops. Knowledge data such as adaptation triggers or inference rules have particular relevance for one specific component. Other knowledge items influence the processes in several components. Examples are training data that can be used to learn user models (Ng et al., 2015; Ouyang et al., 2017) or adaptation rules based on the effectiveness of

past adaptations (Hssina & Erritali, 2019). The sensors and interaction modalities that are available to the managed resource determine which features the monitor can extract from the raw data. They also narrow down the configuration space for the planner. Similarly, both components are influenced by the requirements of the task. For instance, in order to make inferences from gaze data collected while contemplating the dynamic visual scene of a movie, additional processing steps are necessary to remove the bias of motion (Heck et al., 2021). The planner, for its part, should only consider adaptation strategies that do not compromise the completion of the main task. A previous study in which we investigated adaptation strategies for sales terminals demonstrates the significance of this constraint. We identified four content adaptation techniques that are suitable in e-commerce settings: prioritization, filtering, recommendation windows, and product cross-selling (Heck et al., 2019). For self-order terminals in fast food restaurants, we propose applying a filtering strategy to hide irrelevant menu options based on attention cues from eye tracking data. In a different setting, such an approach could be detrimental if content filtering leads to loss of information. One example are educational systems, where students attempting to learn a conceptual topic would have lower learning gains if they were only instructed on a subset of the topic for which they demonstrate interest.

Several frameworks have been developed that specifically focus on handling adaptive presentation in multimodal interfaces (André, 2000; Duarte & Carriço, 2006; Gutierrez et al., 2005; Kerpedjiev et al., 1997; Martin, 1998). However, the specification of how the interface reacts to an adaptation trigger is left to the discretion of the application designer (Maybury & Wahlster, 1999). Since this thesis seeks to define adaptation rules with a strong theoretical foundation, the following section provides an overview of theories from cognitive science. The theories form the pillars on which the adaptation rules presented in the three research essays are built. Each of them is based on the hypothesized effect of how content and modality adaptation affects the usability of multimodal systems.

2.3. Theoretical foundations of adaptation rules

Working memory is a brain system that temporarily stores and manipulates information to solve complex cognitive tasks, which may then ultimately be transferred to the more permanent long-term memory (Baddeley, 1992). Since its capacity is limited, *Cognitive Load Theory* (Sweller, 1988) investigates techniques that reduce the load on working memory. It distinguishes between intrinsic, extraneous, and germane load (Paas et al., 2004).
Intrinsic load increases with the number of logically related elements that have to be processed simultaneously (Paas et al., 2003). Consequently, content filtering can reduce intrinsic load. In digital environments, rich media and emotional designs – often realized through purely decorative elements – generate additional load, but at the same time promote motivation (Skulmowski & Xu, 2021). Essay 3 therefore investigates how content filtering affects engagement. The filtering mechanism specified in **Adaptation Rule 3** effectively reduces the number of related media elements while maintaining their richness and emotional character. This is realized by identifying and prioritizing the user's preferred elements.

In contrast to intrinsic load, extraneous and germane load can be manipulated by how information is presented (Paas et al., 2003). Germane load forms when new information is organized into schemas. While increasing the immediate load on the working memory, the developed schemas allow to automate and accelerate the processing of new information in the long run. Extraneous load, in contrast, is the result of unnecessarily complex instructional material. Redundant information or cross-references that need to be looked up increase extraneous load (Sweller & Chandler, 1991). Since it benefits neither immediate comprehension nor long-term learning, educational material is typically designed to minimize extraneous load (Klepsch & Seufert, 2020; van Merriënboer & Sweller, 2010).
Working Memory Theory (Baddeley & Hitch, 1974) assumes that the working memory consists of multiple largely independent sub-systems that process information perceived and communicated through different senses. Therefore, while working memory capacity is limited, processing in a subsystem is not noticeably disturbed by information from a different modality (Baddeley, 1992).
In multimodal interfaces, the composition of the working memory system leads to competing effects on extraneous load that Sweller et al. (1998) documented in

two principles of Cognitive Load Theory. The *redundancy principle* proposes that replicating information in another modality – for instance, by transcribing spoken discourse into text – can increase extraneous load and should thus be avoided (Kalyuga, 2012). In contrast, the *modality principle* recommends presenting information in multiple modalities in order to use the processing capacity of several working memory sub-systems (Gellevij et al., 2002; Mousavi et al., 1995). The improved understanding of captioned videos, which has been observed for both language learners (Etemadi, 2012; Hayati & Mohmedi, 2011; Perego et al., 2010; Winke et al., 2010; Zareian et al., 2015) and students studying online lectures in their native language (Morris et al., 2016; Zheng et al., 2022), supports the principle. In the evaluation of **Adaptation Rule 2**, we thus test whether the expansion of working memory through multimodal presentation is greater than the additional load imposed by the redundant presentation of content.

Multiple Resource Theory (Wickens, 2002) identifies a limitation of the modality principle. It states that the functionally separate processes of perceiving system output and providing input in response to it – even if performed simultaneously – do not compete for a common resource. Consequently, information processing in the working memory is not necessarily more efficient when input and output are communicated in different modalities.

The symmetry principle, which is rooted in *Gestalt Theory* (Wertheimer, 1938), even suggests that the working memory can process related input and output of the same modality more efficiently. Grounded on Aristotle's claim that "the whole is greater than the sum of its parts", Gestalt theory posits that the combination of perceptual elements gives rise to emergent properties. One of its principles – the principle of symmetry – asserts that humans perceive symmetrical elements as part of a coherent whole. As a result, grouping elements that are communicated through the same sensory modality (and are thus considered to be symmetrical) conserves mental resources (Oviatt et al., 2003).

Further evidence for a beneficial effect of input and output alignment stems from social psychology. *Communication Accommodation Theory* is based on the principle of convergence – a strategy in which individuals assimilate their communication behavior to that of another person in order to be perceived as more likable and make the communication more efficient (Giles et al., 1987). Neurological processes in the mirror and echo neuron systems prepare the human

organism to respond to the actions of another individual by activating brain regions that are associated with these actions (e.g., speaking or moving) (Oviatt, 2017). As a result, communication partners are more likely to respond in the same modality through which they receive information. In human-computer interaction, humans mimic the language (Iio et al., 2015) and movements (Fujiwara et al., 2022) of intelligent conversational agents. In turn, agents that adopt such an assimilation strategy by mimicking the movements of their human communication partner are more persuasive (Bailenson & Yee, 2005). Grounded on these theories, **Adaptation Rule 1** specifies that output is presented in a modality that is considered compatible with the user's input (Schaeffner et al., 2016). Specifically, spoken requests are answered by a voice assistant, whereas manual input is followed by text output. Essay 1 tests the effectiveness of the rule.

While existing adaptation approaches for multimodal interfaces seldom consider the role of cognitive theories, applications have been developed in a variety of domains. In the following, we review how modality and content adaptation has been realized in applications with multimodal interfaces that – while not discussing their theoretical foundations – implement concrete adaptation rules.

2.4. Related literature on user-centric adaptation

Personalized services have been developed for almost all situations in which a person interacts with an information system, most notably education (Apoki et al., 2020; Ennouamani & Mahani, 2017), recommender systems (Burke, 2007), and automotive applications (Murali et al., 2022; Sikander & Anwar, 2019; Wang et al., 2006). Most often, however, they rely on explicit input that the user has to provide – for example in the form of a questionnaire – before being able to use the service (Rothrock et al., 2002). In contrast, the focus of this thesis lies on adaptation that immediately responds to an implicitly inferred user state. In the following, we review related literature on applications that adapt to implicitly inferred, dynamic user states. Some exemplary applications that adapt to the users' general preferences – typically identified through long-time observations – or to explicitly reported user states are discussed briefly. However, the list is by no means exhaustive and only serves the purpose of giving a complete picture of the domains that hold potential for adaptive applications.

TABLE 2.1: Overview of applications with user-centric adaptation rules by application domain. Survey papers are marked in *italic* with an asterisk.

	USER MODEL								ADAPTATION							USER INPUT							
Applications by domain	interaction preference	interest	emotion	fatigue/ distraction	incomprehension	knowledge	demographics	ability	cognitive style	output modality	graphic/ vocal design	augmentation	content order/ selection	notification	help dialogue	ambient adjustment	explicit	task interactions	cursor/ keyboard	speech	facial expression	eye tracking	physiological	body movements
Automotive:																								
*Murali et al. (2022)**	●		●	●						●	●			●		●		●	●	●	●	●	●	●
*Sikander and Anwar (2019)**				●										●				●			●	●	●	
*Wang et al. (2006)**				●										●				●				●	●	
Tchankue et al. (2011)				●									●	●				●						
Nass et al. (2005)			●								●							●						
Nasoz et al. (2010)			●				●		●					●		●							●	
Moniri et al. (2012)							●				●		●							●				
Manufacturing:																								
Josifovska et al. (2019)	●						●	●		●	●		●				●	●						
Smart Office:																								
Adam et al. (2017)			●									●	●		●									
Maat and Pantic (2007)					●							●	●		●							●		
Vertegaal et al. (2003)		●											●									●		
Smart home:																								
Coelho et al. (2011)	●				●			●		●	●		●		●		●	●						
Reithinger et al. (2003)	●									●	●						●			●				●
Education & learning:																								
*Apoki et al. (2020)**	●	●	●			●		●	●	●		●	●				●							
*Brusilovsky (2003)**										●		●	●											
*Ennoumani et al. (2017)**	●	●	●		●	●	●	●	●	●	●	●	●											
*Mousavinasab et al. (2021))**	●	●	●			●	●	●	●	●	●		●	●	●									
Garcia Barrios et al. (2004)			●	●	●							●		●								●		
Duarte and Carriço (2006)	●									●		●		●				●						
Recommender systems:																								
Mobasher et al. (2000)		●											●					●						
Cheng and Liu (2012)		●											●									●		
Kozma et al. (2009)		●											●									●		
Hardoon et al. (2007)		●											●									●		
Salojärvi et al. (2004)		●											●									●		
Xu et al. (2008)		●											●									●		
Linden et al. (2003)		●											●					●						
Kim et al. (2001)		●					●						●				●	●						
Qvarfordt and Zhai (2005)		●										●										●		
Duarte and Carriço (2006)	●									●			●						●	●				
Cinematography:																								
Bolt (1981)		●										●										●		
Starker and Bolt (1990)		●										●										●		
Hansen et al. (1995)		●											●									●		
Vesterby et al. (2005)		●											●									●		
Netflix (2018, 2021a, 2021b)		●											●				●							
Peng et al. (2018)			●								●		●				●							

Table 2.1 summarizes the reviewed applications. Possible forms of *adaptation* range from modifying the presentation form over content changes, up to ambient adjustments. Adaptation targeting the presentation form may select the most suitable output modality, modify the graphic or vocal design, or augment specific elements with either additional information or visual and/or acoustic highlighting. Content changes typically entail some form of ordering and/or selection or hints through notifications and help dialogues.
The decision on how the interface should adapt to achieve a predefined goal is based on a *user model*. Dynamic states that are monitored include interaction preferences, interests, emotions, fatigue or distraction, incomprehension, and knowledge. Additionally, the model can store more permanent characteristics of the user such as demographics, abilities, and cognitive style. While these variables may change over time, they are unlikely to vary over the course of a single interaction.
The *input* for the user model can either be provided explicitly by the user, or implicitly during their interaction with the application. Implicit cues are typically collected from the user's interaction with the application, cursor movements, or from physical sensors that monitor changes in the user's voice, expression, gaze, physiological measurements, or body movements.

2.4.1. Automotive

Car manufacturers are integrating more and more intelligent in-vehicle systems. In addition to infotainment systems that can switch from manual entry on a graphics display to spoken dialogues, warning systems that react to fatigue, emotions, or distraction are emerging (Murali et al., 2022; Sikander & Anwar, 2019; Wang et al., 2006). A number of alternative responses to critical states have been proposed, including visual, acoustic, or vibrotactile alerts, engaging conversations, or ambient adjustments. For example, the multimodal interface for mobile info-communication (MIMI) developed by Tchankue et al. (2011) infers distraction from the driver's performance. In critical situations, the car prompts a warning sound and blocks phone calls to avoid interruptions.
Based on the empirical finding that drivers are more attentive and have less accidents when the car's infotainment system responds in a voice that matches their emotions, Nass et al. (2005) propose affective voice assistants. While not

yet implemented, the authors suggest that emotions may be inferred from facial expressions, sensors attached to the steering wheel, or vocal cues.
Nasoz et al. (2010) monitor physiological sensor readings from a wristband to infer emotions. Depending on the driver's personality traits, the system responds to emotional states with one out of multiple adaptation options. Possible measures include changing the radio station, suggesting a rest stop or relaxation exercise, rolling down the window, splashing water into the driver's face, telling them to calm down, and making a joke. Through reinforcement learning, the driver's emotional reactions to the applied measures are used to gradually refine the adaptation strategy.

Although less prevalent, adaptation to static user characteristics has been explored. Moniri et al. (2012) use acoustic cues from spoken input to infer the gender and age of drivers in a rental car. Gender information is used to select a voice assistant that matches the driver's gender and filter for female parking slots. For senior drivers, written text is enlarged, the voice assistant speaks at a slower rate, and warning messages are prompted sooner to account for longer reaction times.

2.4.2. Manufacturing

The smart production plant interface developed by Josifovska et al. (2019) extracts user information from a Digital Twin to dynamically adapt its navigation, layout, modality, or complexity of the content. Relevant characteristics stored in the Digital Twin – if explicitly entered by the user – include demographic data, communication preferences, as well as physical, cognitive, and sensory abilities. The authors outline exemplary adaptation rules which include, inter alia, switching from visual to vocal communication when interacting with a vision impaired user. Whenever a worker is standing at some distance from the display, the interface adapts the visual layout or switches to audio presentation if the worker is not within viewing range. The speed and volume of audio output, as well as the amount of presented information may be adapted to the worker's abilities which are inferred from their age and previous interactions.

2.4.3. Smart office

Adaptation in *enterprise systems* aims to reduce the complexity of tasks and offer proactive support. Adam et al. (2017) developed a framework with generic guidelines for stress-sensitive adaptive enterprise systems (SSAES). Stress may be measured through any unobtrusive sensor. While the authors define no concrete adaptation rules, they propose to implement measures including email filtering, help dialogues, and decision support elements when the system senses that the user is stressed.
In Gaze-X (Maat & Pantic, 2007), adaptive support is offered when the user pauses an activity or looks at an element for a long time. For each event, multiple adaptation options are defined. For example, when the user tries to open a file within an application, the system may either highlight recently opened file names, or launch a desktop search application. When attempting to edit a table, it can direct the user towards a help tab or table-related menu options. The user's preferences for each option are dynamically updated based on explicit feedback or reactions – inferred from facial expressions – to an adaptation.

To improve communication in geographically distributed work groups, Vertegaal et al. (2003) present a *virtual meeting* tool that attenuates the audio output from attendants that the user is not currently looking at. To enhance visual communication, an eye tracker monitors the direction of the user's gaze and the video stream is transmitted from the camera – selected from three available devices – that captures the user's face from a frontal view.

2.4.4. Smart home

GUIDE (Coelho et al., 2011) is a multimodal interface that aims to support elderly people in tasks such as controlling their television and home appliances, participating in video conferences and telelearning, as well as media and social interactions. It maintains a user model of their physical and cognitive abilities, knowledge, preferences, affective states, and attention. The user model is initialized with data that the users enter when first accessing the system, and is gradually refined through interactions with the application. GUIDE adapts the layout and content, and always presents information in the user's preferred modality. Examples include increasing the volume in response to a disoriented user, filtering

menu options and increasing their size if the user is inactive for a long time, or deactivating help messages for experienced users.
The smart home application of the embodied agent SmartKom (Reithinger et al., 2003) distinguishes between two modi. If the user is paying attention to or interacting with the display, detailed information is projected onto the screen and a voice assistant only summarizes the details of the user's requests. Whenever the user turns away from the screen, SmartKom retrieves only the most relevant information, presented exclusively through speech.

2.4.5. Education and learning

Adaptive educational hypermedia systems personalize instruction strategies by adapting content selection, navigation support, and presentation to the student (Apoki et al., 2020; Brusilovsky, 2003; Ennouamani & Mahani, 2017). Content selection identifies learning objects that are relevant to the student's learning goals and are compatible with their knowledge and background. Navigation support shows or – depending on the knowledge level – hides links to additional learning content and arranges them in an order that guides the user to the most relevant information. Adaptive presentation can manipulate the layout or level of detail of the learning material, or choose one of several alternative presentation formats (e.g., text, audio, or video). In addition to learning goals and knowledge levels, the user model typically maintains information about the student's content, media, and layout preferences, cognitive abilities, learning style, or their emotional and physical state (Apoki et al., 2020; Ennouamani & Mahani, 2017).
Intelligent tutoring systems typically respond to the learners' performance, knowledge level, and navigation behavior (Mousavinasab et al., 2021). Less frequently, variables including learning style, preferences, emotions, cognitive factors, or culture are used to characterize learners. The focus of the adaptation lies on individualized task selection, feedback and explanations, help dialogues – often provided in a scaffolded manner – and learning path navigation. Most of the time, the information is entered explicitly by the student or inferred from completed learning assignments. One notable exception is AdELE (Garcia Barrios et al., 2004), a learning platform that collects gaze data to provide adaptive functionalities. Apart from tracking the students' attention to the learning material, gaze patterns give insights into levels of concentration, excitement, and fatigue.

The authors envision the information to be used, for instance, for providing additional explanations if understanding difficulties are detected, or suggesting better learning strategies. However, the implementation of the system has not been completed.

Duarte and Carriço (2006) apply a generic adaptation framework to a *digital book* with adaptive annotations. The framework defines adaptation rules as matrices that map all possible combinations of input modality (voice/ pointing), output modality (voice/ graphical), content (e.g., complexity of instructions), and layout. One adaptation rule stops the display of alerts that hint at supporting material if the user initially ignores them. Conversely, if the user frequently consults the indicated content, the interface starts presenting supporting material without previously issuing alerts. Another rule specifies that the output modality of annotations shall depend on the user's preference for their immediate or delayed display. Immediate display is provided as visual output, whereas delayed annotations are announced by an audio-visual signal. However, the authors emphasize that the adaptation rules are only exemplary, and that it is the task of the designers to specify the adaptation matrix for their application.

2.4.6. Recommender systems

Recommender systems assist users in situations where their knowledge of the available alternatives is insufficient for making a decision (Resnick & Varian, 1997). An example of an adaptive solution that seeks to improve *web browsing* is described by Mobasher et al. (2000). Through embedded links, users can access other websites that, given their browsing history, might interest them.

Gaze-based *search engines* identify images (Cheng & Liu, 2012; Kozma et al., 2009) or documents (Hardoon et al., 2007; Salojärvi et al., 2004) that users dwell on as relevant and retrieve items with similar attributes. Xu et al. (2008) recommend videos with content that is related to the key frames of previously watched videos that particularly attracted the user's attention.

Online retailers such as Amazon recommend products related to items that the customer has placed in the shopping cart (Linden et al., 2003). Using a similar strategy, Kim et al. (2001) propose a solution for personalized advertisements in online shops based on past transactions and demographic data.

Smart *tourist guides* provide assistance for trip planning. For example, iTourist (Qvarfordt & Zhai, 2005) displays supplementary audio descriptions whenever the user dwells on a landmark. Duarte and Carriço (2006) implement a tourist guide that responds to voice or gesture input by, for instance, switching from visual to audio presentation, and from simple to detailed instructions. The authors specify presentation modality, instruction complexity, synthesized speech type, and layout as possible adaptation parameters, but define no concrete adaptation rules.

2.4.7. Cinematography

The groundwork for *adaptive cinematography* was laid by Bolt (1981) with a setup of multiple video streams projected simultaneously onto a large screen. When looking at a stream for several seconds, the window is enlarged and its soundtrack activated. Starker and Bolt (1990) describe an attentive graphics world that narrates the story of Saint-Exupéry's 'The Little Prince'. If the user looks at the general scene, a speech-enabled avatar of the little prince tells a general story about his planet. Whenever the user glances at a specific object for a prolonged time, the prince starts explaining details about the object.
The vision of full-fledged self-adaptive movies described by Hansen et al. (1995) was later implemented by Vesterby et al. (2005). Starting with a scene of two people who leave a room in opposite directions, the person to whom the viewer pays more attention is shown in the next scene.
In contrast to these subliminal adaptations, interactive Netflix movies pause at decisive scenes and let the viewer decide how the plot continues (Netflix, 2018, 2021a, 2021b).

Pursuing a different objective, *self-reflective movies* metaphorize past experiences of the user. Peng et al. (2018) narrate the adventures of a dog traveling to the moon. During the week leading up to the presentation of the movie, the users submit daily reports of their mood and behavior to an online questionnaire. In the movie, the obstacles and social interactions that the dog encounters along the journey vary with the user's reported experiences of the past week. The user's mood is reflected by the dog's behavior and expressions, as well as the visual and acoustic design of the environment.

2.5. Summary of theoretical foundations & implications for further research

The review of related literature identified a respectable number of adaptive applications, demonstrating their potential in a variety of domains. However, most of the solutions lack both empirical and theoretical support. An experimental investigation of how the adaptation affects usability has only been reported for a few applications from the automotive domain (Tchankue et al., 2011), search engines (Cheng & Liu, 2012; Kim et al., 2001; Qvarfordt & Zhai, 2005), smart office solutions (Maat & Pantic, 2007), and cinematography (Peng et al., 2018).

To fill this gap, each of the three essays proposes and evaluates a rule for user-centric adaptation of multimodal interfaces. The formulation of the adaptation rules is rooted in empirically supported (and at times self-contradictory) cognitive theories. Figure 2.3 illustrates how each theory-driven adaptation rule determines the optimal presentation of content to achieve an adaptation goal. The adaptation is conditional on a user state, which is inferred from sensor input.

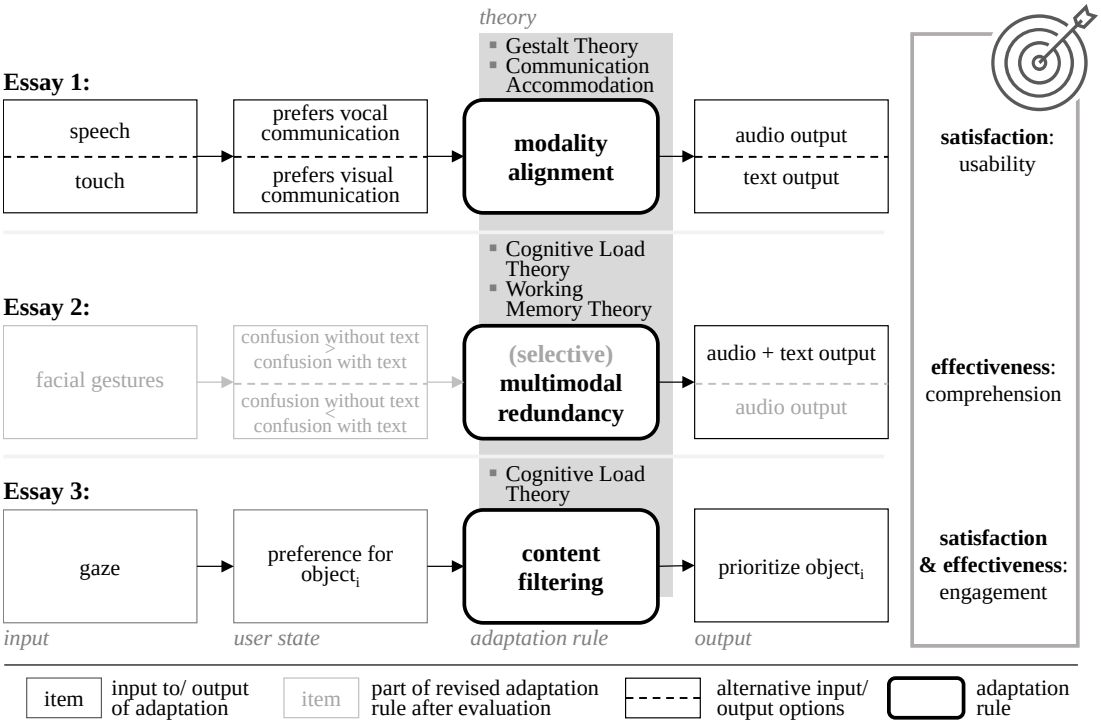


FIGURE 2.3: Logical connection between adaptation input and output. Theory-driven adaptation rules determine how output shall be presented given the user's state – which is inferred from user input – to achieve an adaptation goal.

Essay 1 tests whether modality alignment (Rule 1) – grounded in *Gestalt Theory* and *Communication Accommodation Theory* – increases satisfaction by providing higher usability to the user. If a sensor receives spoken (versus manual) input, the user's preferred modality is predicted to be auditory (versus visual-manual). Consequently, auditory (versus written) output is created.

Essay 2 applies non-adaptive multimodal redundancy (Rule 2), which receives support from *Working Memory Theory* and the modality principle of *Cognitive Load Theory*. Since the rule does not meet its effectiveness goal of improved comprehension, a revised version is proposed. With a stronger focus on the redundancy principle of *Cognitive Load Theory*, selective multimodal redundancy (Rule 2-2) captions spoken discourse only if the user's facial expression reveals that the multimodal presentation indeed reduces confusion.

Essay 3 tests whether content filtering (Rule 3) increases satisfaction by reducing intrinsic load, as implied by *Cognitive Load Theory*. Preferred content in videos is inferred from gaze and prioritized as the plot progresses.

3. Essay 1: Does Using Voice Authentication in Multimodal Systems Correlate With Increased Speech Interaction During Non-critical Routine Tasks?

The widespread availability of voice control with software components such as Siri (Apple, 2021), Cortana (Microsoft, 2021), or Amazon Alexa (Amazon, 2021) have introduced another convenient interaction method, complementing the traditional manual input on personal devices. With the redundant implementation of all available functionalities in different modalities (Oviatt, 2003a), users can easily switch between multiple interaction options. Multimodal interfaces thus allow users to interact with their devices in a modality that suits the immediate context. Given the dynamics of most people's busy lifestyles, a context may change almost instantly (Oviatt et al., 2000). Yet, despite a long-standing history of research on multimodal interfaces (Dumas et al., 2009; Jaimes & Sebe, 2007), their design is still a challenging task, as each user's interaction with such a system is different (Oviatt et al., 2003). While it is possible to activate multiple communication channels simultaneously, this may not be desirable. For instance, audio output can be disturbing in public spaces, or even pose a threat to security when the device reads out loud sensitive information (Cowan et al., 2017). In an attempt to overcome these issues, adaptive multimodal interfaces use cues from the user's chosen input modality to infer the most appropriate output channel in a given situation (Coelho & Duarte, 2011; Maat & Pantic, 2007).

In most applications, the user's first point of interaction is authentication. However, given its sensitivity to security issues, users may log in using a different modality than they prefer for non-critical routine tasks. In particular, most people have reservations about speaking a passphrase out of fear of being overheard (Trewin et al., 2012). We therefore investigate whether the input for authentication differs from interaction preferences during non-critical tasks. We study this question in

the context of a smart home application that can be controlled using touch or speech input. In a user study with 41 participants, we evaluate whether the input modality that is used during login is continued to be used for the remainder of the interaction, and whether users who choose voice authentication find speech input more usable. We further test whether it is beneficial to the user if system output is displayed in a compatible modality (cf. Schaeffner et al., 2016). Thus, depending on the input, instructions and system responses are presented as written or auditory output.

3.1. Related work: Multimodal integration patterns

Previous work that attempts to understand user behavior when interacting with multimodal systems has mainly focused on studying under which contexts users prefer multimodal over unimodal input. A consensus exists that the usability of multimodal interaction is primarily driven by the activity (Morris, 2012; Oviatt, 1997; Oviatt et al., 2004; Williams et al., 2020). However, the views on how task complexity relates to the suitability of multimodal interaction are conflicting. Observations from experimental investigations suggest that users prefer unimodal touch or speech input when executing simple tasks, but revert to multimodal input as the tasks become more cognitively complex (Oviatt, 1997; Oviatt et al., 2004). Consistent with these findings, Morris (2012) showed that unimodal interaction is preferred for issuing simple commands to a speech and gesture controlled television. When interacting with the considerably more complex image editing application PixelTone, in contrast, the users preferred multimodal interaction over using only speech or touch input (Laput et al., 2013). Conversely, elderly users have shown a preference towards multimodal input for controlling a simple television application (Coelho et al., 2011) and a recent study by Williams et al. (2020) suggests that a combination of speech and gestures is more strenuous for complex tasks. Apart from external task related factors, user specific variables influence the usability of multimodal interaction (Bubalo et al., 2016; Oviatt et al., 2003). The more experienced the users are, the more often they use multiple input modalities (Bubalo et al., 2016). Once a user has developed a preference for unimodal or multimodal interaction, their integration patterns remain consistent (Oviatt et al., 2003). When asked explicitly to use both speech and manual input, the modalities

are used either simultaneously (multimodal) or sequentially (unimodal). Attempts to change the integration strategy (e.g., by introducing frequent errors) only strengthened the previous strategy.

Input preferences. When given the choice between multiple interaction modalities, manual input has been found to be preferred over speech in human-robot (Profanter et al., 2015) and in human-computer interaction including applications for computer-based drawing (Alibay et al., 2017) and the documentation of electronic health records in hospitals (Seligman & Dillinger, 2006). Empirical evidence suggests that the observed preference persists even when speech is more efficient and effective. Users of an interactive voice response system for reporting public safety issues judged speech to be more efficient, but most often opted for keyboard input when given the choice (Breetzke & Flowerday, 2016). In a study with a smart television (Ibrahim et al., 2001), the traditional remote control was preferred over speech interaction. Gestures were the modality of choice for selecting virtual objects in an augmented reality (AR) application (Lee et al., 2013). While this may be attributed to the higher cognitive load of speech input (Schaffer et al., 2011), speech can be preferred if the benefits in terms of efficiency and effectiveness are sufficiently high. Schaffer et al. (2015) report that users of a restaurant booking application for mobile phones preferred speech when it led to less interaction steps and had a low recognition error rate. In a study investigating text input on smartphones, Smith and Chaparro (2015) report the lowest error rates for speech input and physical keyboards. Both modalities were also ascribed a high usability by the study participants. In an interactive map application, the users chose manual input for simple location specifications, but preferred speech for more extensive and not clearly defined object descriptions (Oviatt, 1997). Similarly, interaction with the meeting documentation system Archivus suggest that the mouse is used for simple navigation tasks, whereas speech is preferred for free text entry (Melichar & Cenek, 2006). Since efficiency and effectiveness depend on individual factors like demographics and prior knowledge, modality preferences can differ from person to person (Jokinen & Hurtig, 2006). Dynamic context variables such as the user's environment and cognitive load influence the effectiveness of a modality (Morris, 2012). Ideally, multimodal interfaces should therefore adapt to the individual user and current situation.

Output preferences. Preferences for output modalities depend on the interaction context (Coelho et al., 2011). Adaptive multimodal interfaces therefore use input from keyboard and mouse, cameras, and microphones to dynamically infer relevant context variables, including the user's affective state, attention, and task (Coelho & Duarte, 2011; Maat & Pantic, 2007). Output is then presented in the modality that is most appropriate for the situation. While increasing usability for inexperienced users, output adaptation has not been found to benefit domain experts (Maat & Pantic, 2007).

Our short glimpse into the research landscape evidences the importance of task complexity and contextual factors for determining the suitability of a communication modality. In contrast, the effect of individual preferences on the interaction behavior during different tasks is a topic that has not been systematically researched. Specifically, little is known about the link between the interaction modality that a user chooses for authentication and subsequent inputs for non-critical routine tasks. We therefore analyze whether the behavior during login is indicative of a user's preferred modality for non-critical routine tasks.

3.2. The smart home application

We developed a simple prototype for a desktop application of a smart home display. As telework became the norm during the COVID-19 pandemic, about 75% of office workers now wish to at least partly work from home (Poulton, 2020). Assuming that the users prefer to control their home appliances through the same device that they are already using during office hours, we conducted the study with a desktop application. The application can be operated using touch or speech input. Adaptive functionalities were integrated following the FAME development guidelines for adaptive multimodal applications (Duarte & Carriço, 2006):

1. **Identify adaptation variables.** *Which variables introduce variations from outside of the system?* Users can interact with the system through mouse input or speech. Following the taxonomy of sensory modalities by Turk (2014), we henceforth use the term 'touch input' when referring to mouse clicks. The available input modalities were chosen to maximize usability and user confidence. Touch input is still the default control mechanism

on consumer devices (Hoffmann et al., 2019). However, speech-controlled digital assistants such as Alexa (Amazon, 2021) or Siri (Apple, 2021) have penetrated the market, so that most users are now comfortable using speech commands. In a study investigating the usability of different input modalities for smart home appliances, users were most experienced with touch and speech input, and found these modalities the easiest and most enjoyable to use (Hoffmann et al., 2019). The system adapts exclusively to the input modality that the user chooses for authentication. Environmental variables such as background noise do not influence the adaptation.

2. **Identify adaptable variables.** *What variable system components should respond to outside variations?* Instructions and system responses are presented either as text, or read aloud by the text-to-speech (TTS) engine.

3. **Select model attributes.** *What information requirements should be stored in models?* A user model registers whether speech or touch input was used during the login task. The interface continuously listens for speech input throughout the entire interaction. However, the model is static, i.e., it is initialized when the user logs into the system and is not updated if the user subsequently communicates in a different modality.

4. **Design interaction model templates.** *How is information and the relationships between components represented?* The speech-to-text (STT) engine informs the TTS module and the GUI about the modality through with authentication was completed.

5. **Define adaptation rules.** *What rules and methods define the adaptation?* If speech (touch) is used for login, the TTS module is activated (disabled), and all text instructions on the GUI are disabled (activated).

The system was implemented in Python. The GUI was realized with `Tkinter` which provides convenient functionalities for rapid prototyping (Python Software Foundation, 2021). TTS conversion was realized with Window's native `Microsoft Speech API (SAPI 5.3)` (Microsoft, 2012). It was accessed through the `pyttsx3` Python library. The online Python library `SpeechRecognition` establishes an interface to the `Google Speech Recognition` engine. It has been found to deliver superior recognition performance compared to other state-of-the-art speech

recognition engines (Këpuska & Bohouta, 2017). The speech recognition engine runs in the background and continuously listens for speech input.

3.2.1. Authentication procedure

When starting the application, the experimental instructions are displayed on the screen and simultaneously read aloud by the system. After 20 seconds, the participants are automatically routed to the login page. Authentication is a critical security mechanism for smart homes to protect services such as paid television channels from unauthorized access (Prange et al., 2021). It is of particular importance for voice assistants due to the often sensitive nature of their services and is typically the first point of interaction with the system. Unlike personal computing devices, smart home applications cannot delegate the authentication task to password management systems that automatically fill in the password, because they would grant access to any person inside the house.

We implemented two alternative authentication tasks that are common for consumer devices. Both tasks use a secret (i.e., user-specific knowledge) for authentication, independent of the chosen input modality. It should be noted that this text-dependent authentication method differs from biometric voice authentication, where the user's identity is verified based on vocal parameters (Sae-Bae et al., 2019). We chose this approach so that authentication is based on the same type of credential, independent of whether touch or speech is used. In the study, one authentication task was selected by the system at random. This allowed us to investigate whether the choice of the authentication modality was biased by the task. *Login 1* uses identical tasks for touch and speech input. The user enters the PIN '5678', either on a virtual number pad or by saying the number sequence out loud. *Login 2* uses a different task for each input modality. If choosing touch, the user draws a simple pattern onto the display. This is similar to the widely used authentication on Android smartphones (Uellenbeck et al., 2013), and thus provides both familiarity and high usability. Alternatively, the user can speak the sentence "I wish to enter". In both tasks, the cognitive effort of touch is comparable to speech input (cf. Chapter 3.4.1). We can therefore assume that the choice of the input modality is not influenced by the amount of intrinsic cognitive load (i.e., the mental effort induced by the task itself).

By using two authentication tasks (*Login 1* versus *Login 2*), we account for a legacy bias that may be associated with the task. PIN authentication traditionally requires touch input, since speech can induce a security breach in public spaces. Users might therefore be more inclined to use touch for PIN authentication on the smart home display, even though in the safety of their own home, spoken input does not reveal the secret PIN to unauthorized individuals.

Since numerical digits have an empirically low recognition rate of about 75% (Baig & Kavakli, 2018), we anticipated performance issues for spoken PIN authentication. In order to not discourage participants from using speech and thus bias their modality selection, we applied a Wizard of Oz experimental design in which authentication verification was omitted. We did not verify whether the spoken PIN entry was correct. In contrast, speech commands after authentication needed to be recognized correctly in order for the system to respond.

3.2.2. Smart home functionalities

In our user study, we aimed to obtain a balanced sample of participants using each of the available input modalities for the routine task. At the same time, we did not want to bias their choice by using a task that could objectively be accomplished more conveniently with one modality. The chosen form of interaction should be entirely the result of a personal preference. Thus, we created a scenario in which both modalities were equally convenient to use. Given that, in the telework scenario, the users are already seated in front of their computer, the mouse lies within easy reach. Speech interaction is typically only preferred under specific circumstances (Coelho et al., 2011; Melichar & Cenek, 2006; Morris, 2012; Oviatt, 1997). In particular, it has been proven to be especially useful when executing simple tasks that require only minimal input (Aldridge & Lansdown, 1999; Luger & Sellen, 2016). The system thus provides control panels for three smart home appliances that can be navigated with short and simple commands: By selecting 'Weather Forecast', the user can request a detailed weather report for one of the seven days of the week, which is then retrieved with the Python library `Pyowm`. The 'Air Conditioning' functionality allows the user to increase or decrease the room temperature by pressing the respective button, or by saying "up" or "down". In the 'Light' control panel, the lights of three rooms can be switched on or off.

All functionalities can be executed with touch or speech commands, with the latter corresponding to the button labels. Instructions and system responses to user actions are provided as audio or text output, depending on the modality through which authentication was completed.

3.3. Study design: Modality alignment

The objective of the experimental investigation was to test whether users who favor voice over touch authentication on a multimodal interface also find speech more useful during non-critical routine tasks:

> **H 1A.** Voice authenticators are more inclined to use speech input in non-critical tasks.

The hypothesis is based on the empirical finding that ease of use is the most important criterion when selecting the modality for authentication, even more so than confidence and privacy (Toledano et al., 2006). Since this is also the case for non-critical tasks (Connell & Lynott, 2011), **H 1A** assumes that the same modality is preferred for authentication and non-critical tasks.

Modality switches induce additional cognitive load (Connell & Lynott, 2011; Sandhu & Dyson, 2012). Buisine and Martin (2003a) therefore suggest that system output must be presented in audio format if the user of a multimodal system chooses speech input (*symmetry principle*). Symmetric multimodality has been applied to multimodal systems (Wahlster, 2003), but there is still a lack of evidence on how it affects usability. We thus investigated whether those who communicate through speech prefer audio output over written text. We formulate the following hypothesis:

> **H 1B.** Users prefer system output in the same modality they use for input commands.

To test the hypotheses, a between-subjects experimental design with one treatment group and one control group was adopted. Participants could authenticate

themselves using either touch or speech. In the treatment group, system output was presented in a modality that is compatible with the input from the login task (Schaeffner et al., 2016). Thus, participants who authenticated themselves by touch received exclusively visual output. For those who had logged into the system using speech, the TTS engine converted all text elements into audio output and written text was removed. For example, if a user logged in using speech input, the weather forecast was presented exclusively as a spoken report. In the control group, in contrast, system output was presented in the opposite – and thus incompatible – modality of authentication (Schaeffner et al., 2016).

3.3.1. Apparatus

Participants were seated in front of an HP laptop with a 1.6 GHz i5-8250U processor and 16 GB RAM. The GUI was projected onto the laptop display (1920x1080 pixels). We ensured that a stable WiFi connection persisted throughout all experimental trials to prevent network induced performance issues of the speech recognition. In compliance with existing COVID-19 regulations, participants were wearing face masks while interacting with the smart home display. To compensate for muffled voice input, a Jabra Evolve 40 headset was connected to the laptop.

3.3.2. Participants

We recruited 41 campus residents and their personal contacts (23 female, mean age: 26.1 $\pm$ 3.4) for the experiment. Of the participants, 11 were enrolled in an undergraduate program at our university, and 27 were currently pursuing or had already completed a graduate degree. The participants were from 14 different nationalities (22 South or East Asian, 19 European), with none speaking English as their first language. No participant had any sensory impairment that could affect the usability of an interaction modality. Participation was voluntarily, and no monetary or equivalent incentive was given.

3.3.3. Procedure

Before starting the experiment, the subjects were informed that they were participating in a research study in which they would interact with a smart home

display. They were told that all mouse and speech input was recorded. After completing the experiment, they were given a detailed explanation of its purpose.

Before using the system, the participants were given the login credentials. Since the experiment consisted of a single session and no individual user accounts were created, the same PIN was used for all participants. The experimenter then left the room so that the participants would not be disturbed. Upon starting the application, the following instructions were displayed on the computer screen:

> *"Thank you for trying out Smart Home Display. You can log in using spoken commands or the mouse. After logging in, please feel free to browse through the menus and explore Smart Home Display."*

Once the participants had finished exploring the system, they were presented a questionnaire to rate their perceived workload for the authentication task and the usability of speech input and output.

3.3.4. Metrics

We evaluated the usability of the interaction along three constructs that measure the *workload of authentication*, *usability of speech input*, as well as *usability of text and audio output*. Table 3.1 lists the items that were used to collect subjective user ratings. All items were measured on a seven-point Likert scale.

Perceived *workload of authentication (**A**)* was used as a control variable to test whether the modality choice was biased by the intrinsic cognitive load of the task itself. The construct was measured with six items from the NASA Task Load Index (TLX) (Hart & Staveland, 1988).

*Usability of speech input (**U**)* was used to measure the users' attitude towards speech interaction. It was collected with *effort* and *performance expectancy* measures adapted from the UTAUT model on user acceptance of information technology (Venkatesh et al., 2003).

TABLE 3.1: Items used to measure usability of input and output modalities. The items measuring *perceived workload of authentication* were collected with the NASA TLX (Hart & Staveland, 1988). *Usability of speech input* as well as *usability of text and audio output* were measured with UTAUT items (Venkatesh et al., 2003). All items were measured on a seven-point Likert scale.

Item		Construct
A-1	How mentally demanding was the login?	**Workload of authentication**
A-2	How physically demanding was the login?	
A-3	How hurried or rushed was the pace of the login?	
A-4	How successful were you in accomplishing the login?	
A-5	How hard did you have to work (mentally and physically) to accomplish the login?	
A-6	How insecure, discouraged, irritated, stressed and annoyed did you feel during the login?	
U-1	Using speech interaction enables me to accomplish tasks more quickly than with traditional mouse interaction.	**Ease of use of speech input**
U-2	Using speech makes it easier to interact with the smart home application than using the mouse.	
U-3	Using speech to interact with the system is cumbersome.	
U-4	My interaction with the smart home application is clear and understandable.	**Confidence with speech input**
U-5	Using speech gives me greater control over the system than using mouse-based inputs.	
U-6	The smart home application responded to my speech input in a timely manner.	**Perceived success with speech input**
U-7	I wished that the system better recognized my speech.	
O-1	I find it useful to receive spoken instructions.	**Audio usability**
O-2	Receiving spoken instructions enables me to accomplish tasks more quickly than with text instructions.	
O-3	Spoken instructions make using the system more interesting.	
O-4	It scares me to think that I could miss important information if the instructions were only displayed as spoken messages.	
O-5	I would have wished to receive more spoken audio instructions.	**Audio deficiency**
O-6	I would have wished to receive more text instructions.	**Text deficiency**

Additionally, the logged interaction behavior served as an objective measure of the users' attitude towards speech input. All user input was logged along with the corresponding timestamp. For speech input, additional parameters were logged in order to assess the quality of the speech recognition. More precisely, we recorded all predicted speech alternatives and their confidence values. Confidence scores (ranging from 0 to 1) indicate how reliable the STT conversion is (Jiang, 2005). From the raw log data, we extracted three metrics related to the users' *modality choice*, and three additional indicators for the *speech recognition quality*. The metrics and their calculation are summarized in Table 3.2.

TABLE 3.2: Interaction metrics extracted from log data. The metrics were used as objective measures to evaluate the usability of speech and touch input.

Metric	Definition	Construct
clickCount	Total number of clicks during the interaction	**Modality choice**
speechCount	Total number of recognized speech inputs from the entire interaction	
speechRatio	Ratio of speech input to total input: $\frac{\text{speechCount}}{\text{speechCount} + \text{clickCount}}$	
modalitySwitches	Number of times the user changed from speech to touch, or from touch to speech input	
speechConfidence	Mean speech recognition confidence for a user, calculated from the confidence value of the most likely speech alternative	**Speech recognition quality**
executableCommands	Number of speech inputs that resulted in the successful execution of the associated command (excluding input where some spoken input was recognized, but could not be associated with a command, either because a wrong keyword was used, or because the words were not recognized correctly)	
commandRatio	Ratio of executable to total speech input: $\frac{\text{executableCommandCount}}{\text{speechCount}}$	

Subjective preferences for output modalities were measured with six additional UTAUT items. The items were adapted to assess the *usability of text and audio output (**O**)* of the smart home display.

3.4. Results: Usability of modality alignment

Participants spent on average 3.61 minutes (σ = 2.72 minutes) exploring the smart home application. After excluding one sample due to missing data in the log file, we retained 40 valid data records.

The participants' modality choice for authentication was fairly evenly distributed. 16 subjects used speech to authenticate themselves, and 24 subjects opted for touch input (cf. Table 3.3).

TABLE 3.3: Descriptive statistics of authentication modalities. Depicted is the number of subjects who chose the specified input modality for the specified authentication task.

Authentication task	Touch input	Speech input	TOTAL
Login 1 (PIN)	17	4	**21**
Login 2 (phrase/ pattern)	7	12	**19**
TOTAL	**24**	**16**	**40**

3.4.1. Task validation

We first assessed whether the modality choice was influenced by dissimilar effort for authentication with speech versus touch. We analyzed whether participants logging into the system with speech perceived the cognitive workload for the task different from those using touch. A two-sided t-test showed that cognitive workload does not significantly differ between the two input modalities (p-value = .351). An additional analysis in which we assessed each authentication task separately did not reveal a significant modality effect either, neither for *Login 1* (p-value = .481) nor for *Login 2* (p-value = .187). The tasks can therefore be assumed to evoke similar cognitive workload, independent of whether they are executed using speech or touch. We concluded that the authentication procedures that were used in the study did not bias the participants' selection of an input modality.

3.4.2. H 1A: Voice authenticators are more inclined to use speech input in non-critical tasks

In our quest to answer **H 1A**, we investigated the link between the participants' selection of a login modality and their attitude towards voice input as a control mechanism for the smart home display. We tested the relationship for each of the two constructs measuring the usability of voice input. Responses from the follow-up questionnaire were used as a subjective measure of the users' attitude towards speech input. Additionally, we used the logged interaction behavior as an objective measure.

Correlation with perceived usability. To determine whether the modality that a user chooses for authentication is related to the perceived usability of speech input during later interactions, we summarized the usability metrics of speech input into three constructs representing *Ease of Use* (**U-1**, **U-2**, **U-3**), *Confidence*

(**U-4**, **U-5**), and *Perceived Success* (**U-6**, **U-7**) of speech input. Table 3.4 shows that speech authenticators assign slightly higher usability scores to *Ease of Use* and *Confidence*, while evaluating *Perceived Success* lower than touch authenticators. However, two-sided t-tests revealed that the effect is not significant for any of the usability constructs. Two-sided t-test applied to each of the individual usability metrics separately reported no significant effect either. Therefore, we conclude that speech authenticators do not ascribe a higher usability to speech input than touch authenticators.

TABLE 3.4: Subjective usability ratings of speech input per chosen authentication modality. Significance of group differences is calculated with two-sided t-tests.

	Touch input		**Speech input**		**t-test**	
Speech usability construct	μ	(σ)	μ	(σ)	t	sig.
Ease of Use (U-1, U-2, U-3)	3.26	(1.58)	3.73	(1.07)	-1.083	.285
Confidence (U-4, U-5)	3.77	(.135)	4.06	(1.06)	-0.742	.463
Perceived Success (U-6, U-7)	2.29	(1.39)	2.06	(1.39)	0.468	.642

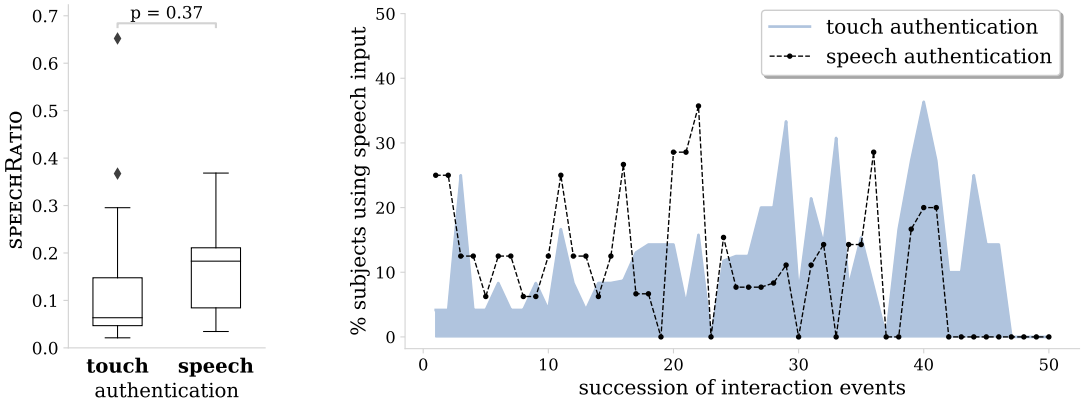
Correlation with interaction behavior during routine tasks. We analyzed whether the participants' selection of an input modality for authentication is representative for their later input behavior. Table 3.5 summarizes the average values of the interaction metrics for the two experimental groups (i.e., touch versus speech authentication).

TABLE 3.5: Usage of input modalities from interaction metrics per chosen authentication modality. Significance of group differences is calculated with two-sided t-tests.

	Touch input		**Speech input**		**t-test**	
Metric	μ	(σ)	μ	(σ)	t	sig.
clickCount	43.17	(44.08)	32.06	(16.08)	1.092	.283
speechCount	5.08	(5.13)	5.56	(2.83)	-0.370	.714
speechRatio	.1285	(.1426)	.1656	(.0911)	-0.979	.334
modalitySwitches	8.75	(5.76)	8.25	(4.02)	0.315	.755

The box plot in Figure 3.1a shows that the median speechRatio of voice authenticators is higher than for participants who used touch input to log into the system. Yet, a two-sided t-test shows that the use of an input modality after authentication does not significantly differ between the experimental groups (t = 0.900, p-value = .374). The line graph of the modality usage over time in Figure 3.1b reveals that for voice authenticators, the proportion of speech input

does not gradually decline. Rather, directly after authentication, the majority of the users switch to touch interaction. Touch input is the overall dominant input modality, irrespective of whether touch or speech is chosen for authentication. Yet, most users frequently revert to speech input multiple times throughout the interaction, with on average 8.6 modality switches.



(a) Proportional use of speech input

(b) Sequential development of the number of participants using speech and click input during the first 50 interactions.

FIGURE 3.1: Frequency of touch and speech input after authentication by experimental group. Significance of group differences was assessed with a two-sided t-test.

Interaction effects. Given that the complexity of the task remained constant throughout the entire interaction, there is no evidence for a causal relationship between the participants' frequent switching behavior and intrinsic (i.e., task induced) load. We therefore explored whether other factors related to the task caused the observed modality switches. Contrary to observations that have been reported in the literature (Aldridge & Lansdown, 1999; Bierschwale et al., 1989; Buisine & Martin, 2003b), the participants in our study used speech input more often for directional commands like "up"/"down" (56.6%) than for heavily semantic commands such as selecting the 'Weather Forecast' control panel. For touch input, we observed the reverse: 57.4% of all clicks were attributed to semantic commands. To understand the motivations behind the inconsistencies in the empirical evidence, we took a differentiated look at the contexts in which each type of control command was used. A preference for touch input has typically been related to directional commands for controlling continuous functions (Bierschwale et al., 1989), whereas the directional commands of our smart home display are used for singular adjustments (e.g., switching on the light). Observations that speech

input is preferred for semantic commands are based on input that is considerably longer (Aldridge & Lansdown, 1999) or not clearly defined (Buisine & Martin, 2003b). In contrast, we used very short commands throughout the experiment.

With the experimental setup of our user study and the configurations of the smart home display, we did not find clear modality preferences for a specific command type that would explain the switching behavior. We therefore verified whether the quality of the speech recognition had an effect on the selection of an input modality. Studies have shown that users tend to choose the modality that they expect to be the least prone to errors and subsequently switch to another modality when an error occurs (Oviatt et al., 1998). The logged speech input data shows that the speech recognition quality is highly dispersed across the participants, with a COMMANDRATIO ranging from 0.2 to 1.0 executable speech input commands (mean = .687, stdv. = .265). The COMMANDRATIO indicates the number of speech inputs that resulted in the successful execution of the associated command in relation to the total number of registered speech inputs. Its observed mean value translates into an average recognition error of 31.3%. The SPEECHCONFIDENCE ranges from 0.538 to 0.933 (mean = .847, stdv. = .070). However, Pearson's correlation coefficient provides no evidence that a higher speech recognition quality increases the use of speech input. As can be seen in Figure 3.2, a higher ratio of successfully executed speech commands is even negatively correlated with SPEECHRATIO (corr. = -.11, p-value = .52).

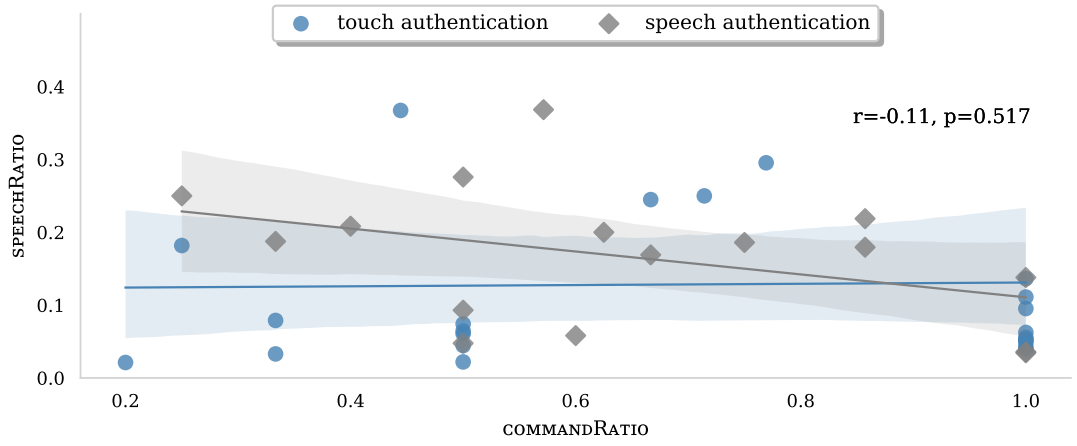
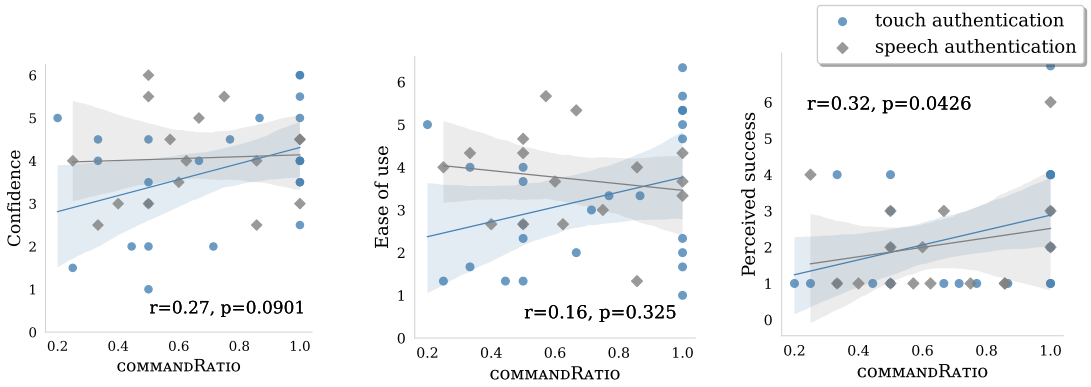


FIGURE 3.2: Correlation between the number of speech commands and their successful execution. Significance is calculated for the combined data from both experimental groups. No positive relationship exists between the ratio of successfully executed commands and the use of speech input.

However, the distribution of the usability ratings in Figure 3.3 demonstrates that perceived usability correlates with the quality of the speech recognition. Both *Confidence* (corr. = .27, p-value = .09) and *Perceived Success* (corr. = .32, p-value = .04) are positively correlated with the number of successfully executed speech commands, measured by the COMMANDRATIO. The relationship is significant at 95% confidence. It thus appears that, while speech recognition quality does not influence the users' actual use of speech input, it does have an effect on their perceived usability of speech as an input modality.



(a) Confidence **(b)** Ease of Use **(c)** Perceived Success

FIGURE 3.3: Correlation between speech input usability ratings and recognition quality. Speech recognition quality is measured by COMMANDRATIO, i.e., the proportion of speech commands that resulted in the correct execution of an event.

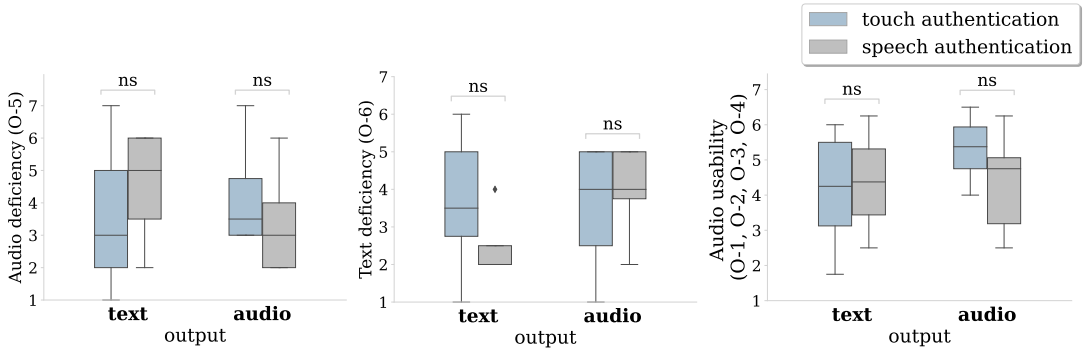
Based on the evidence from the statistical analyses, we reject the hypothesis that the chosen mode of authentication reveals a stronger inclination to use speech input in non-critical tasks (**H 1A**). Instead, attitudes towards voice control are formed gradually as the users perform the routine task and are mainly driven by the quality of the speech recognition.

3.4.3. H 1B: Users prefer system output in the same sensory modality they use for input commands

The previous analyses reveal that users do not necessarily continue using the communication channel through which they authenticate themselves. We therefore first test **H 1B** for the chosen authentication modality, and then repeat the analysis for the dominant input modality during subsequent non-critical tasks.

Correlation of authentication input with usability of speech output. We test whether participants of the treatment group (i.e., system outputs match the authentication modality) evaluate the usability of the smart home display different than participants of the control group (i.e., system outputs do not match the authentication modality).

The box plots in Figure 3.4 show how strongly the participants felt that output in one modality was missing, i.e., to what degree they experienced an *Audio deficiency* (**O-5**) or *Text deficiency* (**O-6**). Additionally, their ratings for *Audio usability* (**O-1**, **O-2**, **O-3**, **O-4**) are depicted.



(a) Audio deficiency **(b)** Text deficiency **(c)** Audio usability

FIGURE 3.4: Usability ratings for text and audio output by authentication modality. Significance was tested with a two-sided t-test (significance level: *p < .1%, **p < .05, ***p < .01).

The distributions in Figures 3.4a and 3.4b demonstrate that voice authenticators feel more strongly than touch authenticators that they should have received more audio output when presented exclusively with text, and wish they had received less text. Touch authenticators are generally satisfied with the output they receive, independent of whether it is presented in text or audio format. However, a two-sided t-test reveals that, for both **O-5** (audio output: t = 0.983, p-value = .381, text output: t = -1.058, p-value = .348) and **O-6** (audio output: t = -0.502, p-value = .642, text output: t = 1.224, p-value = .468), the difference between the experimental groups is not statistically significant.

Paradoxically, *Audio usability* was evaluated slightly higher by the control group, where the output did not match the chosen authentication modality (cf. Figure 3.4c). However, the effect is not significant (audio output: t = 1.444, p-value = .201, text output: t = -0.231, p-value = .829). A fine-grained analysis of

each usability measure (**O-1**, **O-2**, **O-3**, **O-4**) reveals that, even when examined in isolation, the difference between the experimental groups is not significant for any of the measures. Table 3.6 presents the complete descriptive statistics.

TABLE 3.6: Usability evaluation of audio output by authentication and activated output modality. Mean values (μ) were calculated from the subjective ratings of all participants. Standard deviations (σ) are given in parenthesis.

	Text output						**Audio output**					
	touch authentication (N = 20)		speech authentication (N = 4)		t-test		touch authentication (N = 4)		speech authentication (N = 12)		t-test	
Item	μ	(σ)	μ	(σ)	t	sig.	μ	(σ)	μ	(σ)	t	sig.
O-1	4.60	(1.91)	5.50	(1.66)	-0.855	.437	6.00	(0.00)	5.00	(1.78)	1.864	.089
O-2	3.75	(1.97)	4.25	(1.92)	-0.418	.697	5.25	(1.48)	4.50	(2.06)	0.710	.501
O-3	4.80	(1.97)	5.25	(1.48)	-0.569	.595	6.00	(1.00)	4.67	(1.93)	1.627	.135
O-4	4.35	(1.74)	5.50	(0.78)	-1.797	.112	4.00	(1.41)	4.58	(1.75)	-0.600	.571
O-5	3.40	(1.77)	4.50	(1.66)	-1.058	.348	4.25	(1.64)	3.25	(1.23)	0.983	.381
O-6	3.55	(1.60)	2.50	(0.87)	1.694	.135	3.50	(1.66)	4.00	(0.91)	-0.502	.642

Correlation of routine task input with usability of speech output. Given that the statistical tests provide no support for a link between the authentication modality and the frequency of speech interactions in subsequent non-critical tasks, we additionally investigated the role of subsequent interactions. Specifically, we tested whether users who more frequently use speech input throughout the entire interaction have a more positive perception of the usability of speech output. While a small positive correlation between the logged SPEECHRATIO and *Audio usability* exists (cf. Figure 3.5), it is not significant (corr. = .11, p-value = .52).

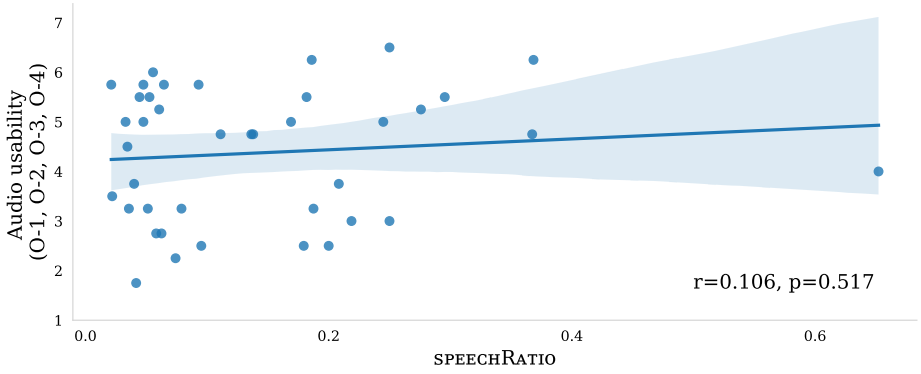


FIGURE 3.5: Correlation between *Audio usability* and speechRatio. The interaction metric SPEECHRATIO serves as an indicator for the frequency of speech interactions during non-critical tasks.

Based on the statistical evidence from the experimental investigation, we find no support for the hypothesis that users prefer system output in a modality that is compatible with their input (**H 1B**). This finding is consistent across all evaluated task types, and thus applies to security sensitive authentication as well as non-critical routine tasks.

3.5. Discussion & limitations of modality alignment

The data from the 40 participants of our experimental investigation who interacted with the smart home display reveals that users who authenticate themselves with speech do not necessarily perceive speech input and output as more usable. What is more, they do not use speech interaction more frequently during non-critical tasks than touch authenticators. Therefore, the experimental investigation provides no support for the hypothesis that the interaction behavior during login is representative for a user's inclination to use speech input in non-critical tasks (**H 1A**). The finding is robust to the quality of the speech recognition engine: A high number of recognition errors does not prevent the users from issuing spoken commands, although it does negatively impact the perception of their usability. Similar behaviors have been observed in a study by Rebman et al. (2001), where users of a meeting support application were dissatisfied with the speech recognition quality, but would still use the technology for future interactions. In contrast, Schaffer et al. (2015) report opposing results from a study in which the users of a restaurant booking system for mobile phones were less likely to choose speech over touch input if the speech recognition quality was poor.

We further found that users who more frequently favor speech input over touch do not evaluate speech output more positively than those who have a general preference for touch input (**H 1B**).

The study was conducted in a private space where security threats are minimized. Yet even in this protected environment, we did not find a correlation between voice authentication and the users' attitude towards voice-based interaction in non-critical tasks. While we anticipated concerns about speaking a PIN aloud, the participants' answers to A6 ("How insecure, discouraged, irritated, stressed and annoyed did you feel during the login task?") indicate that the participants who

used the PIN based speech authentication felt the most secure, even compared to touch authenticators. Thus, we observed no influence on the users' trust in the system. It is therefore safe to assume that the authentication modality will also not correlate with inputs for non-critical tasks in other contexts, including public spaces.

These findings have two important implications for the design of multimodal systems that adapt the output format to the user's interaction behavior. First, the sensory modality that the user chooses for the first few inputs does not necessarily match their preferred output modality. This is independent of whether the input is used for authentication or for the execution of non-critical routine tasks. Instead, preferences are formed gradually. We therefore anticipate that a static one-time adaptation of the output format to the user's input during the first few interactions would not be beneficial to the user. Thus, both output modalities should be provided until a clear preference is discernible. Second, multimodal interfaces should listen to all input channels throughout the interaction. Since we observed many modality switches, it would be detrimental to the usability if the system stopped listening to one input channel. Instead, input fusion techniques (Dumas et al., 2009) should be considered.

The study that we present in this paper concludes the exploratory phase of a long-term project. The insights we gained from this study will allow us to further improve the prototype application and remove technical barriers to using speech interaction. The current version of the application requires the users to say the exact words that are written on the actionable element. A detailed analysis of the log files showed that 63% of input recognition errors were caused not by poor performance of the speech recognition engine, but the use of an incorrect keyword. This is consistent with the findings from previous studies which suggest that some users prefer to activate a button by saying the command written on the element, while others refer to its order position on the screen (Coelho et al., 2011; Morris, 2012). In the next iteration, we will therefore allow for more flexible commands.

We will build on the findings from this study in order to conduct a field experiment with a large and heterogeneous sample. Observing interactions over a larger time span and in the users' natural environment will reduce experimental effects which might bias the users' choice of an interaction modality (Morris, 2012). This will allow us to see whether the input modality that dominates over a prolonged

usage allows to draw conclusions about the user's preferred output format in specific situations. If such a relationship exists, the output format could be dynamically adapted to the user's situational preferences. In addition to the chosen communication channel, situational preferences take into account the concomitant contextual factors such as ambient noise and the current location.

3.6. Conclusion: Summary of Essay 1

We conducted a study with 41 participants to assess whether the use of touch or speech input during authentication with a smart home application reveals the user's attitude towards speech interaction during non-critical routine tasks. We found that, even in the secure environment of a private home, users do not necessarily authenticate themselves with the modality they prefer for subsequent system inputs. The users' authentication behavior is therefore no reliable indicator for inputs during non-critical routine tasks. We further found that matching the output presentation (i.e., text versus audio) to the communication channel that was used to issue the first few commands (i.e., touch versus speech) does not increase the system usability. This finding is consistent for all task types, including both security sensitive authentication and non-critical routine tasks.

Building on these findings, we will extend the study in a long-term field experiment to observe interaction patterns over an extended period of time. We hope that this will allow us to conclude whether contextual input preferences from prolonged interactions can be used for dynamic adaptation of multimodal systems.

4. Essay 2: Evaluating the Potential of Caption Activation to Mitigate Confusion Inferred from Facial Gestures in Virtual Meetings

During the COVID-19 pandemic, virtual meetings have evolved into an indispensable tool for collaboration in industry, academia, and other parts of society. As a substitute for face-to-face meetings (Shockley et al., 2021), they should approximate the quality of physical encounters, especially in complex situations like contract negotiation or technical discussions. Yet, with the new technology, new challenges surfaced in the form of inefficient communication, low attention, and fatigue (Karl et al., 2022). Often, the result is incomprehension and confusion.

Online lectures suffer particularly from the loss of direct communication. In the classroom, teachers directly notice from the students' facial expressions and body response when they have difficulties following a lecture. In virtual lectures, the subtle reactions are easy to overlook (Chung et al., 2020) and students are often embarrassed to admit confusion (Kiyota, 2022). Innovative approaches like 'Mudslide' ask students to revise the lecture slides and mark those that they found confusing (Glassman et al., 2015). Needless to say, this relies on the students' willingness to provide feedback. Yang et al. (2015) infer confusion of online course students from forum discussions. While this requires no additional action, it only captures the feedback of learners who actively engage in the discussions.

Outside of learning, there is a striking lack of strategies to enhance mutual understanding in virtual meetings. Attendants of business meetings or conferences may suffer from the same comprehension issues as students, but in professional settings even more than in education, barely anyone is willing to admit them (Wilding, 2016). Especially when teams with diverse cultural backgrounds meet, incomprehension often remains undetected. For example, while Indians tend to say "yes" out of respect, Europeans will understand this as a sign of comprehension

(Klitmøller & Lauring, 2013). It is therefore imperative to address unexpressed confusion without exposing anyone to the judgment of others.

Research has demonstrated a beneficial effect of multimedia presentation, where auditory material is supported by text and graphic elements (Brett, 1997; Chun & Plass, 1997; Subaidi bin Abdul Samat & Aziz, 2020). One example are movie captions. What started as an assistive technology has evolved into a mainstream feature that is used by many to better understand dialogues (Jacobs, 1999).

In this work, we propose to make use of the functionality to improve comprehension in virtual meetings. Most video conferencing tools provide auto-generated captions. However, they are typically not activated by default, partly because they can distract from visual content (Jiménez et al., 2011; Winke et al., 2010). In movies, dynamic subtitle placement can mitigate these effects (Brown et al., 2015). However, determining a suitable position for captions in virtual meetings with a shared screen can be challenging or even impossible if the entire screen is covered with text. Moreover, auto-generated captions – unlike in movies – may not be perfectly accurate and timed (Dialpad, 2021). Therefore, while it has been firmly established that captioned movies improve comprehension (Zheng et al., 2022), the same may not hold true for virtual meetings. Consequently, displaying captions at all times may not be desirable, especially if they occlude important content or the speaker's face (Oviatt, 2006; Sueyoshi & Hardison, 2005). Adaptive interfaces offer an alternative solution which entails dynamically activating captions when needed. Thus, users are not distracted by captions unnecessarily, but also do not have to search for a functionality they might not even be aware of or consider as a solution. However, this requires a mechanism that is able to identify situations in which captions can be beneficial.

To research how captions can be used effectively in virtual meetings, we conducted a user study with two objectives in mind: First, to test whether auto-generated captions can mitigate confusion; second, to understand whether it is possible to detect confusion in virtual meetings without prompting for active feedback. Since the tools already have camera access if the required permission is granted, we tested if facial gestures extracted from video frames contain hints of confusion.

The contributions of this work are as follows:

1. We provide empirical evidence for the usefulness of auto-generated captions as a tool to mitigate confusion.
2. Facial gestures are identified that reveal confusion in virtual meetings, both during purely auditory presentation and when the visual component of a caption is added.

4.1. Related work: Audio-visual presentation & confusion

Confusion is the result of a cognitive disequilibrium caused by inconsistencies or missing knowledge that is required to understand presented information (Arguel et al., 2017). In videos, audio captions have evolved into a popular strategy to overcome knowledge gaps that can occur when a user misses parts of an auditory transmission by transcribing it into text. The first part of this section therefore reviews the use of audio-visual presentation – with a special focus on captions – as a strategy to improve the understanding of videos.

While captions may improve comprehension, they can also distract from visual content (Jiménez et al., 2011; Winke et al., 2010). Therefore, detecting confusion has been of particular interest in digital learning research to identify concrete events that require corrective actions (Brusilovskiy, 1994). Traditionally, confusion is inferred from interaction data or conversations with intelligent tutors. Yet, the data is only available if students actively engage in discussions (D'Mello et al., 2008). Recently, physiological measures have been explored, including electroencephalography (EEG), heart rate variability, electrodermal activity, facial electromyography, and eye movements (Arguel et al., 2017). An alternative to physiological indicators – which require that all users have access to and wear the necessary sensors – are vision-based cues which can be extracted from video frames recorded with an ordinary web camera. As, typically, only the face is recorded, facial expressions and cues from eye activity – in particular pupil dilation and blink rate – are rich sources of information. While empirical evidence relates pupil dilation to a person's cognitive state, it is highly sensitive to luminance (Charles & Nixon, 2019). Facial expressions and blinks are more robust in uncontrolled

settings. The second part of this section thus reviews recent advances in confusion detection from blink rate and facial expressions.

4.1.1. Audio-visual presentation

The beneficial effect of augmenting speech output with a visual representation has received ample empirical support. In the following, we discuss two distinct forms of visualizations: (1) Visual speech in the form of lip movements, and (2) video captions.

Visual speech. McGurk and MacDonald (1976) showed that accompanying an auditory presentation with a video recording of the speaker affects perception. The concomitant faster processing of the stimulus – commonly known under the term *McGurk effect* – has been validated in a number of empirical studies (Munhall et al., 1996; Sekiyama, 1997; Sekiyama & Tohkura, 1991; van Wassenhove et al., 2005). The effect is particularly strong when auditory comprehension is compromised by background noise, unclear articulation, or fast speech (Munhall et al., 1996; Sekiyama & Tohkura, 1991). Van Wassenhove et al. (2005) found that lip movements reduce the processing time substantially when they are easy to distinguish and match the auditory speech exactly. Due to their phonetic particularities, Asian languages including Japanese (Sekiyama & Tohkura, 1991) and Chinese (Sekiyama, 1997) produce a weaker effect than the clearly distinguishable sounds of the English language.
However, displaying a video of the speaker is not always possible or desirable. For example, speakers often present important visual content on a shared screen, or do not wish to turn on their camera due to privacy concerns (Bennett et al., 2021). In such situations, an alternative visual augmentation strategy is to provide a written transcription of the speech output.

Video captions. Video captions transcribe speech and sounds to overlay a synchronized textual representation on the video (Spina, 2021). While, in the United States, 'subtitle' typically refers to the written translation of auditory output into another language, we use the terms 'caption' and 'subtitle' interchangeably. In contrast to open captions, which are burned into the video and are always visible, closed captions can be activated and deactivated as needed. Captions for movies

are typically manually created, time-coded transcriptions, so that the correct text appears at the right time (Dialpad, 2021).

While originally developed to make videos accessible to people with hearing impairments (Kulkarni, 2019), a positive effect of captions on language learners has been demonstrated repeatedly. Learners of English are better able to understand captioned movies (Etemadi, 2012; Perego et al., 2010; Zareian et al., 2015) and recall their content (Hayati & Mohmedi, 2011). Recent studies suggest that captions are equally beneficial for native speakers. In a study with native Mandarin speakers, Zheng et al. (2022) observed better comprehension of audio lectures when a caption was provided. Morris et al. (2016) report a positive effect of captured lecture videos. Qualitative student feedback revealed that the subtitle was particularly helpful for clarifying segments where the audio was unclear or background noise complicated audio comprehension.
Virtual meeting platforms like Zoom (Larkin, 2021), Microsoft Teams (Microsoft, 2022), and Google Meet (Google, 2022) allow their users to activate live captioning. However, the auto-generated captions are error-prone. Depending on the platform and quality of the audio, accuracy ranges from 82% to 97% (Graham & Choo, 2022). Therefore, we conducted an experimental investigation to research whether auto-generated captions in virtual meetings are sufficiently good to foster comprehension.

4.1.2. Confusion detection from facial gestures

While research on the response of facial gestures to confusion is scarce, the effects of cognitive load – which is tightly related to confusion – have been investigated in a number of empirical studies. Cognitive load is defined as the amount of effort imposed on the working memory (Sweller et al., 1998). Confusion – while not a synonym – produces mental effort (Poehnl & Bogner, 2013). Given the transitive relation between the concepts, we extend our review of the literature to the effects of high cognitive load.

Blink rate. There is a consensus that blink inhibition occurs at moments of active task execution (Fairclough et al., 2005; Gao et al., 2013; Siegle et al., 2008). Observations of blink reactions to high cognitive load over longer periods are inconsistent and appear to depend primarily on whether the load increasing

activity requires constant visual attention. Blink inhibition occurs when the complexity of vision-based tasks increases. In a presumably subconscious attempt to minimize eye occlusion, the blink rate is reduced. The effect has been observed during control tasks including strategic games (Chen et al., 2011; Fowler et al., 2019; Mallick et al., 2016; Orden et al., 2001), driving (Borghini et al., 2012; Faure et al., 2016), flight control (Brookings et al., 1996; Ryu & Myung, 2005; Veltman & Gaillard, 1996), or the performance of parallel monitoring and tracking in a Mulit-Attribute Task Battery (MATB) (Fairclough et al., 2005; Fournier et al., 1999). It has also been found during text analysis (Ahmad et al., 2020; Bafna et al., 2020; Peitek et al., 2018) and visual search (Zagermann et al., 2018).

Primarily mental activities that require only sporadic visual checks have the opposite effect. For example, car drivers solving mental arithmetic problems as a secondary task had higher blink rates compared to a drive-only scenario (Faure et al., 2016; Tsai et al., 2007). The same effect has been observed on computer administered arithmetic tasks of increasing difficulty (Chen & Epps, 2013; Jyotsna & Amudha, 2018). Jyotsna and Amudha (2018) used a sequence of five equations to evoke spikes of cognitive load. Chen and Epps (2013) displayed addends sequentially, so that mental load was high at the onset of each number. The stepwise execution of a complex emergency procedure in a nuclear power plant led to higher blink rates than a simplified procedure (Chen et al., 2019; Gao et al., 2013). The findings from these studies imply that blinks are placed routinely at strategic points, e.g., after the visual display of an operand. Blinking whenever it is not imperative to have a clear visual focus – even if the biological state of the eye does not require it – leads to the observed acceleration of the blink rate. In line with these observations, Cho (2021) reports a lower blink entropy – implying less random blink sequences – during mental calculations as the difficulty of arithmetic problems increases. Similar blink synchronization mechanisms have been observed with movies, where blinks accumulate during scenes that require less visual attention (Nakano et al., 2009).

In summary, cognitive load causes blink inhibition when visual attention is essential. If visual focus is not permanently required, an overcompensation effect after load peaks may lead to an overall higher blink rate. Whether the compensation effect dominates blink inhibition during peaks depends on the extent to which visual focus is required. An example for the two conflicting effects was reported by Siegle

et al. (2008). Using a Stroop task – which requires to select the color in which a color word is written – they found no difference between incongruent stimuli where color words are written in a mismatched color and congruent stimuli.
Given that confusion occurs at moments of high cognitive load, we postulate that the same effects manifest whenever a user feels confused. Therefore, we expect confusion to produce a lower blink rate in captioned virtual meetings and – assuming that the overcompensation effect dominates – a higher blink rate when only auditory output is transmitted.

Facial expressions. Teachers and students alike consider facial expressions as the most relevant nonverbal indicator of comprehension in the classroom (Sathik & Jonathan, 2013). They agree that students display positive expressions when they have a good understanding of the lecture, and negative when they are confused. Shi et al. (2019) demonstrated that neural networks (NNs) can recognize relevant changes in the facial expressions of students participating in online courses. The authors used a convolutional neural network (CNN) to extract high-level features from images of the students' faces. A support vector machine (SVM) trained on the extracted features detected confusion with 90% precision and 78% recall.
Emotion databases with annotated images typically only differentiate between the six basic emotions happiness, anger, sadness, disgust, surprise, and fear (Li & Deng, 2020) which Ekman and Friesen (1971) identified as being universally recognized across cultures. Since there is a deficiency of training data for other emotions, research has taken an interest in determining which facial muscles are particularly active during confusion. Using electromyography, Durso et al. (2012) were able to detect confusion from electrical signals of muscular movements, but observed that the signals from some facial muscles such as the zygomaticus major add noise to the data, rather than providing helpful cues of confusion.
The Facial Action Coding System (FACS) categorizes these muscular movements into action units (AUs) (Ekman & Friesen, 1978). Early research relied on human judges to code AUs (D'Mello et al., 2009; Grafsgaard et al., 2011). In a study with AutoTutor, D'Mello et al. (2009) observed a correlation of confusion with AU4 (Brow Lowerer) and AU7 (Lid Tightener). The experienced affect was reported verbally by the students in 10-second intervals. In a follow-up study, two trained judges, a peer participant, and the students themselves assigned affect labels to a recording of the session in intervals of 20 seconds, or whenever they observed an

affect change. While AU4 and AU7 remained sensitive to confusion, the study did not corroborate the marginally significant effect on AU12 (Lip Corner Puller).
Grafsgaard et al. (2011) conducted a study in which students exchanged chat messages with a human tutor during a coding task. Assuming that AU4 is related to confusion, the authors identified behavior that indicates confusion. AU4 was activated whenever a student gave a shallow answer, or when the tutor pointed out a mistake without providing further explanations.
Recent advances in computer vision have fostered the emergence of software for automatic AU extraction. Borges et al. (2019) trained a long short-term memory (LSTM) with 20 AUs extracted with FaceReader (Loijens & Krips, 2021). Confusion in a navigation task that was designed to induce miscommunication between the navigator and follower was identified by three independent coders based on facial expression changes or delayed answers. Through lesioning, where some features are successively excluded from the model, the authors identified AU4 (Brow Lowerer), AU15 (Lip Corner Depressor), AU25 (Lip Part), AU26 (Jaw Drop), and AU27 (Mouth Stretch) as primary indicators of confusion.
Yasser et al. (2021) implemented a system that recognizes 18 AUs, seven of which were more frequently activated during confusion: AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), and AU23 (Lip Tightener). Logistic regression and quadratic discriminant classifiers trained on the seven AUs distinguished 'confused' from 'not confused' interview responses to personal social questions with 96% accuracy, although the authors do not disclose how confusion was labeled.
From the combined empirical evidence, a consensus forms that AU4 and AU7 imply confusion. The role of other AUs is more controversial – presumably, at least in part, a result of individual differences. Whitehill et al. (2008) report substantial inter-subject variability in the correlations of perceived lecture difficulty – which is linked to confusion – with action units. Difficulty self-reports were obtained by replaying the lectures on a frame-by-frame level and asking the participants to assign a difficulty rating on an eleven-point scale. While no clear pattern of AU activation emerged, AU2 (Outer Brow Raiser), AU15 (Lip Corner Depressor), and AU17 (Chin Raiser) were most frequently activated in the difficult lectures.

Clearly, uncertainties persist about the influence of the task and personal factors on the usefulness of facial gestures for predicting confusion. We seek to close

this gap by researching the effect of confusion on blink frequency, expressions of positive or negative emotional valence, and action units during two types of activities: a purely auditory listening task, and a captioned listening task which adds a visual component.

4.2. Study design: Confusion in audio-visual presentation

We conducted a user study with two objectives: First, to assess whether auto-generated captions improve the comprehension of an auditory report. Grounded on the beneficial effect of movie captions (cf. Section 4.1.1), we wanted to understand whether the – not always correct or appropriate – auto-generated captions have similar positive effects. We tested the following hypothesis:

> **H 2A.** Auto-generated captions improve comprehension in virtual meetings.

Instances of confusion were caused by orally relating news articles in which some words were replaced by phrases that are similar in sound, but out of context. Using a between-subjects experimental design, we counted how often participants who activated the captions reported confusion and verified whether sessions without captions produced fewer confusion reports.

Captioning virtual meetings by default may not be desirable for all users, especially if it covers parts of a shared screen, or occludes the speaker's face and obscures visual cues from lip movements which often foster listening comprehension (Brown et al., 2015; Sueyoshi & Hardison, 2005). The second goal was therefore to determine whether it is possible to identify concrete moments in which the benefits of captions are particularly promising. Specifically, we investigated whether facial gestures can detect confusion. We formulated the following hypothesis:

> **H 2B.** Facial gestures reveal confusion of virtual meeting participants.

A within-subjects experimental design was adopted. In an offline analysis, we assessed the effect of confusion on three types of facial gestures: (1) blinks, (2) expressions emotion, and (3) facial action units.

Assuming that the non-understanding issues are not a transient state, but instead are likely to persist for the rest of the meeting, we propose to implement a proactive adaptation to improve comprehension of the following spoken discourse.

Our analyses seek to raise awareness and help to understand the relevance of captions in virtual meetings (**H 2A**). By identifying facial gestures that are affected by confusion (**H 2B**), we supply software developers with a tool for recognizing – and, through the ad hoc activation of captions, counteracting – confusion in virtual meetings.

4.2.1. Auditory material

The study was designed to induce confusion as defined by the epistemic affective state which occurs when an individual attempts to process contradictory or incongruent information (D'Mello & Graesser, 2012; Vogl et al., 2020). Similar to the study by Durso et al. (2012), confusion was triggered by sentences in an article that appear incongruent. While the earlier study used written sentences, we read out loud two articles written in English. In order to account for mediating effects of their affective value, one sad and one funny article were chosen. The funny article relates an online interview with a professor which is rather inopportunely interrupted by his children. The sad article tells the story of a girl who lost her mother to the war in Ukraine. In each article, three sentences were altered so that they appear out of context. Unintended causes of confusion such as complicated names of people and places were removed to minimize the risk that the participants felt confused throughout the entire session. The presentation of each article took approximately 90 seconds. They were presented as follows, with underlined confusion triggers:

Article 1 (FUNNY):

Political expert Robert Kelly has given an update on his kids, five years after they caused him to go viral on the internet becoming better known as “BBC Dad". In the viral clip, unbeknownst to Robert, a small guest decided to join his interview halfway through, and viewers watched in shock as a cute child in a yellow jumper entered the room. **The BBC News predator [presenter] said: “I think one of your chickens [children] has just walked in"** as Robert tried to move his daughter out of the view of the camera, while still trying to maintain professionalism and talking about the subject at hand. But, another child comes whizzing into the room in a baby walker, meaning both his kids were now stealing the focus of the interview. He apologized as he said: “Pardon me, my apologies" **before the wolf was climbing and growling in [his wife also came bursting into] the room** to try and bustle the kids back out of the study. However her presence only managed to cause even more chaos as the child in the yellow jumper fell off the bed and the baby in the walker crashed into the door. Once she finally got the children out, the lady crawls back in to shut the door. Later Dr. Kelly said **he feared the ghost that was [incident would] mean** that he’d never be invited to do a TV interview again, but now, years after the incident, he is able to laugh about the memorable moment.

[adapted from Jones, Liverpool Echo, 1 March 2022]

Article 2 (SAD):

A letter written by a little girl reveals the heartbreaking toll of the war in Ukraine. Nine-year-old Galia penned the note inside what appears to be a day planner to her mother, who died in one city of Ukraine. A photo of the note was shared on Twitter by a reporter **from urine’s missing smells and fairs [Ukraine’s minister of internal affairs]**. Galia wrote the letter on March 8, less than two weeks after Russian President Vladimir Putin launched the deadly attack. This is a part of what she wrote. “Mum, this letter is your present. If you think that you nurtured me for no reason, you are not right. Thank you for the 9 years of my life. Thank you so much for my childhood. You are the best mother in the world! I will never forget you!" The reporter said that the little girl’s mother **died in a Japanese [Ukrainian]* city that has been making terrific dishes [facing horrific conditions]** since Russian troops invaded. Ukrainian President Volodymyr Zelenskyy said last week that the situation there is even worse than those in its neighboring city, where horrific images have surfaced of slaughtered civilians. On Wednesday, Zelenskyy said that “tens of thousands of people have already been killed" in Ukraine. “Russian army uses **all types of artists [artilleries], all types of muscles [missiles], aerial bombs, including phosphorus bombs and other tradition [ammunition]* banned by the interpersonal [international] law.**"

[adapted from Cohen, CBS News, 13 April 2022]

* Incongruities marked with an asterisk went unnoticed by most participants.

4.2.2. Pilot study

A pilot study was conducted with 14 participants to verify that the selected articles trigger discrete instances of confusion. In the pilot study, the articles were presented as an audio recording read by a native American English speaker. We observed that almost none of the participants experienced confusion at the onset of a trigger. Multiple subjects stated that the audio recording felt monotonous, which made it difficult for them to pay attention. In the main study, the articles were therefore read out loud by the experimenter – a second language English speaker of Asian background.

4.2.3. Participants

The study was performed remotely through Google Meet. Only subjects were selected that met the technical requirements: (1) The meeting should be joined from a laptop or personal computer with a webcam functionality. Participation from mobile devices was not possible. (2) A quiet environment with stable internet connection should be ensured. (3) If using visual aids, the participants were asked to wear contact lenses or take off glasses for the experiment. This was imposed after blink detection in the pilot study failed with glasses.

Through convenience sampling, we were able to recruit 45 university students from multiple departments and personal associates who met the requirements for participation (21 male, mean age = 28.4 $\pm$ 7.9). Most were undergraduate (N = 20) or graduate (N = 17) students. Participants with diverse cultural backgrounds were selected (29 European, 10 South East Asia, 4 South Asia, 1 Central Asia, 1 Africa). All were non-native English speakers with self-reported proficiency levels ranging from A2 to C2 (42 at least C1). The participants volunteered with no monetary or equivalent incentive. They were informed that the meeting was recorded and that the anonymized audio and video material would be used for research. Written consent was obtained prior to data collection.

4.2.4. Study design

We used a mixed design with between-subject variation of subtitle (captions, no caption) and two within-subject affect conditions for the articles (funny, sad).

Following a balanced design, participants were randomly assigned to the treatment group with captions (N = 23) or the control group (N = 22). Each group listened to both articles whose order was randomized to ensure that increased familiarity with the procedure did not affect the confusion reports.

4.2.5. Setup and procedure

The participants joined the one-on-one meetings through a Google Meet invitation link. The experimenter's camera was turned off during the experiment. This setup was chosen to ensure that only the effect of captions on comprehension was measured, and not the effect of visual speech (McGurk & MacDonald, 1976). In order to minimize participation barriers, the session was recorded on the experimenter's device. The Google Meet window with the participant's video stream and chat was recorded in full-screen mode at 30 fps with a resolution of 1920x1080 pixels (cf. Figure 4.1).

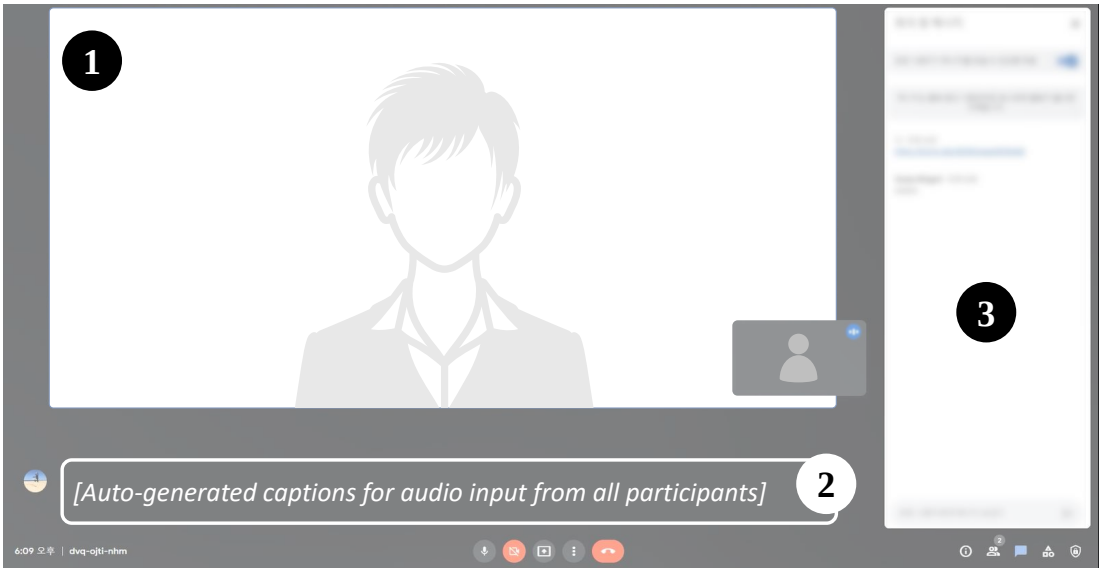


FIGURE 4.1: Exemplary Google Meet session recording, captured from the experimenter's view. The application settings were adjusted so that (1) the participant's video stream, (2) the captions (if available), and (3) the chat window were clearly visible.

Before the recording started, the participants were informed that the objective of the study was to collect data for developing a context-aware meeting tool. They then submitted demographic information in an online questionnaire. The treatment group was asked to turn on the captions and consult them whenever

they appeared helpful to them. All participants were instructed to open the chat window and enter 'x' whenever they felt confused or had troubles understanding the audio material that they would be presented. In order to avoid behavioral changes when pressing the keys, they were instructed to maintain one finger on each the 'x' and 'enter' key at all times. Research investigating gesture input in related contexts suggests that two-handed input is faster and preferred over one-handed operations (Rickel et al., 2022). Therefore, we chose a two-handed key sequence to minimize the cognitive effort of the key press. To simulate the context of a short presentation given during a video conference, no additional task apart from the confusion reports was given.

Subsequently, one of the two articles was read out loud by the experimenter. After listening to the first article, the recording stopped, and the participants completed a questionnaire to rate their understanding of the text and recall any passages that appeared out of context. The treatment group additionally rated the usability of the captions. The procedure was repeated for the second article. Overall, the experiment took about 15 minutes for each participant.

4.2.6. Data analysis

Confusion. The articles were manipulated to trigger confusion resulting from incongruent information, independent of English skills. However, since all subjects were non-native English speakers and the audio quality further challenged comprehensibility, multiple reported additional unintended confusion (i.e., non-understanding of spoken words), and others did not recognize the intentional incongruities. From the incongruent text passages that the participants recalled in the questionnaire it appeared that subtle alterations ("phosphorus bombs and other tradition", "banned by the interpersonal law") remained mostly unnoticed. In previous studies, confusion was often defined by the participants (Shi et al., 2019; Whitehill et al., 2008) and/or independent raters (Afzal & Robinson, 2010; Borges et al., 2019; D'Mello et al., 2009) by reviewing a video recording of the experimental session. The reliability of this method has seen mixed results (Conati et al., 2013), as it risks capturing only very obvious displays of confusion, while missing subtle expressions. Therefore, we chose an adapted version of the real-time self-reports proposed by Lallé et al. (2016) and used the chat entries as ground

truth. On average, subjects reported 3.29 (stdv = 2.53) confusing sentences in Article 1. Article 2 triggered 2.82 (stdv = 2.54) confusion reports. The reports were extracted on a per frame basis. A Python script was developed using motion analysis and template matching to identify the entry of an 'x'. The frames were manually revised to ensure that all timestamps were correctly exported.

Subtitle effectiveness. To appraise **H 2A**, we performed a series of statistical tests in which we compared the *Confusion Reports* – i.e., the number of times 'x' was typed into the chat – of the treatment group with captions to the control group. Table 4.1 summarizes the tested dependent and independent variables.

TABLE 4.1: Effect variables and fixed factors for testing subtitle effectiveness. Separate analyses for the confusion concepts *Non-Understanding* and *Incongruity Confusion* were performed to test whether captions have a different effect depending on the cause of the experienced confusion.

	Variable	Description
DEPENDENT	**Confusion Reports**	Number of times a participant reported confusion by typing 'x' in the chat window.
	Incongruity Confusion	Confusion from contradictory information. Measured by *Confusion Reports* submitted within 10 seconds after confusion trigger.
	Non-Understanding	Confusion from poor audio quality or language deficiencies. Measured by *Confusion Reports* that were submitted outside of the 10 second window after a confusion trigger.
INDEPENDENT	**Subtitle**	Availability of subtitles *(0: no caption, 1: captions)*
	Subtitle Usability	Reported usability of the auto-generated captions *(1: not at all, 2: sometimes, 3: absolutely)*

Subtitle: The effect of a *Subtitle* on *Confusion Reports* was tested with a two-sided t-test. To assess whether captions have a different effect depending on the cause of the experienced confusion, we additionally performed separate analyses for each of the two concepts of confusion that the experiment may trigger: (1) *Incongruity Confusion* resulting from contradictory information was defined as the number of *Confusion Reports* that occurred within 10 seconds following the display of an intentional confusion trigger; (2) *Non-Understanding* includes *Confusion Reports* that were submitted outside of this time window, and are thus attributed to poor audio quality or language deficiencies.

Subtitle Usability: We additionally wanted to understand the role of incorrect or inappropriate captions. We therefore performed a one-way between-subjects ANOVA to examine how the perceived *Subtitle Usability* affects the *Confusion Reports*. To measure *Subtitle Usability*, we administered an adapted version of Lewis' Computer System Usability Questionnaire (Lewis, 1995, Item 11) to the treatment group, asking whether they found the captions helpful for understanding the speech (1: not at all, 2: sometimes, 3: absolutely). The analysis of *Subtitle Usability* was again repeated for both confusion concepts.

Comprehension intervals. To research the effect of confusion on facial gestures (**H 2B**), we defined *Comprehension* intervals for a baseline, confusion, and reconciliation phase. Windows of equal size, each corresponding to five seconds, were chosen. Previous work has attempted to detect whether confusion occurs within intervals of 10 seconds (D'Mello et al., 2009) or 20 seconds (Afzal & Robinson, 2010; D'Mello et al., 2009). However, in order to determine whether a change in confusion is effectively the result of an interface adaptation (i.e., the activation of captions), it must be detected with minimal latency. We therefore tested whether changes in facial gestures manifest between any two five-second intervals coinciding with, following, or preceding a confusion trigger.
An inspection of the confusion reports showed that a reaction typically occurred with a delay of approximately 5 seconds (Article 1: median = 154 frames, Article 2: median = 162 frames). Figure 4.2 shows the timed onset of the confusion triggers and the subsequent confusion report for each participant.

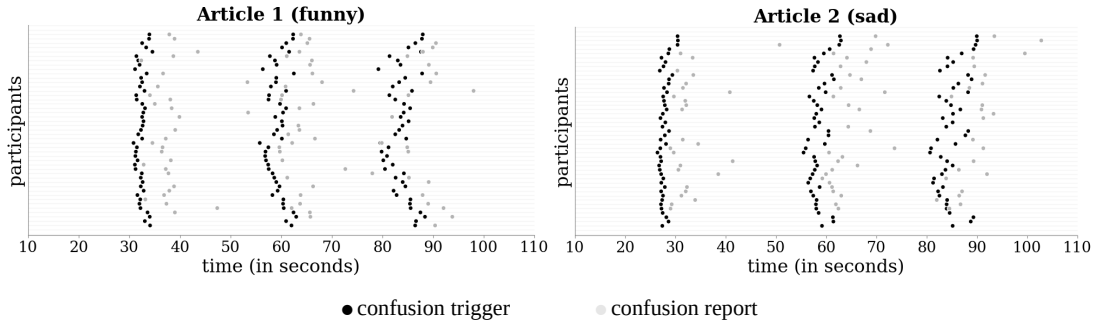


FIGURE 4.2: Reaction time from onset of a confusion trigger to report. Participants reported confusion with a median delay of approximately 5 seconds (Article 1: 154 frames, Article 2: 162 frames) after the onset of the trigger.

A 'confusion' interval thus started 5 seconds before the participant reported confusion, which typically coincided with the onset of the confusion trigger. It

always ended with the confusion report. Since the participants were explicitly instructed to report confusion, we presume that the report allowed them to attribute the perceived incongruities to the experimental manipulation. Knowing about their intentionality would allow them to overcome confusion in the following 'reconciliation' interval. The 'baseline' interval was defined as the 5 seconds preceding a confusion interval. Figure 4.3 illustrates the temporal sequence of the *Comprehension* intervals.

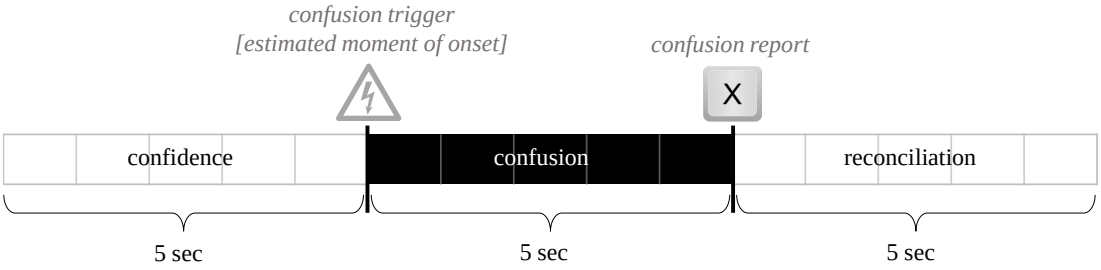


FIGURE 4.3: Comprehension intervals. Facial gestures during the *baseline*, *confusion*, and *reconciliation* were calculated for 5 second windows.

Facial gestures. Blinks, expressions of emotion, and AUs were extracted for each *Comprehension* interval. Table 4.2 summarizes the facial gestures that were considered as confusion indicators.

Blinks: We used facial landmark detectors from the `dlib` library[1] to extract eye contours for calculating the Eye Aspect Ratio (EAR) as proposed by Čech and Soukupová (2016). While a high EAR indicates widely opened eyes, a low EAR does not necessarily imply that a blink occurred. It may also capture other facial expressions such as yawning. Since blink detection using fixed EAR thresholds therefore typically perform poorly, we used the detection algorithm proposed by Genchi et al. (2019)[2], which implements a pre-trained SVM. Blinks were extracted on a per-frame basis, using the first frame in which the eyelid was completely closed as a reference. The extracted blinks were revised manually by two independent researchers by inspecting the recorded videos and comparing observable blinks with the model output. On average, 42.98 (stdv = 24.03) blinks were observed for each participant, and the model identified 40.67 (stdv = 16.56) blinks. Blinks that were not identified by the model were mostly a result of the low image resolution which prevented successful eye contour detection. Applications that extract facial

[1]Dlib C++ library: http://dlib.net/ (King, 2009).
[2]Blink detection model: https://github.com/rmenoli/Eye-blinking-SVM.

TABLE 4.2: Facial gestures considered as confusion indicators. Extraction of the metrics was automated using computer vision techniques.

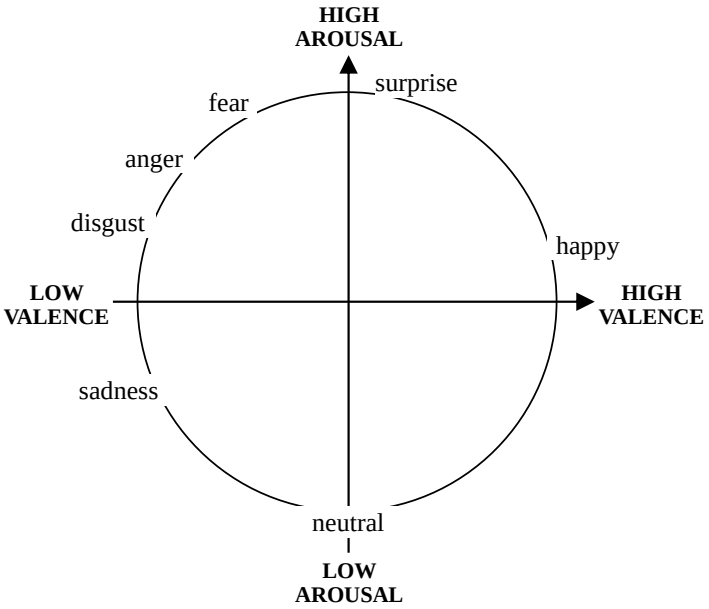
Metric	Description	Construct
BLINKRATE	Number of blinks per second.	**Blinks**
POSITIVE, NEGATIVE, NEUTRAL } EMOTION	Percentage of frames with recognized expression of {*positive, negative, neutral*} emotion.	**Emotion**
EMOTIONFLUCTUATION	Number of times the expression switches to neutral, positive, or negative emotions. Normalized over the number of frames.	
AU1	Inner Brow Raiser	**Action units**
AU2	Outer Brow Raiser	
AU4	Brow Lowerer	
AU5	Upper Lid Raiser	
AU6	Cheek Raiser	
AU7	Lid Tightener	
AU9	Nose Wrinkler	
AU10	Upper Lip Raiser	
AU12	Lip Corner Puller	
AU14	Dimpler	
AU15	Lip Corner Depressor	
AU17	Chin Raiser	
AU20	Lip Stretcher	
AU23	Lip Tightener	
AU25	Lip Part	
AU26	Jaw Drop	
AU28	Lip Suck	
AU45	Blink	

gestures on the user's device or forward the original camera images directly to a remote server for further processing can easily overcome quality issues. Therefore, we corrected the model output for unidentified blinks to ensure that the blink activity is accurately represented. To measure BLINKRATE, the number of blinks per second was calculated.

Emotion: Based on the observation that confusion shows in negative facial expressions (Sathik & Jonathan, 2013), we wanted to understand whether automatic facial expression recognition (FER) can be used to detect confusion. Facial expressions were extracted with the open source Python model developed by Shenk et al. (2021). The software implements a pre-trained CNN and is able to distinguish between the six basic and neutral emotions. We applied the model to each frame and extracted the most intense emotion.

Even though the use of neural networks has considerably improved the iden-

tification of facial expressions, distinguishing between similar emotions is still challenging, especially when low image resolution introduces noise (Zhang et al., 2022). In the classifications that we obtained, the participants only rarely demonstrated disgust (0.3%), anger (3.4%), fear (4.4%), and surprise (0.9%). Therefore, we distinguish between positive, negative, and neutral emotions on an aggregated level. Discrete emotions were classified as positive (happy, surprise) or negative (anger, disgust, fear, sadness) according to the classification scheme proposed by Plutchik (1991) (cf. Figure 4.4). It is based on the Circumplex Model of Affect (Posner et al., 2005) which defines emotions as a linear combination of arousal (i.e., emotion intensity) and valence (i.e., positive/ negative emotion). To measure the intensity of each class, we counted the number of frames with POSITIVEEMOTION, NEGATIVEEMOTION, and NEUTRALEMOTION. As confusion is a fleeting state, we expected that it would cause frequent changes between expressions of different emotions. Therefore, we additionally calculated EMOTIONFLUCTUATION as the number of times the valence of the recognized expression changes.



POSITIVEEMOTION	happy surprise
NEGATIVEEMOTION	anger disgust fear sadness
NEUTRALEMOTION	neutral

FIGURE 4.4: Circumplex Model of Affect (Posner et al., 2005). Emotions were assigned to the valence of their quadrant (adapted from Plutchik (1991)).

Action units: While empirical evidence acknowledges that confusion shows in expressions of emotional valence (Sathik & Jonathan, 2013), we expected that the fine-granular changes in AUs deliver more robust predictions. We used `OpenFace`[3]

[3]OpenFace: https://cmusatyalab.github.io/openface/ (Amos et al., 2016).

for automatic AU extraction. As a measure of their activation, we extracted the intensity for each of the 18 AUs that are recognized by `OpenFace`. AU45 (Blink) was included in addition to the EAR based BLINKRATE, as it considers all activations in the muscles that are associated with blinking, including small twitches of the lid. In contrast, the EAR model only extracts blinks if the lid is closed for a defined number of frames which is determined by the SVM.

Interaction effects. We tested whether the effect of confusion on facial gestures (**H 2B**) is conditional on the *Article Topic* or the *Subtitle*. Table 4.3 summarizes the main and interaction effect variables.

TABLE 4.3: Main and interaction effect variables of *Comprehension*. *Subtitle* and *Article Topic* were included to test whether they interact with the main effect of *Comprehension*.

	Variable	Description
MAIN EFFECT	**Comprehension**	Five second interval of subjects' comprehension state, identified by a confusion report. • *confusion*: interval starting four seconds before, and ending one second after a confusion report • *recovery*: interval directly following a confusion interval • *confidence*: interval directly preceding a confusion interval
INTERACTION	**Subtitle**	Availability of subtitles *(0: no caption, 1: captions)*
	Article Topic	Experimental trial identified by the topic of the presented article *(Article 1: funny, Article 2: sad)*

Article Topic: While the influence of emotions on facial expressions is undisputed, empirical evidence suggests that positive emotions also reduce the BLINKRATE (Lange et al., 2022). Therefore, we wanted to understand whether the affective value of the *Article Topic* (funny, sad) determines whether confusion manifests itself in facial gestures.

Subtitle: Reading has been associated with a reduced blink rate (Bentivoglio et al., 1997; Lenskiy & Paprocki, 2016) and multiple AUs – particularly those related to eyebrow movements – change during attentive reading (Li et al., 2016). Therefore, we included the *Subtitle* (captions, no caption) as an additional control variable.

4.3. Results: The role of confusion in audio-visual presentation

The meetings with all 45 participants were successfully recorded and could be used for the analysis. Between 2426 and 2774 frames (mean = 2616 ± 77) were recorded from the presentation of one article. The analyses were performed on the aggregated data from both articles.

4.3.1. Effectiveness of auto-generated captions (H 2A)

In the treatment group, where captions were displayed, 74% indicated that they consulted the captions when feeling confused. We used a two-sample t-test to compare the *Confusion Reports* from the treatment group (1: captions) to the control group (0: no caption). The role of perceived usability of the captions was tested with a one-way between-subjects ANOVA on *Confusion Reports*, using *Subtitle Usability* (with an additional 'no caption' group) as fixed factor. The test measures whether the subjective feeling that captions are helpful for understanding the spoken reports affects how often a user experiences confusion. Post-hoc comparisons were performed with Šidák corrected pairwise t-tests.

Experimental manipulation check. To appraise the validity of *Confusion Reports* as an indicator for comprehension of the articles, we tested its correlation with the participants' ratings of whether the content was understandable (1: barely, 2: moderately, 3: completely). Subjects who indicated a good understanding tended to issue fewer *Confusion Reports*. Pearson's product-moment correlation confirmed a significant negative relation (corr. = -.262, p-value = .013).

Effect of Subtitle on Confusion Reports. A two-sided t-test showed no significant effect of *Subtitle* on *Confusion Reports* (t = 0.792, p-value = .431). Additional t-tests performed on each of the two confusion concepts separately confirmed that auto-generated captions have no significant effect on *Incongruity Confusion* (t = 0.570, p-value = .570). However, the effect was marginally significant for *Non-Understanding* (t = 1.735, p-value = .086). This implies that captions that replicate the information of an incongruent auditory report cannot alleviate confusion caused by the contentual contradiction. In contrast, the

cognitive disequilibrium resulting from poor audio quality or language deficiencies can be mitigated.

Effect of Subtitle Usability on Confusion Reports. The one-way ANOVA on *Confusion Reports* revealed a significant effect of *Subtitle Usability* (F = 4.327, p-value = .007, η^2 = .131). Participants who found the captions 'absolutely' helpful submitted less *Confusion Reports* than the control group with no captions (t = 2.995, p-value = .024, η^2 = .089). Differences between other groups were not significant.

Differentiated analyses of the two confusion concepts confirmed that *Subtitle Usability* has a significant effect on *Non-Understanding* (F = 4.636, p-value = .005, η^2 = .139). Compared to the control group with no captions, subjects of the treatment group reported less instances of *Non-Understanding* if they found the captions 'absolutely' (t = 3.034, p-value = .021, η^2 = .090) or 'sometimes' (t = 3.258, p-value = .011, η^2 = .096) useful. The ANOVA on *Incongruity Confusion* reported a significant effect of *Subtitle Usability* (F = 3.413, p-value = .021, η^2 = .106). However, post-hoc pairwise comparisons showed no significant differences between any of the groups.

H 2A SUMMARY: A comparison of the experimental groups revealed that auto-generated captions reduce confusion if they are perceived as useful. In particular, participants who found the captions 'absolutely' or 'sometimes' helpful reported significantly less *Non-Understanding* caused by poor audio quality or language deficiencies than the control group with no captions. In contrast, auto-generated captions do not mitigate *Incongruity Confusion* resulting from contradictory auditory reports – independent of whether they are perceived as useful.

4.3.2. Confusion in facial gestures (H 2B)

To research the effects of confusion on facial gestures in the presence or absence of captions, a two-way mixed ANOVA with *Comprehension* and *Subtitle* as fixed factors was performed for each facial gesture. Influences of the affective value of the article were tested in additional two-way repeated-measures ANOVAs with *Comprehension* and *Article Topic* as fixed factors. Since the interaction between *Article Topic* and *Subtitle* is immaterial to the research question, each confounding factor was evaluated in a separate ANOVA. Mauchly's test of sphericity was

applied to all facial gestures to test whether the assumption of equal variances of group differences was met. The p-values of groups with unequal variances in repeated-measures ANOVAs were Greenhouse-Geisser corrected by adjusting the degrees of freedom for F-value calculations by their sphericity. Post-hoc analyses of significant group effects were performed using pairwise t-tests with Šidák correction. Table 4.4 summarizes the statistics of significant pairwise comparisons. For better readability, we only report significant interaction and main effects of *Comprehension*. The complete statistical output is included in Appendix A.

TABLE 4.4: Significant effects of *Comprehension* on facial gestures. Group differences were tested for the baseline, confusion, and reconciliation intervals. Significance was determined with two-way mixed ANOVAs using *Comprehension* (C) and *Subtitle* (S), and two-way repeated-measures ANOVAs with *Comprehension* and *Article Topic* (T) as fixed factors. For the interaction effects C*S and C*T, group differences are reported if they are significant within the specified interaction group. Pairwise t-tests with Šidák correction were used for post-hoc comparisons.

		ANOVA			**groups**				**pairwise t-test**		
	Factor	F	sig.	η^2	interaction	A	B	[A] μ (σ)	[B] μ (σ)	t	sig.	η^2
BLINK RATE	C	5.285	.007	.107		baseline	– confusion	.401 (.272)	.333 (.255)	2.952	.015	.016
						confusion	– reconcile	.333 (.255)	.404 (.282)	2.753	.025	.017
	C*S	2.778	.068	.008	[caption]:	baseline	– confusion	.411 (.296)	.288 (.227)	3.832	.005	.052
					[caption]:	confusion	– reconcile	.288 (.227)	.387 (.309)	2.586	.097	.032
NEUTRAL EMOTION	C*S	5.559	.005	.002	[caption]:	baseline	– reconcile	.581 (.359)	.540 (.363)	2.641	.086	.003
AU4	C	3.083	.066	.065		baseline	– reconcile	.519 (.520)	.563 (.542)	2.195	.097	.002
AU6	C	3.041	.058	.065		baseline	– reconcile	.246 (.336)	.278 (.346)	2.201	.096	.002
AU7	C	3.830	.033	.080		baseline	– reconcile	.628 (.696)	.684 (.712)	2.354	.068	.002
AU10	C	3.914	.034	.784		baseline	– reconcile	.298 (.396)	.333 (.388)	2.327	.072	.002
AU17	C	5.094	.020	.104		baseline	– confusion	.291 (.329)	.333 (.357)	3.899	.001	.004
						baseline	– reconcile	.291 (.329)	.356 (.439)	2.590	.038	.007
	C*T	3.370	.039	.017	[sad]:	baseline	– confusion	.283 (.361)	.344 (.459)	2.799	.045	.006
						confusion	– reconcile	.274 (.339)	.399 (.589)	2.510	.091	.016
AU23	C	6.983	.002	.137		baseline	– confusion	.063 (.074)	.096 (.121)	3.335	.005	.026
						baseline	– reconcile	.063 (.074)	.118 (.184)	2.945	.015	.037
	C*T	2.959	.057	.033	[funny]:	baseline	– confusion	.077 (.123)	.106 (.144)	2.758	.050	.012

Blinks. The BLINKRATE showed substantial variations across participants. Intervals between blinks lasted between 27 frames ($\sim$ 1 second) and 446 frames ($\sim$ 15 seconds).

The main effect of *Comprehension* (F = 5.285, p-value = .007, η^2 = .107) and the interaction with *Subtitle* (F = 2.778, p-value = .068, η^2 = .007) were significant. Pairwise comparisons revealed a lower BLINKRATE during confusion compared to the baseline (t = 2.952, p-value = .015, η^2 = .016) and reconciliation (t = 2.753, p-value = .025, η^2 = .017). The interaction effect indicated that only subjects

with captions blinked less during confusion compared to the baseline (t = 3.832, p-value = .005, η^2 = .052) and reconciliation (t = 2.586, p-value = .097, η^2 = .032).

Emotion. The FER model failed to detect facial landmarks for one participant due to the low contrast of skin against background. Since the model failed even after applying multiple filters, the sample was dropped for the emotion analysis. The main effect of *Comprehension* was not significant for any emotion class or EMOTIONFLUCTUATION. ANOVA tests reported a significant interaction with *Subtitle* for NEGATIVEEMOTION (F = 4.090, p-value = .020, η^2 = .002) and NEUTRALEMOTION (F = 5.559, p-value = .005, η^2 = .002). Post-hoc tests confirmed the effect for NEUTRALEMOTION (t = 2.641, p-value = .086, η^2 = .003). Expressions of NEUTRALEMOTION were more frequent during the baseline compared to the reconciliation phase if captions are supplied.

Action units. ANOVA tests reported a significant main effect of *Comprehension* on **AU4** (F = 3.083, p-value = .066, η^2 = .065), **AU6** (F = 3.041, p-value = .058, η^2 = .065), **AU7** (F = 3.830, p-value = .033, η^2 = .080), **AU9** (F = 3.397, p-value = .038, η^2 = .071), **AU10** (F = 3.914, p-value = .034, η^2 = .784), **AU17** (F = 5.094, p-value = .020, η^2 = .104), **AU23** (F = 6.983, p-value = .002, η^2 = .137), and **AU45** (F = 2.796, p-value = .067, η^2 = .060). The interaction effect with *Article Topic* was significant for **AU17** (F = 3.370, p-value = .039, η^2 = .017) and **AU23** (F = 2.959, p-value = .057, η^2 = .033).
Post-hoc comparisons revealed a more intense activation during reconciliation compared to the baseline for **AU4** (t = 2.195, p-value = .097, η^2 = .002), **AU6** (t = 2.201, p-value = .096, η^2 = .002), **AU7** (t = 2.354, p-value = .068, η^2 = .002), **AU10** (t = 2.327, p-value = .072, η^2 = .002), **AU17** (t = 2.690, p-value = .038, η^2 = .007), and **AU23** (t = 2.945, p-value = .015, η^2 = .037).
A higher activation during confusion compared to the baseline was observed for **AU17** (t = 3.899, p-value = .001, η^2 = .004) and **AU23** (t = 3.335, p-value = .005, η^2 = .026). The analysis of *Article Topic* interaction showed that only the sad article triggers the effect on **AU17** (t = 2.799, p-value = .045, η^2 = .006). The effect on **AU23** was limited to the funny article (t = 2.758, p-value = .050, η^2 = .012). For the sad article, we additionally observed a lower activation of **AU17** during confusion than reconciliation (t = 2.510, p-value = .091, η^2 = .016). In summary, the results of the statistical analyses imply that confusion shows in multiple action units, but typically only after a confusion report.

H 2B SUMMARY: Confusion activates **AU4** (Brow Lowerer), **AU6** (Cheek Raiser), **AU7** (Lid Tightener), **AU10** (Upper Lip Raiser), **AU17** (Chin Raiser), and **AU23** (Lip Tightener). With captions, it additionally reduces the BLINKRATE and expressions of NEUTRALEMOTION. For most features, the effect surfaces in the interval five seconds after confusion is reported. **AU17** and **AU23** are additionally activated – albeit not as strongly – directly at the onset of a confusion trigger. Yet, the immediate effect is confined to sad topics in **AU17**, and funny topics in **AU23**. In contrast, the BLINKRATE drops immediately after the onset of a trigger, and then quickly returns to its original frequency.

4.4. Discussion & implications of multimodal redundancy

The analysis of auto-generated captions (**H 2A**) revealed their potential for reducing non-understanding (i.e., confusion from poor audio quality or language deficiencies) in virtual meetings, but only if perceived as useful. Perceived usefulness may stem either from individual preferences for uni- or multimodal interaction (Oviatt et al., 2003, 2005; Xiao et al., 2002), or from the quality of the captions. This shows an important limitation of auto-generated captions compared to videos with verified captions (Hayati & Mohmedi, 2011; Winke et al., 2010). Confusion caused by contradictory auditory reports, in contrast, is unfazed by auto-generated captions which merely replicate the incongruities of the auditory presentation.

The examination of facial gestures (**H 2B**) identified them as a promising contender for detecting confusion in virtual meetings. It is reflected in action units AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU10 (Upper Lip Raiser), AU17 (Chin Raiser), and AU23 (Lip Tightener). Inferred confusion can thus be used as a trigger to activate captions in moments of need. The summary of findings from related studies in Table 4.5 shows that, previously, only AU4 and AU7 were unanimously associated with confusion (Borges et al., 2019; D'Mello et al., 2009; Yasser et al., 2021). While Yasser et al. (2021) also report signs of confusion in AU6, AU10, and AU23, our findings for AU17 are unprecedented. In captioned meetings, confusion additionally led to a lower BLINKRATE and less expressions of NEUTRALEMOTION. Findings from previous studies in which high cognitive load was associated with blink inhibition during reading (Ahmad et al., 2020; Bafna et al., 2020; Peitek et al., 2018) can thus be relayed to confusion. This identifies the BLINKRATE as an important – and so far mostly overlooked –

confusion indicator. Blinks are subtle and may identify confusion even if strong emotions or concentration reduce the predictive power of other features. Yet, the confinement to vision-based tasks has important implications for its application in practice. Moreover, blinks occur at intervals, whereas changes in action units can be identified on the granularity of single frames. Therefore, the combination of different facial gestures is another promising research direction.

TABLE 4.5: Comparison of facial gestures identified as relevant for confusion prediction in related research. Only related work that studies confusion (excluding cognitive load studies) and performs an analysis of relevant facial gestures is compared.

Research	Study & task	Ground truth collection	Analysis method	Relevant features
Yasser et al. (2021)	120 participants are interviewed on personal social issues	Segmentation into 'confused' and 'not confused' response	n/d	AU4, AU5, AU6, AU7, AU10, AU12, AU23
D'Mello et al. (2009)	**Study 1:** 7 students study a computer literacy topic in AutoTutor for 90 minutes	Students verbally report when they experience emotions	Correlation with AUs extracted 3 seconds prior to ground truth emotion	AU7, AU4, AU12
	Study 2: 28 students study a computer literacy topic in AutoTutor for 35 minutes	Students and three judges code emotions in 20 second intervals		AU7, AU4
Borges et al. (2019)	13 subjects receive map navigation instructions designed to cause misunderstandings	Three judges identify confusion based on changes in facial expression or delayed answers	Lesioning with LSTM prediction	AU4, AU15, AU25, AU26, AU27
Our work	45 subjects listen to article readings, half of the sample with captions	Participants enter 'x' when they are confused	ANOVA	**All sessions:** AU4, AU6, AU7, AU10, AU17, AU23 **Captioned sessions only:** blinkRate, neutralEmotion

The observed changes in AUs and neutralEmotion occurred in the interval after confusion was reported. An immediate response to an emotion trigger was only visible in AU17 if a sad, and AU23 if a funny topic was presented. In contrast, D'Mello et al. (2009) detected effects on AU4, AU7, and AU12 three seconds before an oral emotion report. It thus appears that the ground truth collection procedure determines its temporal relation to changes in facial gestures and should receive careful consideration in future studies.

Since facial gestures do not reveal the cause of confusion – in particular, incongruent information versus non-understanding from poor audio quality or language deficiencies – a static activation of captions in response to a one-time detection of

confusion may not be desirable. In fact, it can even be detrimental to users who are only confused by contentual incongruities if the captions divert their attention away from – potentially clarifying – information on a shared screen (Jiménez et al., 2011; Winke et al., 2010). We thus propose that confusion levels should be monitored continuously in the background. If the captions do not reduce confusion, they should again be deactivated.

4.4.1. Threats to validity

Conclusion validity: *Are there issues that might affect the relation between experimental treatment and outcome?* The one-on-one meetings of the study may not represent a typical meeting setup. However, we expect users to behave similarly in meetings with multiple participants. Moreover, since confusion detection and adaptive responses are performed for each user independently, the number of meeting participants should not interfere with the adaptation component.
While offline extraction of AUs and emotions similar to the procedure that was used in the study would limit their suitability for consumer applications, real-time extraction is possible with open-source tools such as `CERT`[4] and `OpenFace`[5].

Construct validity: *Do measurement errors or variable definitions compromise the relation between theory and observation?* Even though they were instructed to face the screen, some subjects turned their head during the experiment. While this may have reduced the quality of the recordings, it provides a realistic benchmark and promises the features to be robust in unconstrained settings.
Similarly, low image resolution prevented the blink detection model from correctly extracting all blinks. Applications that extract facial gestures on the user's device or a remote server can have access to high-quality images. The reported results, which are based on manually corrected blinks, are therefore subject to the assumption that the user's camera produces images of sufficient quality to accurately detect blinks.
A controversial issue is the definition of confusion labels. We noted that not all subjects reported confusion after the onset of intentional triggers. Psychology research has shown that humans are capable of understanding incomplete sentences

[4] `CERT`: https://inc.ucsd.edu/mplab/users/marni/Projects/CERT.htm (Littlewort et al., 2011).
[5] `OpenFace`: https://cmusatyalab.github.io/openface/ (Amos et al., 2016).

by filling in missing words (Cohen & Faulkner, 1983). Since the manipulated sentences remained phonetically close to the original text, some subjects may have subconsciously disentangled the correct meaning and thus did not report confusion. Since recognizing incomplete sentences is mentally demanding (Cohen & Faulkner, 1983), the effort may nevertheless be reflected in some facial gestures. To understand the implications of such an effect, we repeated the analysis using the intentional triggers as ground truth. Given that this resulted in a substantially reduced effect on facial gestures we presume that, even if a cognitive process takes place, its effect on facial gestures is sufficiently disjunct from confusion.

Internal validity: *Are there any moderating effects?* The physical effort from pressing a key to report confusion may have caused unintended effects on facial gestures. While less noticeable than mental effort, physical exertion changes the facial expression (Tian et al., 2005). However, in the interval following a confusion report – where we observed changes in facial gestures – the effects of physical exertion should already have subsided.

External validity: *Can the results be generalized to other target groups?* Age, gender, and cultural background influence the success of face detection (Abdurrahim et al., 2018) and the intensity of facial expressions (Sohail et al., 2022). While we did not test for differences between demographic subgroups, we carefully selected a diverse sample to identify facial gestures that are independent of demographic characteristics.
Since we only tested English captions and all participants were advanced non-native speakers, it is difficult to estimate whether captions have the same effect on native speakers or users with lower language skills. However, empirical support for a beneficial effect on native speakers with English (Morris et al., 2016) and Mandarin (Zheng et al., 2022) captions suggests a broad generalizability.

4.4.2. Relevance for research

The findings from the experimental investigation confirm that activating captions at all times may not be desirable if personal preferences lean towards unimodal presentation, or the quality of the auto-generated captions is unsatisfactory. It may even be detrimental if captions occlude visualizations on a shared screen or the speaker's face (Brown et al., 2015; Oviatt, 2006; Sueyoshi & Hardison,

2005). Dynamic caption activation may thus be one measure that adaptive multimodal interfaces can implement to improve comprehension, but should be used in conjunction with additional steps. This is especially crucial if challenging speech input (e.g., from a non-native presenter) results in incorrect captions. Since inadequate captioning may even increase confusion, special care must be taken to ensure that the captions have no negative effects. As a countermeasure, confusion should be continued to be monitored after activating the captions. The real-time feedback on the user's level of confusion can then be used to evaluate the effect of the multimodal presentation, and the application should revert to unimodal output if there is no improvement in comprehension after the captions are activated. Another option – in line with the trend towards more human control (Shneiderman, 2020) – is to recommend the captions to the user, but leave the final decision to them. As the quality of the captions depends on the enunciation of the speaker (Benzeghiba et al., 2007), it is advisable to re-evaluate its usefulness for each speaker separately.

4.4.3. Relevance for practice

The dynamic activation of captions (**H 2A**) can be beneficial in multiple use cases outside of the virtual meeting scenario that we studied – most notably in online learning and captioned movies.
Moreover, the ability to detect confusion in virtual meetings (**H 2B**) enables not only dynamic caption activation, but also additional functionalities such as feedback for presenters.

Online learning. Timestamped confusion detection in video tutorials enables innovative features, including interactive subtitles such as SubMe which uses eye tracking to infer English skills (Fujii & Rekimoto, 2019). Captioning only difficult words proved to be not enough to understand a movie. However, captions of the entire audio with additional definitions for difficult words received positive evaluations. Alternatively, Schneegass et al. (2019) propose creating personalized vocabulary lists with words from audio-visual material that cause incomprehension.

Captioned movies. Netflix reports that 80% of its subscribers use subtitles or captions at least once a month (Robison, 2019). Some sporadically use captions to better understand quickly spoken or poorly enunciated dialogues (Office of

Communications, 2006). Since their need for captions depends on the movie – and may even change with the speaker within one movie – they would benefit from automatic caption activation whenever comprehension issues are detected.

Feedback for presenters. Emotional reactions in online meetings are often hard to read, and presenters lament the lack of direct feedback (Frisch & Greene, 2020). The only options are usually polling tools, but actively providing input interrupts the presentation. Alternative polling techniques with hand gestures (Koh et al., 2022) or attention tracking through body posture (Revadekar et al., 2020) are less intrusive. Meeting platform providers are therefore working on innovative features such as the Microsoft Teams AffectiveSpotlight which tracks facial expressions and movements to highlight emotional responses to the presenter (Murali et al., 2021; Stokel-Walker, 2021).
Moreover, receiving feedback about the current mood in the audience may not only benefit the presenter, but other meeting participants as well (Fessl et al., 2012). Knowing that others feel equally confused might lower the barriers to express comprehension issues.

4.5. Conclusion: Summary of Essay 2

Many attendants of virtual meetings experience confusion, but are hesitant to express their doubts. In movies, captions are an established tool to improve comprehension. In contrast, the quality of auto-generated captions in virtual meetings is often poor and their benefit unclear. To research how captions can be used effectively in virtual meetings, we conducted a user study with 45 Google Meet users. The study revealed two major insights:

1. **Captions:** Auto-generated captions are only moderately helpful for alleviating confusion, as they are only effective if perceived as useful – which varies according to individual differences in information processing. Moreover, their beneficial effect is limited to confusion caused by poor audio quality or language deficiencies. In contrast, they have no effect on confusion resulting from contentual incongruities.

To mitigate negative side effects such as occlusion or distraction from important visual information when captions are not strictly needed, we posit that their activation should be confined to situations in which the user experiences confusion.

2. **Confusion detection:** Facial gestures are able to identify confusion in virtual meetings within a time window as short as five seconds. During non-visual tasks, confusion is reflected in AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU10 (Upper Lip Raiser), AU17 (Chin Raiser), and AU23 (Lip Tightener). When captions demand visual focus, a lower BLINKRATE and less expressions of NEUTRALEMOTION are additional indicators of confusion.

By extracting facial features, confusion can be monitored continuously in order to determine whether activating captions is effective. If comprehension does not improve, the application may revert to auditory-only presentation. The findings further highlight the importance of selecting appropriate features for the detection of confusion depending on whether the presentation is purely auditory or supported by visual material – text in particular. More research is needed to investigate whether complementary measures can further mitigate confusion.

5. Essay 3: EyeDirect: A Gaze Contingent System for Personalized Video Display

> *Imagine coming home from a long day at work and eventually finding the time to watch a movie. Having been truly captivated by the crime story, you tell your co-worker about it and urge her to watch it. The next day, she thanks you for the recommendation. You are surprised to hear that she was intrigued by the victim's construction of an escape plan, but seems to have missed the part where the villain prepared a lethal coup. As you continue discussing the movie, you notice that your co-worker's elaborations on some scenes do not reflect how you remember the plot.*

What happened in this scenario is no coincidence. It is the making of the movie which allows the spectators to direct it themselves as they are watching the film. Without deliberately taking an action, they subconsciously decide about each turn in the plot while the movie evolves. This futuristic vision that changes the way we watch videos and movies could soon become reality. In 2018, Netflix launched the interactive movie 'Black Mirror: Bandersnatch', where viewers make choices about subsequent scenes (Netflix, 2018). As the movie branches out into alternative continuations, viewers may arrive at five different endings, depending on their choices. The movie 'You vs. Wild' and its sequel 'Animals on the Loose: A You vs. Wild Movie', in which the viewer's choices determine protagonist Bear Grylls' survival in nature, were released soon afterwards (Netflix, 2021a, 2021b). Interactive documentaries such as iOtok encourage viewers to interact by clicking on supplementary material or discussing the topic with others (Ducasse et al., 2020). While this allows viewers to control what they see, having them actively select their preferred plot or content contradicts the purpose of movies, which is to provide leisurely entertainment. We thus propose twisting the storyline to the

viewers' taste without them becoming actively engaged – turning the viewer into the subconscious director of the movie.

We developed *eyeDirect*, a system for creating videos with an adaptive storyline based on gaze data. An eye tracker records the viewer's gaze while watching a video. Making use of the relation between gaze distribution and preference for fixated objects, *eyeDirect* selects a variant of the subsequent scene in which the viewer's preferred object takes the lead (cf. Figure 5.1).

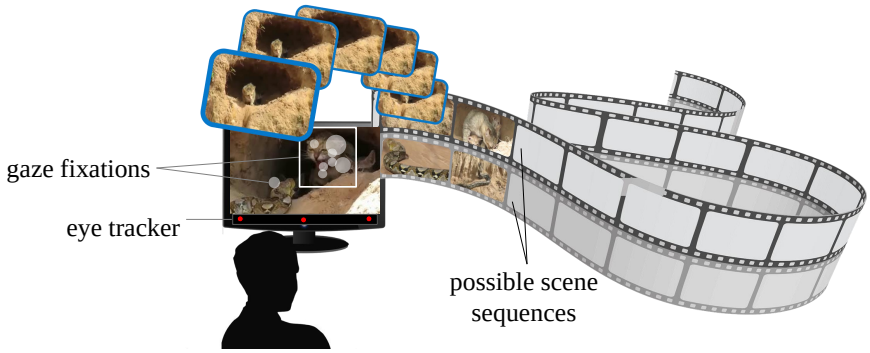


FIGURE 5.1: Illustration of a gaze-informed adaptive movie. The next scene is dynamically selected based on the viewer's distribution of gaze to the objects in the current scene.

We make the following contributions:

- We present *eyeDirect*, a gaze-informed system for adaptive video delivery. The system uses commercial products, and can thus be implemented using readily available technology.
- We conduct a user study (N = 175) to analyze the effect of different personalization strategies on the user engagement attributes *focused attention*, *involvement*, and *novelty*.

The *eyeDirect* system may serve as a reference architecture for designing adaptive videos based on eye tracking data.

5.1. Related work: Gaze-based content adaptation

Gaze-contingent systems presuppose that preferences can be inferred from gaze. Evidence for a correlation between visual attention and preference for elements was already found in the 1960s. Mackworth and Morandi (1967) report that

important aspects of pictures receive more fixations. Engelke and Le Callet (2015) identified a strong correlation between the two concepts. Yet, attention does not necessarily imply preference for an element. While top-down attention is goal-driven, i.e., it is controlled by the person (Burke & Leykin, 2014), bottom-up attention responds to salient elements (e.g., bright colors). Nevertheless, citing the results of several psychological studies, Bednarik (2005) asserts that a link between cognitive processes and eye movements can be assumed. Market research has long made use of this link to analyze the determinants of consumers' decision-making process (Chang & Chen, 2017; Clement et al., 2013; Duchowski, 2002; Hwang & Lee, 2018; Muñoz-Leiva et al., 2019; Otterbring et al., 2014; Wedel & Pieters, 2006). While market research traditionally analyzes gaze data offline, *eyeDirect* seeks to infer preferences dynamically to adapt the content.

In the following, we provide an overview of personalized content retrieval and augmentation with attentive user interfaces in general, before reviewing approaches for video personalization. Table 5.1 summarizes the related literature.

5.1.1. Personalized content retrieval

Personalization requires feedback from users to determine their preferences. Explicit feedback is proactively submitted by the user, for example in the form of product ratings. This potentially alters the natural system usage (Claypool et al., 2001), and preference predictions are only as good as the users' ability to express their preferences (Franke et al., 2009). Implicit feedback therefore extracts information from natural interactions with the system. For instance, online stores use transaction data to recommend relevant products (Bigornia, 2015; Kim et al., 2001; Linden et al., 2003). At Amazon, customers are recommended products that others usually buy together with the items in their shopping cart (Linden et al., 2003). Facebook enables brands to issue advertisements to the accounts of users who show interest in a product on their website, as well as to other Facebook users with similar profiles (Bigornia, 2015).

Since past user data becomes obsolete at a fast pace and often does not adequately represent immediate preferences, dynamic approaches determine the user's psychological state at the moment of the interaction (Matz & Netzer, 2017). Browsing data such as search queries (Lai & Hwang, 2010; Langheinrich et al.,

TABLE 5.1: Overview of related literature on gaze-contingent systems

	ADAPTATION					PREDICTION							
	subject		target			input		method			research output		
Research	image/ document	video	augmentation	search result	plot adaptation	explicit	eye tracking	majority voting	rule-based	machine learning	implementation	efficacy study	
Meißner et al. (2019)	•		•				•	•			•	•	
Starker and Bolt (1990)	•		•				•	•	•		•		
Qvarfordt and Zhai (2005)	•		•				•		•		•	•	
Cheng and Liu (2012)	•			•			•		•		•	•	
Kozma et al. (2009)	•			•			•			•	•		
Salojärvi et al. (2004)	•			•			•			•			
Hardoon et al. (2007)	•			•			•			•			
Xu et al. (2008)	•	•		•			•	•			•		
Netflix (2018, 2021a, 2021b)		•			•	•							
Ducasse et al. (2020)		•	•			•							
Gifreu (2013)		•	•			•							
Peng et al. (2018)		•			•	•							
Bolt (1981)		•	•				•	•			•		
Vertegaal et al. (2003)		•	•				•	•			•		
Kandemir et al. (2010)		•	•				•			•			
Hansen et al. (1995)		•			•		•		•				
Vesterby et al. (2005)		•			•		•	•	•		•		
eyeDirect		•			•		•	•		•	•	•	

1999; Shatnawi & Mohamed, 2012; Trusov et al., 2016) or click data (Claypool et al., 2001; Hauser et al., 2009) have been identified as reliable preference predictors in interactive web applications.

In applications with little user interaction such as video streaming, feedback from browsing data is scarce. Gaze-informed recommender systems thus analyze eye movements to extract relevant images (Cheng & Liu, 2012; Kozma et al., 2009), or predict document relevance from the users' gaze patterns while reading text (Hardoon et al., 2007; Salojärvi et al., 2004). Xu et al. (2008) additionally personalize recommendations for videos. They measure attention to key frames of a scene and then predict attention to other videos with similar key frames. In contrast to *eyeDirect*, attention is measured on the granularity of entire videos, instead of individual elements in a scene.

5.1.2. Attentive user interfaces

Users' visual focus, usually captured by an eye tracking device, has been exploited to augment elements that attract their attention. Starker and Bolt (1990) designed a graphics world inspired by Saint-Exupéry's 'The Little Prince' in which items that the user looks at are augmented. The interactive map iTourist uses gaze data to dynamically create auditory reports for landmarks of interest (Qvarfordt & Zhai, 2005). Meißner et al. (2019) display information for products that users in a virtual reality shopping scenario are looking at.

Using eye tracking to augment dynamic elements that a user looks at was first proposed by Bolt (1981) in a multimedia environment with multiple muted video streams displayed simultaneously on a large screen. Looking at a stream for multiple seconds triggered an adaptation to enlarge the window and play its soundtrack. The video communication system developed by Vertegaal et al. (2003) replays only the speech of whoever the user is currently looking at. Kandemir et al. (2010) display information tags over relevant elements in a video frame.

5.1.3. Personalized videos

Netflix started launching interactive movies in 2017. Throughout the movie, the viewer has to make choices that determine how the plot evolves (Netflix, 2017). Originally targeting children, the multimedia company subsequently launched the interactive movies 'Black Mirror: Bandersnatch' (Netflix, 2018) and 'You vs. Wild' (Netflix, 2021a, 2021b) to cater for a broader audience.
Interactive documentaries display supplementary content if viewers click on an element (Ducasse et al., 2020). They may even become co-authors by entering additional audiovisual or text content while watching the video (Gifreu, 2013).
While such interactive approaches are adaptable and thus offer each user a personalized experience with videos, they are not adaptive. Instead, they require an explicit selection intervention by the user while consuming the video.

The self-reflective animated movie 'A Trip to the Moon' narrates the adventures of a dog traveling to the moon (Peng et al., 2018). The obstacles and social interactions that the dog encounters on the journey vary with the mood and events that the user reports to have experienced throughout the past week. The

user's mood is reflected in the dog's behavior and expressions, as well as in visual and acoustic parameters of the environment. The adaptation does not require any user input at consumption time, but is still dependent on user reported emotions.

Hansen et al. (1995) were the first to suggest using eye tracking for movies with an adaptive storyline. They describe a scene in which two people leave a room in opposite directions. The person to whom the viewer pays more attention is shown in the next scene. The conceptual framework for gaze-informed interactive movies was implemented in later work by Vesterby et al. (2005).

While the reviewed systems demonstrate that it is feasible to create gaze-informed adaptive movies, none of them has been both implemented and evaluated with regard to its usability. In the following, we thus present *eyeDirect*, a system for dynamic video adaptation. The system was implemented and used to assess the effect of different preference inference strategies on user engagement.

5.2. The eyeDirect system

EyeDirect is a system for creating adaptive movies with standard eye tracking equipment and classification algorithms. The system analyzes gaze data collected at runtime and selects the next scene based on the viewer's preferences for the objects in the current scene (cf. Figure 5.2). It may serve as a reference architecture for creating gaze-informed adaptive videos.

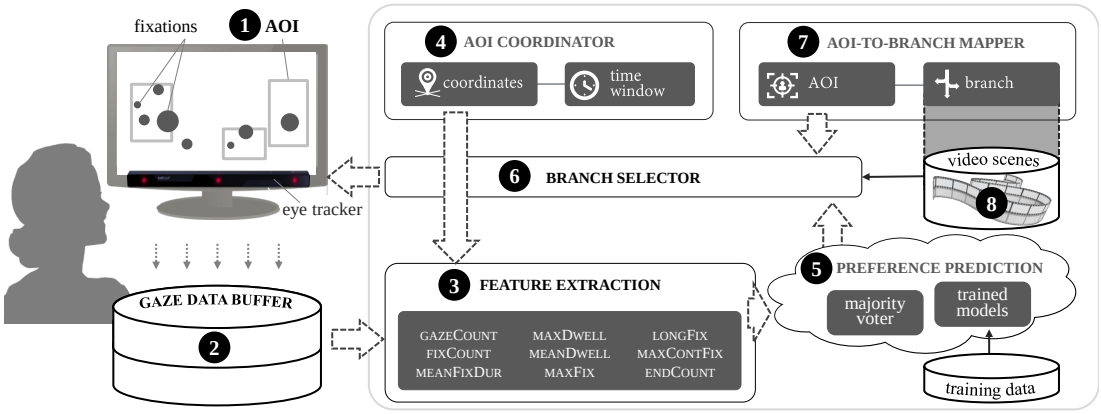


FIGURE 5.2: The *eyeDirect* system architecture for personalized videos. The system analyzes real time gaze data to predict user preferences for elements in a video. The next scene for the video is selected based on the predicted preference.

5.2.1. Gaze data collection

Gaze data is recorded during preference indicative sequences in which all ❶ *areas of interest (AOIs)* are visible. An AOI is a rectangular region containing an element about which the system makes an inference. It may encompass one video of a tiled screen, or a single object like a person. For moving objects, the definition of AOIs can be automated with object tracking (Wang & Yeung, 2013).
A screen-based eye tracker starts recording the viewer's gaze at the beginning of a scene and continuously forwards the gaze data to the ❷ *gaze data buffer*. The buffer stores the raw gaze data for the entire scene. Gaze recording stops two seconds before the scene ends to ensure that data processing does not delay the display of the next scene.

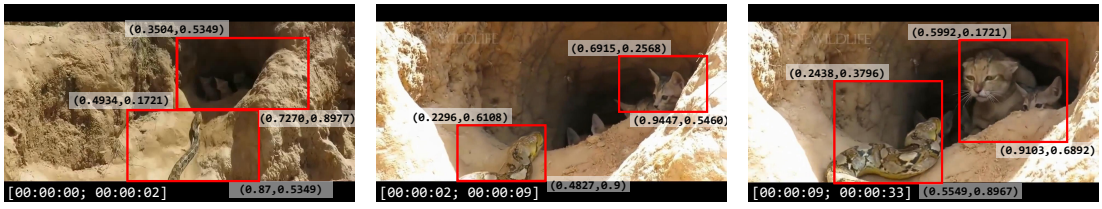
5.2.2. Feature extraction

At the end of each scene, the ❸ *feature extraction* module processes the raw data accumulated in the gaze data buffer into a standard set of gaze features. Applying the strict average method (Olsen, 2012), invalid gaze recordings are discarded, and the gaze coordinates of the left and right eye are averaged. The gaze points are then processed into fixations and saccades. During fixations, the viewer focuses on a specific point on the screen. Using the values recommended by the manufacturer of the Tobii eye tracker (Olsen, 2012), we defined fixations as eye movements with a velocity below a visual angle of 30° per second. Appendix B.1 contains the pseudocode of the algorithm that was used to extract fixations. As they are typically associated with the acquisition of information, fixations can be leveraged for preference predictions (Cheng & Liu, 2012; Jacob, 1991; Majaranta & Bulling, 2014). During the saccadic jumps between fixations, in contrast, only little information is absorbed (Joachims et al., 2017b). We therefore discard saccades and only further process the extracted fixations. The set of included features comprises one sequential, four duration, and four frequency features that are commonly used in gaze-contingent systems. Frequency and duration features ignore the chronological order of the recorded gaze data. In contrast, sequential features take into account changes in the gaze patterns over time. Table 5.2 describes the features. The methods for their calculation in pseudocode are included in Appendix B.2.

TABLE 5.2: Frequency, duration, and sequential gaze features. Frequency and duration features are time invariant. Sequential features take into account changes in the gaze patterns over time.

Frequency Features	**Description**
gazeCount	Total number of valid gaze points
fixCount	Total number of fixations
longFix	Number of fixations longer than 400ms *(threshold from Tobii (Olsen, 2012))*
maxContFix	Maximum number of continuous fixations on one AOI
Duration Features	
meanFixDur	Mean duration of a fixation (in ms)
maxDwell	Maximum time that the gaze dwells on one AOI (in ms)
meanDwell	Mean time that the gaze dwells on one AOI (in ms)
maxFix	Duration of the longest fixation (in ms)
Sequential Features	
endCount	Number of gaze points from the last 400ms

The feature extraction module then examines the distribution of gaze across the AOIs. Information about the screen coordinates of the AOIs is retrieved from the ❹ *AOI coordinator*. Since the coordinates of a moving target may change within a scene, this module associates each AOI with its corresponding time window (defined in terms of milliseconds from scene start). An example of a video with dynamically changing AOI coordinates is illustrated in Figure 5.3.



(a) AOIs for $t_{00} - t_{02}$ **(b)** AOIs for $t_{02} - t_{09}$ **(c)** AOIs for $t_{09} - t_{33}$

FIGURE 5.3: Example of a video with dynamic areas of interest. AOI coordinates are stored along with their time intervals (defined in milliseconds from scene start).

5.2.3. Preference prediction

The extracted gaze features are forwarded to the ❺ *preference prediction* module which determines the user's preferred AOI.

We integrate and test two prediction strategies. The first strategy uses majority voting to analyze the distribution of a single feature over the AOIs. The viewer's preferred object is defined by the AOI with the highest value for the gaze feature.

Past research has shown that several properties of images including size, color, and placement of visual elements bias gaze distribution (Buscher et al., 2009; Engel, 1974; Mackworth & Morandi, 1967). The second strategy therefore tests whether training machine learning models with gaze data from subjects who have previously watched the same scenes improves the effectiveness of the system.

5.2.4. Branch selection

Different rationales can be applied to map visual objects to alternative sequences of the plot. One possibility is to manipulate the chronology so that scenes that have a high correlation with the viewer's preferences are displayed first (Verma et al., 2021). This may prove beneficial for longer movies and videos with cognitively demanding content, where concentration tends to dwindle over time.
In contrast, *eyeDirect* implements alternative branching. Initially, all users watch the same segment. Depending on the recorded gaze, the storyline then unfolds into one out of multiple alternative narratives so that, for each user, their most appreciated object takes the center stage (Hansen et al., 1995; Vesterby et al., 2005). The corresponding variant of the scene is retrieved by the ❻ *branch selector*. It issues a query to the ❼ *AOI-to-branch mapper*. This module holds a tree-based record of all AOIs with a link to the next scene branch. The mapping of AOIs to branches is a priori specified by the movie director. When the scene ends, the retrieved branch is displayed to the user.

While alternative branching may result in a complex movie structure, interactive movies like 'Black Mirror: Bandersnatch' (Netflix, 2018) and the 'You vs. Wild' series (Netflix, 2021a, 2021b) have demonstrated that broadcasters possess both the resources and the capabilities to create such films. Vesterby et al. (2005) suggest to maintain the complexity of the video manageable by defining basing points at which alternative branches converge.
To implement such a structure, the adaptive movie should contain two types of ❽ *video scenes*. Scenes on the main branch are preference indicative. During these scenes, gaze is recorded to determine the viewer's preferences for the displayed objects. Scenes on alternative branches are retrieved in response to the viewer's gaze behavior. Alternative scenes that branch out from the same basing point should be substantially different and be logically connected to the preceding and

subsequent scene. Audio cross-fading or volume adjustment can therefore only be applied within the margins of a scene. A branch switch is always accompanied by a change of the camera angle in order for the shift to appear natural. Each alternative branch should be perceived as the natural continuation of the preceding scene. From a technical viewpoint, alternative branches are self-contained units that are independent of all previous and subsequent scenes.

In case no clear preference is detected, the movie director should specify a behavior for exception handling. The viewer may, for instance, be directed to a generic scene on the main branch in which all elements appear. Alternatively, one of the scenes on alternative branches can be defined as the default. Extending *eyeDirect* with additional sensor input such as bio-signals or facial expressions offers further possibilities. By making use of the additional feedback, the viewer's contentment with a scene can be detected even if gaze predictions are inconclusive.

5.3. Study design: Content filtering with eyeDirect

To research how content filtering in cinematographic applications influences engagement, we designed three adaptive videos and implemented them in *eyeDirect*. We tested two adaptation strategies that are based on a causal model in which the relationship between gaze-informed preference predictions and engagement is mediated by the degree of personalization (cf. Figure 5.4).

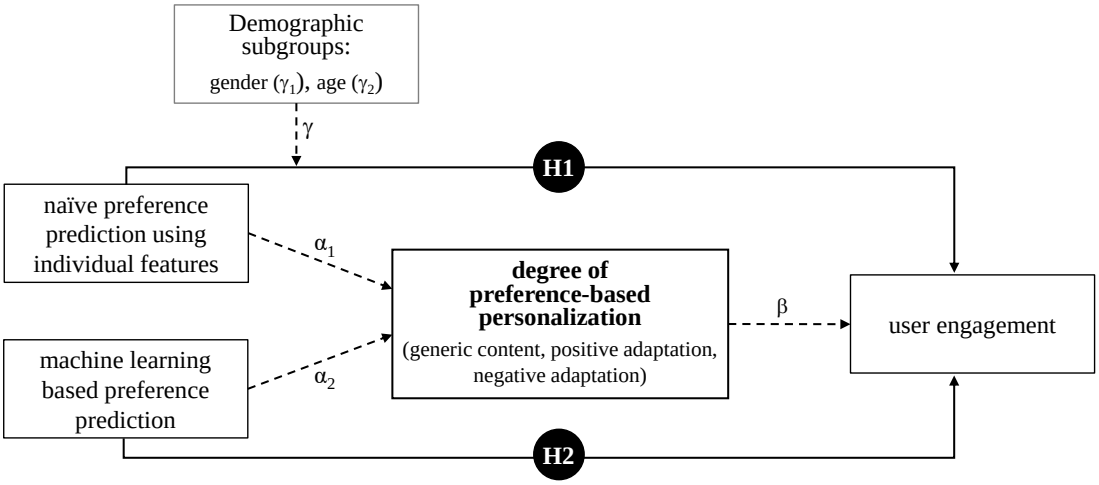


FIGURE 5.4: Conceptual framework of preference-based personalization. Causal model showing the hypothesized relationships between preference predictions and engagement.

Empirical evidence from web applications suggests that users prefer personalized content and services (Ho, 2006; Tam & Ho, 2005; Tran, 2017). We expect that a similar positive effect on engagement can be achieved by personalizing videos (β).

A second fundamental assumption of the system is that recording and processing the gaze data of a person watching a video allows to predict which objects in the visual scene they prefer. Based on this assumption, the vision paper by Hansen et al. (1995) suggests a rule-based adaptation. Personalized branching is initiated whenever attention to an object relative to other visible objects – measured by the distribution of gaze data – exceeds a certain threshold. Vesterby et al. (2005) implemented two rule-based adaptation strategies: The 'winner-takes-it-all' strategy selects the object that correlates with the highest number of recorded gaze points. The 'weight-decaying-interest' strategy assigns higher importance to gaze points that are recorded later in a scene. All these strategies are based on majority voting. Yet, their effect on engagement has not been evaluated so far. In a first step, we therefore test the effectiveness of adapting videos based on a 'winner-takes-it-all' majority voting (α_1). Since demographic variables often influence cognitive processes, we additionally test whether the effectiveness of the strategy differs between individuals depending on their gender (γ_1) and age (γ_2).

State-of-the-art adaptive systems mostly use machine learning to predict user preferences (Hardoon et al., 2007; Kandemir et al., 2010; Kozma et al., 2009; Salojärvi et al., 2004; Schweikert et al., 2018). While these models require prior scene-specific data collection, they are able to filter out saliency bias. Such a bias occurs when attention to objects is steered by visually striking properties such as colors and shapes, instead of preferences (Buscher et al., 2009; Engel, 1974; Mackworth & Morandi, 1967). We thus test whether adaptation to preferences predicted by machine learning is more effective than the majority voting that has been proposed in previous research (α_2).

When showing participants a video continuation in which only one of the objects is present, any subsequent preference statements are likely to be biased. There is a good chance that subjects will indicate a preference for the continued object simply as a result of the prolonged exposure to it. Therefore, we cannot simultaneously evaluate whether an adaptation strategy leads to accurate preference predictions (β) and whether it indeed has an effect on engagement (α). Instead, we evaluate how adapting a video based on to the *predicted* preference affects en-

gagement. We make no claims regarding the effectiveness of video personalization based on the viewer's *true* preferences. We formulate the following hypotheses:

> **H 3A.** Personalizing videos by applying majority voting to gaze features (naïve prediction) increases engagement.
>
> **H 3B.** Personalizing videos by applying machine learning to gaze features increases engagement.

To test the hypotheses, we conducted a user study in which we recorded the participants' gaze while they were watching three videos from the genres *Music*, *Tutorial*, and *Documentary*. An experimental between-subjects design with one treatment group and one control group was adopted. In the treatment group, one out of two possible continuations of the video was chosen at random. Thus, some participants experienced *positive adaptation (P)*. They were shown a continuation in which their preferred object was the central figure. Others experienced *negative adaptation (N)* and continued watching a scene of their least preferred object. In the control group, no branching was initiated. Instead, the subjects were shown a *generic (G)* continuation featuring both video elements. We then compared the engagement ratings of participants watching a video with positive adaptation to (1) negative adaptation, and (2) a non-adaptive generic video.

5.3.1. Participants

In total, 175 members and visitors of our research institution (93 male, mean age: 25.35 $\pm$ 8.39) participated in the user study. The experiment took about five minutes for each participant. The subjects were assigned quasi-randomly to an experimental group according to the date of their arrival. They were informed that their gaze was being monitored, but they were not aware of its use for dynamic video adaptation.

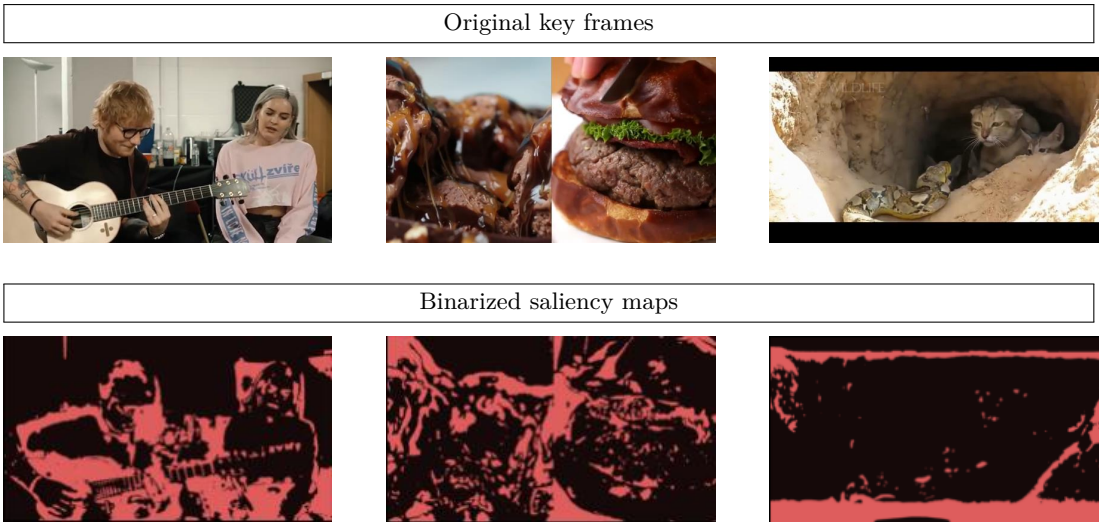
5.3.2. Apparatus

The software ran on a 2.5GHz Intel Core i7-6500 HP PC. The PC drove a 23.6 inch iiyama touch monitor (1920 x 1080 pixels) on which the videos were displayed.

Gaze data was collected with a Tobii Pro X3-120 eye tracker that was mounted to the display. The experiment was performed in our research lab. Participants were seated in front of the display, maintaining a distance between their eyes and the eye tracker of 20-30 inches.

5.3.3. Video material

The experiment consisted of three videos from the genres *Music*, *Tutorial*, and *Documentary*. The categories were chosen to reflect the content that is commonly available on video streaming platforms (cf. Figure 5.5).



(I) Music. Live performance of the song '2002' by Ed Sheeran (AOI_1) and Anne Marie (AOI_2).

(II) Tutorial. Tiled screen with two parallel cooking tutorials for a dessert (AOI_1) and a burger (AOI_2).

(III) Documentary. Filming of a snake (AOI_1) attacking a wildcat mother and her offspring (AOI_2).

FIGURE 5.5: Video material for evaluating gaze-informed content adaptation. Three videos were used from the genres *Music* (Nicholson & Ed Sheeran, 2018), *Tutorial* (Tasty, 2016, 2019), and *Documentary* (Best Of Wildlife, 2019). The binarized saliency maps confirm that the visual saliency factor is comparable for the two AOIs in all three videos.

Each video features two main elements which were defined as AOIs. We used two fullscreen videos (I, III) and one tiled video (II). The latter shows two video streams in parallel. Such a design allows, for instance, to simultaneously present multiple trailer previews on television or on online video streaming platforms. An adaptive system may subsequently display a teaser scene of the movie that received the most attention.

In order for a video to produce gaze patterns that are indicative of the viewer's preferences, it should have clearly differentiated objects with similar visual saliency factors. This ensures that attention to an element is effectively a result of the viewer's interest, and is not biased by the properties of the visual material. We therefore analyzed multiple key sequences in the three videos using a Python implementation of visual saliency maps following the design by Itti et al. (1988)[6]. From the exemplary binarized saliency maps in Figure 5.5 it can be seen that the two main objects have similar saliency factors in all three videos.

5.3.4. Procedure

At the beginning of the experiment, an eye tracker calibration with five calibration points was performed. The participants were asked to fixate five red circles that successively appeared on the screen. Completing the procedure required successful eye detection. If the eye tracker did not detect the subject's eyes, its orientation was adjusted. This ensured that gaze data was collected for all participants.

The participants were instructed to sit in a comfortable position and watch the three video sequences. By not constraining the subjects' freedom of movement, we obtain more realistic estimates of the system's performance in an uncontrolled setting. The order in which the sequences were displayed was identical for all participants. This experimental setup was chosen to maximize the uniformity of conditions between the experimental groups.

Gaze was recorded during the first part of the video, which was identical for all groups. In the first sequence, the subjects were shown a live performance of the song '2002' by Ed Sheeran and Anne-Marie (2018). After 39 seconds, the video branched out into alternative scenes. The treatment group was shown a solo performance of one of the two singers. The control group continued watching the duet. The second sequence displayed two video streams of cooking tutorials in parallel, one for a dessert (Tasty, 2019) and one for a burger (Tasty, 2016). After 28 seconds, the treatment group was shown a video of another recipe for either one of the food categories. The control group was again shown two streams in parallel. The third sequence showed a snake attacking a wildcat family (Best Of Wildlife, 2019). The video halted after 33 seconds, when the wildcat was just

[6] `pySaliencyMap`: https://github.com/akisatok/pySaliencyMap/ (Kimura, 2020)

about to launch a counterattack on the snake. The treatment group was shown the fate of one opponent. Participants assigned to the control group watched the outcome of the fight and followed the paths of both animals.

Directly after watching each video, the participants were presented an on-screen questionnaire to rate their degree of engagement when watching the respective video. At the end of the experiment, the accuracy of the gaze recordings was validated. Each subject was asked to fixate five circles that appeared in each corner and in the center of the display while their gaze was recorded. For validation, the gaze data was compared to the screen coordinates of the visual stimuli. Afterwards, they were presented with an additional on-screen questionnaire in which they were asked to recall their preferred element for each of the three videos.

5.3.5. Metrics

We adapted the User Engagement Scale by O'Brien and Toms (2010) to measure engagement with videos. All items were measured on a six-point Likert scale.

TABLE 5.3: List of items used to measure engagement. The items measuring *novelty*, *focused attention*, and *involvement* were adapted from the User Engagement Scale by O'Brien and Toms (2010).

Item		Construct
Q-1	I would have liked to watch the video for another minute.	**Novelty**
Q-2	I found the video interesting.	
Q-3	I was very attentive.	**Focused attention**
Q-4	I found the video enjoyable.	**Involvement**

The first two items (Q-1, Q-2) measure the construct *novelty*, which O'Brien and Toms (2010) define as a *"variety of sudden and unexpected changes* [...] *that cause excitement and joy or alarm"*. The third item (Q-3) is a measure of *focused attention*, which implies *"concentrating on one stimulus only and ignoring all others"* (O'Brien & Toms, 2010). Item Q-4 measures *involvement*, defined as a *"need-based* [...] *psychological identification with some object"* (O'Brien & Toms, 2010). Table 5.3 lists the survey items.

5.3.6. Data analysis

In an offline evaluation, we simulated two adaptation strategies and assessed their effect on engagement.

Strategy A (NAIVE PREDICTION). The viewer's preferred object in a video is predicted by applying majority voting to individual gaze features. Based on this prediction, we categorized the samples into two groups. Subjects assigned to the positive adaptation group had watched the video continuation that matches their predicted preference. Subjects of the negative adaptation group had watched the continuation that does not match their predicted preference. Finally, we compared the engagement ratings resulting from positive adaptation to those from negative adaptation and to the generic control group.

Strategy B (MACHINE LEARNING). Machine learning might be able to capture interdependencies between objects in a video and the systematic saliency bias. We thus examined whether adaptation with machine learning has a stronger effect on engagement than the naïve approach. To test this strategy, we predicted the viewer's preferred element in the video using machine learning classifiers. This can result in a different assignment of a subject to the positive or negative adaptation group. We assessed the effectiveness of personalization with machine learning by again comparing the engagement ratings from the three experimental groups.

Figure 5.6 illustrates how the participants were assigned to an experimental group.

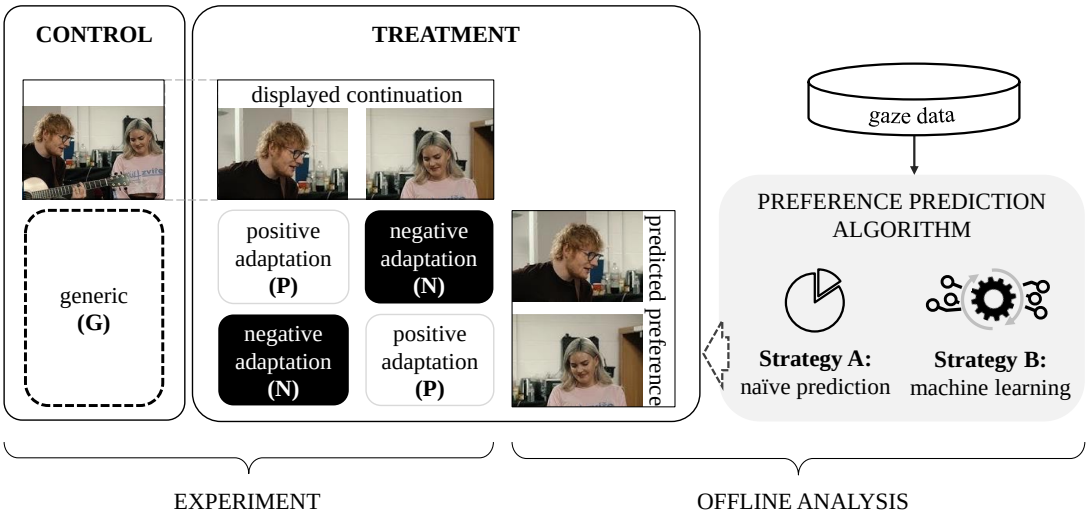


FIGURE 5.6: Procedure for dividing samples into experimental groups. Samples where the predicted preference matches (does not match) the displayed continuation were assigned to the positive (negative) adaptation group.

5.4. Results: Usability & efficacy of content filtering

After excluding 29 samples due to insufficient quality of the gaze data, we retained valid recordings from 146 participants. The gaze accuracy of the retained samples was sufficient for assigning the data to an AOI (mean error $< .9°$ visual angle). 51 participants pertained to the control group that watched a generic scene. The other 95 participants had been exposed to either positive or negative adaptation.

Using the aggregated data from all three videos, we collected on average 10,120 valid gaze points from each participant. An average of 2,641 data points did not contain valid gaze coordinates. Figure 5.7 shows the statistical distributions of the total number of gaze points and additional descriptive measures of the recorded gaze data. The number of transitions from one AOI to another while watching one of the three videos ranges from 4 (Video III) to 83 (Video II), with a mean value of 27.24 across all videos and participants.

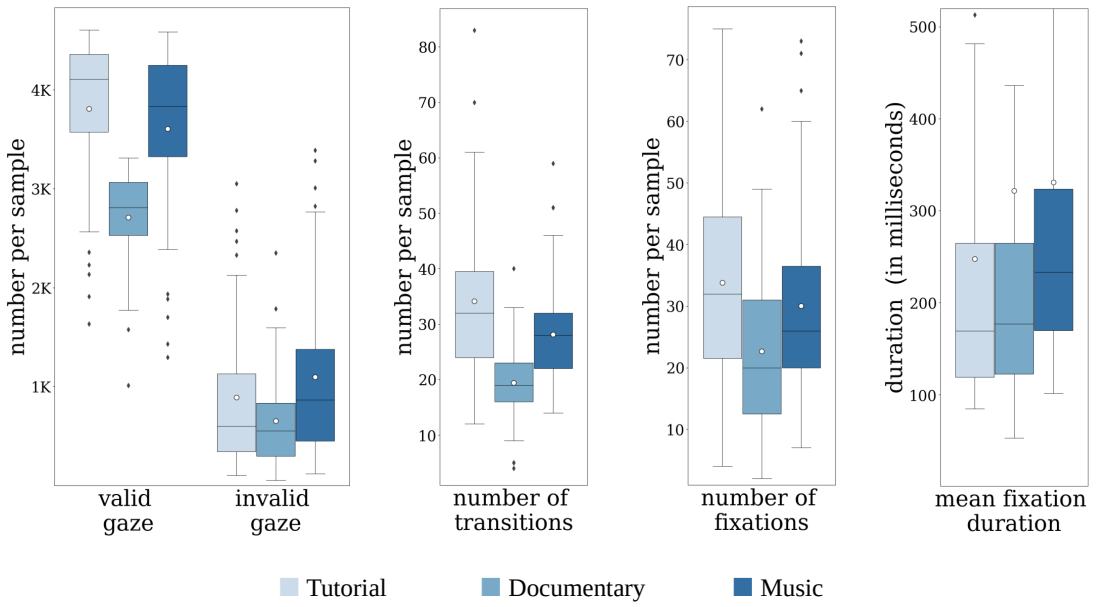


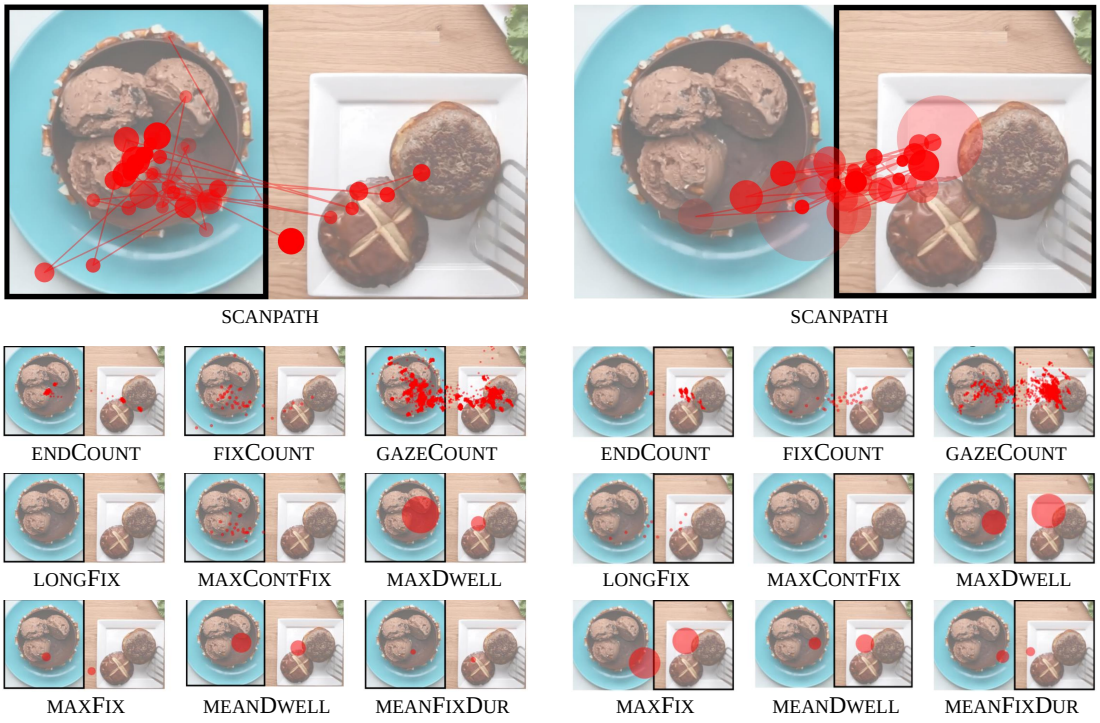
FIGURE 5.7: Descriptive measures of the collected gaze samples per video. The videos produce distinct patterns with regard to the number of valid and invalid gaze points, transitions, as well as the number and duration of extracted fixations. Mean values are visualized as white circles.

From the raw gaze data recorded while the participants were watching all three videos, we extracted an aggregated number of on average 86 fixations per participant with a mean duration of 300 milliseconds. This number is slightly above the average fixation duration of about 200 milliseconds which is usually cited in eye

tracking experiments (Salthouse & Ellis, 1980). The deviation is a direct result of our rigorous data preprocessing, where all fixations with a duration below 60 milliseconds were dropped. We implemented this filter in order to remove very short fixations during which no information is processed (Galley et al., 2015).

5.4.1. Strategy A (naïve prediction)

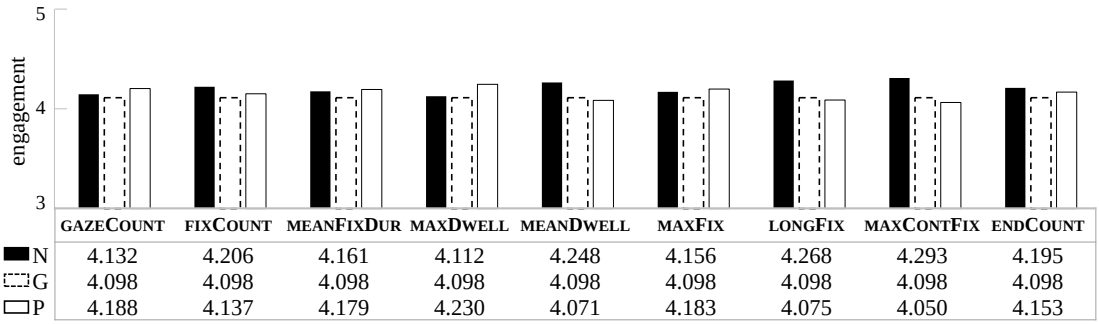
In a first step, we predicted the viewers' preferences based on individual gaze features. This strategy selects the AOI for which the feature returns the highest value, computed for a participant. Figure 5.8 depicts two exemplary visualizations of the scanpath (i.e., the sequential gaze distribution) of participants who reported that they preferred the object in the left (Figure 5.8a) or the right (Figure 5.8b) part of the screen, respectively. Additionally, the gaze features extracted from the raw gaze data are visualized for each participant. In both cases, more gaze points were recorded on the preferred object.



(a) reported preference: *Dessert* **(b)** reported preference: *Burger*

FIGURE 5.8: Visualization of gaze features extracted from two exemplary samples. Circle sizes indicate the duration of a fixation at the corresponding screen coordinates. Early fixations appear faded to visualize sequential gaze distribution.

The main advantage of the naïve prediction is that no further information is needed to make the decision (in contrast to machine learning which requires prior data collection). To evaluate the effectiveness of the strategy, the participants were assigned to either the positive or the negative adaptation group depending on whether, based on the prediction, they had watched a scene with their preferred object or not. We then calculated the average engagement **Q** for each group (including the generic control group) as the mean rating from all engagement items (**Q-1**, **Q-2**, **Q-3**, **Q-4**).



	gazeCount	fixCount	meanFixDur	maxDwell	meanDwell	maxFix	longFix	maxContFix	endCount
N	4.132	4.206	4.161	4.112	4.248	4.156	4.268	4.293	4.195
G	4.098	4.098	4.098	4.098	4.098	4.098	4.098	4.098	4.098
P	4.188	4.137	4.179	4.230	4.071	4.183	4.075	4.050	4.153

FIGURE 5.9: Effect of majority voting personalization on engagement. Illustrated is the reported engagement for gaze-assigned experimental groups. Groups were determined through majority voting based on the values of the specified feature (**N**: negative adaptation, **G**: generic, **P**: positive adaptation). Comparisons are based on mean scores from all engagement items (Q-1, Q-2, Q-3, Q-4).

The distribution of the engagement ratings in Figure 5.9 suggests that the effect of the adaptation differs depending on the gaze feature. Six features (gazeCount, fixCount, meanFixDur, maxDwell, endCount, maxFix) make predictions that result in higher engagement for the positive adaptation group compared to the generic control group. The effect is reversed when predicting preferences with the remaining three features. Compared to the negative adaptation group, only personalization with four features (gazeCount, meanFixDur, maxDwell, maxFix) results in higher engagement. However, a one-sided t-test revealed that the naïve prediction strategy has no significant positive effect on engagement at any common level of confidence, independent of the applied gaze feature. Table 5.4 summarizes the engagement ratings from all three experimental groups by group determination with each of the extracted gaze features.

Effect within demographic subgroups. Since the gaze behavior varies across people of different age and gender (Slessor et al., 2010; Sullivan et al., 2015), we investigated whether naïve predictions are effective at least for some demographic

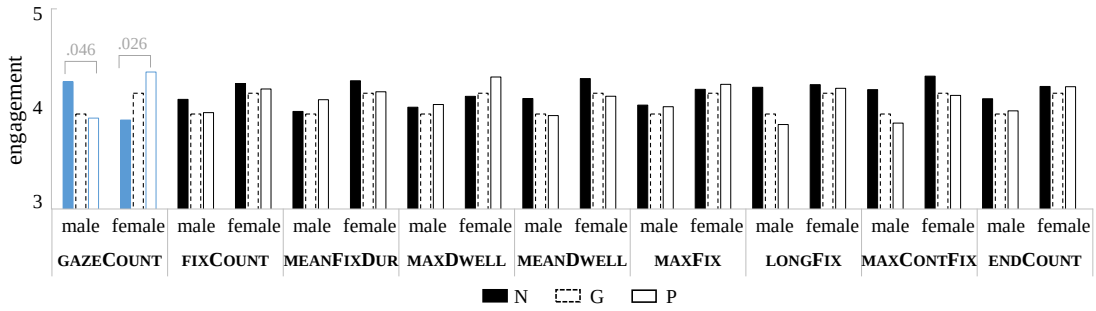
TABLE 5.4: Impact of gaze-based video adaptation from majority voting on engagement. Experimental groups were determined through majority voting based on the values of the specified feature (**P**: positive adaptation, **N**: negative adaptation, **G**: generic). Engagement **Q** is calculated as the mean value from all reported engagement items (Q-1, Q-2, Q-3, Q-4). Statistical significance of the differences between the experimental groups is determined with one-sided t-tests.

	engagement Q (μ)			**P-N**			**P-G**		
Gaze feature	**P**	**N**	**G**	**Δ Q**	**t**	**sig.**	**Δ Q**	**t**	**sig.**
gazeCount	4.188	4.132	4.098	.057	0.354	.361	.090	0.640	.261
fixCount	4.137	4.206	4.098	-.069	-0.466	.679	.039	0.255	.399
meanFixDur	4.179	4.161	4.098	.018	0.123	.451	.081	0.533	.297
maxDwell	4.230	4.112	4.098	.118	0.796	.213	.132	0.861	.195
meanDwell	4.071	4.248	4.098	-.177	-1.173	.880	-.027	-0.165	.566
maxFix	4.183	4.156	4.098	.027	0.185	.427	.085	0.562	.287
longFix	4.075	4.268	4.098	-.193	-1.307	.904	-.023	-0.152	.561
maxContFix	4.050	4.293	4.098	-.243	-1.643	.950	-.048	-0.310	.622
endCount	4.153	4.195	4.098	-.042	-0.281	.611	.055	0.374	.354

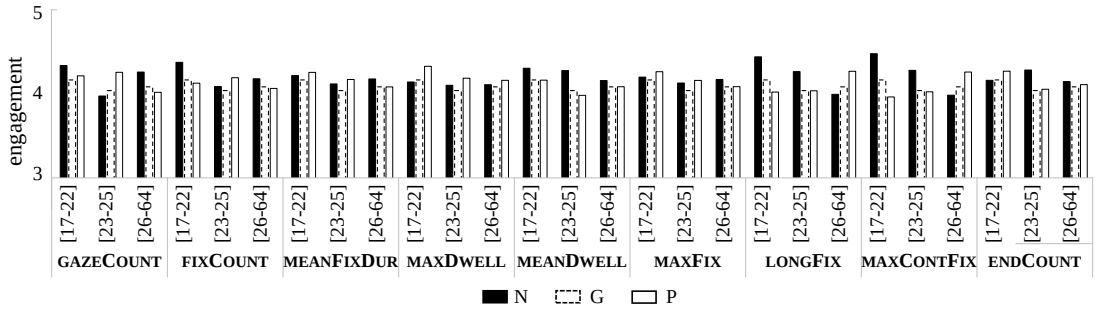
subgroups. The participants were divided into three age groups of approximately equal size: 17-22 (N = 66), 23-25 (N = 63), and 26-64 (N = 46). Participants who identified not as female (N = 79) or male (N = 93) were excluded from the gender analysis due to their small number in the sample. Differences in engagement ratings between the experimental groups were analyzed with a one-sided t-test.

A two-sided t-test revealed that, for females, positive adaptation leads to higher engagement than negative adaptation when majority voting is applied to gazeCount (t = 1.981, p-value = .051). Male participants, in contrast, experience marginally lower engagement when positive (versus negative) adaptation is applied to gazeCount (t = 1.699, p-value = .093). In the overall population, the two opposing effects offset each other. T-tests performed on the remaining features revealed no significant effect. Figure 5.10a summarizes the descriptive statistics. The subgroup analysis implies that, by taking into account the viewer's gender, it may be possible to effectively personalize videos by applying simple majority voting to gazeCount. Since, however, only one gaze feature resulted in higher engagement when applying the naïve prediction strategy, the effect may be conditional on the video material. Additional tests with a variety of videos are necessary to validate the moderating effect of gender.

The analysis of age groups gave no indication for a beneficial effect of the strategy in a subpart of the population (cf. Figure 5.10b). A two-sided t-test revealed no significant improvement in engagement at any common level of significance.



(a) engagement by gender



(b) engagement by age

FIGURE 5.10: Demographic subgroup analysis of the effect of personalization. Illustrated is the reported engagement for the three gaze-assigned experimental groups by demographic variables. Groups were determined through majority voting based on the values of the specified feature (**N**: negative adaptation, **G**: generic, **P**: positive adaptation). Comparisons are based on mean scores from all engagement items (Q-1, Q-2, Q-3, Q-4).

Validity of preference predictions. Based on the findings from the unmoderated hypothesis testing, we reject **H 3A**. Two possible factors could be the cause for this outcome: (1) The preference predictions from the naïve strategy (α_1) may be inaccurate; (2) video personalization may have no effect on engagement (β).

To determine whether inaccurate preference predictions are the primary cause, we verified the accuracy of the naïve predictions. The predicted preferences for the treatment group cannot be objectively verified, because the personalized video delivery may result in biased a posteriori preference statements. Subjects of the control group, in contrast, were exposed to all objects in the video to the same degree. Their reported preferences from the follow-up questionnaire are therefore not corrupted by varying exposure times. We thus applied majority voting to the gaze features extracted from the control group (N = 51) and predicted the participants' preferred object in each video. We then compared the prediction to the reported preference.

Table 5.5 summarizes the distribution of gaze points in videos I - III by reported preference. The distributions indicate how dominant (in percent) a feature was on the left AOI. The values are calculated as the sample mean from all participants of the generic control group who stated a preference for the left (AOI_1) or right (AOI_2) object, respectively. A two-sided t-test was used to identify whether the gaze distribution differs significantly between the two groups.

TABLE 5.5: Gaze distribution from the generic control group. Gaze allocation (in %) is the average value of a feature that is attributed to the left object (AOI_1), calculated from all subjects. Numbers in *italic* indicate inverse preferences, i.e., the feature is more dominant on AOI_1 in the group that reported a preference for the other object (AOI_2). Accuracy is calculated as the percentage of correct predictions from applying majority voting to the indicated gaze feature. Groups are based on reported preferences. Significant group differences were identified with a one-sided t-test (significance level: *p < .1%, **p < .05, ***p < .01).

		gaze features								
Video	**Metric**	gazeCount	fixCount	meanFixDur	maxDwell	meanDwell	maxFix	longFix	maxContFix	endCount
Music	% gaze: prefer AOI_1 (N=32)	53.9	64.7	59.8	*44.8*	*43.6*	64.6	58.3	65.8	58.9
	% gaze: prefer AOI_2 (N=19)	44.3	53.0	54.1	*54.4*	*52.4*	54.2	43.5	49.4	42.2
	t-test	2.58 (**)	2.71 (**)	1.24	-1.69 (*)	-1.67	1.60	1.40	2.32 (**)	1.78 (*)
	accuracy (in %)	52.9	64.7	66.7	43.1	39.2	54.9	68.6	62.7	60.8
Tutorial	% gaze: prefer AOI_1 (N=30)	59.2	63.1	*53.3*	*41.3*	*40.3*	*57.3*	*41.7*	68.1	74.3
	% gaze: prefer AOI_2 (N=21)	49.7	60.2	*55.2*	*46.4*	*49.1*	*60.5*	*48.2*	63.1	51.8
	t-test	2.76 (***)	0.63	-0.62	-0.95	-1.84 (*)	-0.64	-0.58	0.69	3.37 (***)
	accuracy (in %)	62.7	62.7	54.9	45.1	35.3	51.0	56.9	58.8	70.6
Documentary	% gaze: prefer AOI_1 (N=12)	50.4	53.6	44.9	52.4	48.3	45.6	25.0	*8.3*	44.2
	% gaze: prefer AOI_2 (N=39)	41.0	50.5	38.2	46.6	41.4	41.9	25.0	*12.3*	34.8
	t-test	1.16	0.27	0.62	2.11 (*)	1.85	0.33	0.00	-0.41	1.17
	accuracy (in %)	64.7	52.9	62.7	64.7	60.8	60.8	39.2	23.5	68.6
TOTAL	**accuracy (in %)**	**60.1**	**60.1**	**61.4**	**51.0**	**45.1**	**55.6**	**54.9**	**48.4**	**66.7**

In order for a gaze feature to be a valid preference indicator, its dominance on AOI_1 should be higher for participants who also selected this element as their preferred object, compared to the other group. In some videos, we found deviations from this rationale for the duration features meanFixDur, maxDwell, meanDwell, maxFix, longFix, and maxContFix. In contrast, the frequency

features GAZECOUNT and FIXCOUNT, and the sequential feature ENDCOUNT were consistently more dominant on AOI_1 in the group that also favored the object. This suggests that viewers look more often at their preferred object, especially towards the end of a scene when they have established their preference. The duration of individual dwells, in contrast, is not linked to preferences.
However, a one-sided t-test revealed that only the *Music* video produced a significant effect on GAZECOUNT (t = 2.58, p-value = .014), ENDCOUNT (t = 1.78, p-value = .084), and FIXCOUNT (t = 2.71, p-value = .010). For the *Tutorial*, the differences in GAZECOUNT (t = 2.76, p-value = .008) and ENDCOUNT (t = 3.37, p-value = .002) were significant. The *Documentary* produced no significant effect.

The insights from the statistical tests are mirrored in the accuracy of the predictions. The mean accuracy across all videos ranges from 45.10% (MEANDWELL) to 66.67% (ENDCOUNT), and is even below the random baseline of 50% for some features. This confirms that the ineffectiveness of the naïve strategy can – at least in part – be attributed to inaccurate preference predictions. Consequently, the assumption α_1 which stipulates that majority voting allows to predict preferences cannot be sustained. We therefore proceeded to investigate whether machine learning is more effective than the naïve strategy with majority voting.

5.4.2. Strategy B (machine learning)

We evaluated whether machine learning can effectively adapt a video to the viewer's preferences. The classifiers were trained exclusively with data from the control group where no bias was induced by prolonged exposure to an object. We tested five classifiers from the `scikit-learn`[7] Python library which are commonly used to infer preferences: Logistic regression (e.g., Kandemir et al., 2010; Kozma et al., 2009; Slanzi et al., 2017), passive aggressive (e.g., Slanzi et al., 2017), k-nearest neighbor (k-NN) (Kazienko & Adamski, 2007; Ribeiro-Neto et al., 2005; Rothrock et al., 2002), decision tree (e.g., Kim et al., 2001; Rothrock et al., 2002), and SVM (e.g., Hardoon et al., 2007; Horvitz, 1999; Joachims et al., 2017a; Radlinski & Joachims, 2005; Salojärvi et al., 2004; Slanzi et al., 2017). The models were trained for each video individually. This allows them to take into account that some objects draw attention due to salient properties, and not as a result of

[7] `scikit-learn`: https://scikit-learn.org/stable/modules/classes.html (Pedregosa et al., 2011).

the viewer's preference.

We identified the best feature subset by training each classifier with all possible combinations of gaze features on a randomly selected subset of 80% of the training data. Table 5.6 reports the classification accuracy (A), precision (P), recall (R), and F1-score (F1) resulting from 5-fold cross-validation. A training set was randomly drawn in each fold, consisting of a subset of 80% of the samples from the video. Class predictions were made on the remaining 20% of the samples.

TABLE 5.6: Best feature subsets and performance evaluation of preference prediction with machine learning. We report accuracy (A), precision (P), recall (R), and F1 scores (F1). Total performance is calculated as the mean value of the individually trained classifiers for each video. The best feature subset therefore includes all features that are included in one of the models.

		best feature subset									**performance (in %)**			
Video	**Classifier**	GAZECOUNT	FIXCOUNT	MEANFIXDUR	MAXDWELL	MEANDWELL	MAXFIX	LONGFIX	MAXCONTFIX	ENDCOUNT	**A**	**P**	**R**	**F1**
Music	*logistic regression*	•	•	•						•	72.36	83.33	44.00	53.33
	passive aggressive					•					72.36	67.62	41.67	46.10
	k-NN	•	•								74.55	90.00	43.00	54.19
	decision tree	•	•			•					62.55	20.00	14.00	15.71
	SVM	•	•								74.18	90.00	39.00	52.67
Tutorial	*logistic regression*					•				•	74.55	77.67	57.00	62.55
	passive aggressive	•					•			•	68.73	65.33	45.00	46.33
	k-NN	•				•				•	74.73	73.33	67.00	68.10
	decision tree	•				•	•		•	•	64.55	56.51	90.00	67.32
	SVM	•	•							•	74.73	68.67	73.67	70.81
Documentary	*logistic regression*	•									76.36	76.36	100.0	86.36
	passive aggressive									•	76.36	76.36	100.0	86.36
	k-NN		•		•	•					78.18	81.29	91.43	85.70
	decision tree			•							76.36	76.36	100.0	86.36
	SVM	•									76.36	76.36	100.0	86.36
TOTAL	*logistic regression*	•	•	•		•				•	74.42	79.12	67.00	67.41
	passive aggressive	•				•	•			•	72.48	69.77	62.22	59.60
	k-NN	•	•		•	•				•	75.82	81.54	67.14	69.66
	decision tree	•	•	•		•	•		•	•	67.82	50.96	68.00	56.46
	SVM	•	•							•	75.09	78.34	70.89	69.95

The models predicted preferences with an accuracy between 68.73% and 78.18%. The F1 scores reveal substantial performance differences between the classifiers when applied to the fairly balanced dataset from the *Music* video, with values ranging from 15.71% (decision tree) to 54.19% (k-NN). In the less balanced datasets from the other two videos, performance differences are less pronounced

and the F1 scores are considerably higher. SVM classification produced the highest F1 score of 70.81% for the *Tutorial*, and 86.36% for the *Documentary*.

Although a prediction error of about 25% persists, all classifiers deliver significantly better predictions than both the random baseline and majority voting. We thus proceeded to investigate whether, allowing for the remaining prediction error, *eyeDirect* is capable of increasing engagement. All predictions were performed with the best feature subset for the respective classifier.

Effect on engagement Q. We predicted the preferences of participants from the treatment group by using the models that we had previously trained on the gaze data from the generic control group. The groups (positive adaptation, negative adaptation) were again formed based on whether the video had continued for a participant with their preferred element or not. The preference predictions and subsequent group assignments were repeated with each classifier. We determined the average engagement rating for each group and analyzed whether significant differences exist between the experimental groups. In order to derive generalizable conclusions, we report the aggregated results from all videos (cf. Figure 5.11). Adaptation based on decision tree classification produced the largest effect on engagement. Compared to negative adaptation, engagement ratings increased by 3.55% when positive adaptation was applied. The effect is significant at 90% confidence ($t = 1.402$, p-value = .080). Compared to the generic video display, engagement with positive adaptation was on average 2.9% higher, although the effect is not significant ($t = 1.160$, p-value = .123).

We proceeded to analyze each item that we used to measure engagement separately to verify whether video personalization has a positive effect on any of the engagement constructs *novelty*, *focused attention*, or *involvement*.

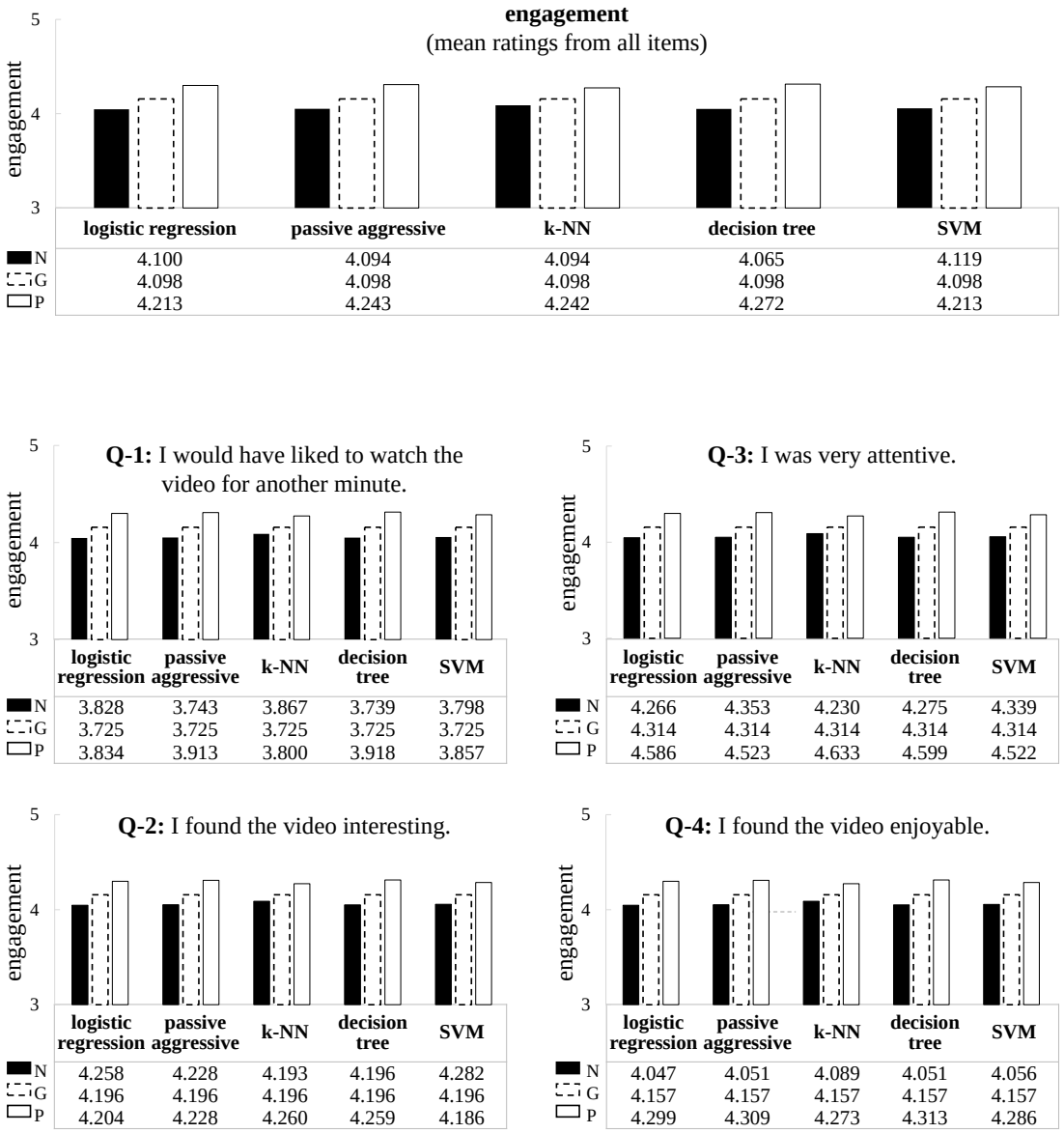


FIGURE 5.11: Effect of machine learning personalization on engagement. Illustrated is the reported engagement for the three gaze-assigned experimental groups. Groups were determined through machine learning classification (**N**: negative adaptation, **G**: generic, **P**: positive adaptation). In addition to the mean scores from all engagement items (**Q-1**, **Q-2**, **Q-3**, **Q-4**), the experimental groups are compared on each of the items separately.

Effect on *novelty*, *focused attention*, and *involvement*. To investigate whether the observed effect on engagement stems from a particular construct related to *novelty*, *focused attention*, or *involvement*, we compared the average ratings from the experimental groups for each of the four engagement measures **Q-1**, **Q-2**, **Q-3**, and **Q-4** individually (cf. Figure 5.11). The comparison was again

performed for the different group assignments resulting from each of the classifiers. Table 5.7 contrasts the average ratings resulting from positive adaptation to the ratings from each of the remaining experimental groups.

TABLE 5.7: Impact of gaze-based video personalization from machine learning on engagement. Experimental groups were determined through machine learning classification (**P**: positive personalization, **N**: negative personalization, **G**: generic). The total engagement rating **Q** is calculated as the mean value from all reported engagement items (Q-1, Q-2, Q-3, Q-4). Statistical significance of the differences between the experimental groups is determined with one-sided t-tests.

		engagement (μ)			**P-N**			**P-G**		
Item	**Classifier**	**P**	**N**	**G**	**$\Delta\ Q_i$**	**t**	**sig.**	**$\Delta\ Q_i$**	**t**	**sig.**
TOTAL Q	*logistic regression*	4.21	4.10	4.10	.131	0.884	.188	.133	0.896	.185
	passive aggressive	4.24	4.09	4.10	.150	1.010	.156	.145	0.975	.165
	k-NN	4.24	4.09	4.10	.147	0.998	.159	.144	0.949	.171
	decision tree	**4.27**	**4.06**	4.10	**.207**	**1.402**	**.080**	.174	1.160	.123
	SVM	4.21	4.12	4.10	.094	0.631	.264	.115	0.775	.219
Q-1	*logistic regression*	3.83	3.83	3.72	.006	0.033	.487	.109	0.589	.278
	passive aggressive	3.91	3.74	3.72	.170	0.901	.184	.187	1.011	.156
	k-NN	3.80	3.87	3.72	-.067	-0.353	.638	.075	0.398	.345
	decision tree	3.92	3.74	3.72	.179	0.951	.171	.193	1.033	.151
	SVM	3.86	3.80	3.72	.059	0.308	.379	.132	0.722	.235
Q-2	*logistic regression*	4.20	4.26	4.20	-.054	-0.310	.622	.008	0.046	.482
	passive aggressive	4.23	4.23	4.20	.000	0.001	.499	.032	0.188	.425
	k-NN	4.26	4.19	4.20	.067	0.388	.349	.064	0.369	.356
	decision tree	4.26	4.20	4.20	.063	0.360	.359	.062	0.364	.358
	SVM	4.19	4.28	4.20	-.096	-0.550	.709	-.010	-0.058	.523
Q-3	***logistic regression***	**4.59**	**4.27**	**4.31**	**.320**	**1.908**	**.028**	**.272**	**1.655**	**.049**
	passive aggressive	4.52	4.35	4.31	.171	1.018	.154	.210	1.255	.105
	k-NN	**4.63**	**4.23**	**4.31**	**.404**	**2.434**	**.007**	**.320**	**1.901**	**.029**
	decision tree	**4.60**	**4.27**	**4.31**	**.323**	**1.938**	**.026**	**.285**	**1.716**	**.043**
	SVM	4.52	4.34	4.31	.183	1.089	.138	.208	1.255	.105
Q-4	***logistic regression***	**4.30**	**4.05**	4.16	**.252**	**1.407**	**.080**	.143	0.847	.198
	passive aggressive	**4.31**	**4.05**	4.16	**.257**	**1.434**	**.076**	.152	0.902	.184
	k-NN	4.27	4.09	4.16	.184	1.031	.151	.116	0.675	.250
	decision tree	**4.31**	**4.05**	4.16	**.262**	**1.464**	**.072**	.156	0.919	.179
	SVM	4.29	4.06	4.16	.229	1.273	.102	.129	0.772	.220

With almost all classifiers, positive adaptation visibly improved *focused attention* (**Q-3**) and *involvement* (**Q-4**). The improvement in *involvement* is significant compared to negative adaptation when applying logistic regression (t = 1.407, p-value = .080), passive aggressive (t = 1.434, p-value = .076), or decision tree classification (t = 1.464, p-value = .072). *Focused attention* increased compared to both negative adaptation and generic videos. The effect is significant with logistic regression (P: t = 1.908, p-value = .028; G: t = 1.655, p-value = .049), k-NN (P: t = 2.434, p-value = .007; G: t = 1.901, p-value = .029), and decision

tree (P: t = 1.938, p-value = .026; G: t = 1.716, p-value = .043).
The effect on *novelty* (**Q-1, Q-2**) is less poignant and is even negative in some of the evaluated scenarios. A one-sided t-test revealed that none of the classifiers adapts the videos in a way that has a significant positive impact on *novelty*.

Consequently, we confirm **H 3B** for the engagement constructs *focused attention* and *involvement*, but reject the hypothesis for *novelty*. In view of the underlying principle of *eyeDirect*, these results are not surprising: Videos are personalized by turning the object that the viewer predominantly looks at into the protagonist. When applying negative adaptation, in contrast, viewers are shown an object to which they previously did not pay much attention. It may therefore hold more *novelty* than an object to which the viewer has already paid close attention. *Focused attention* and *involvement*, on the other hand, are more pronounced for objects to which the viewer has an emotional connection. Displaying a preferred object thus results in higher scores for these constructs.

5.5. Discussion & implications of content filtering

The study demonstrates that *eyeDirect* is able to increase engagement with a video by spontaneously tailoring the plot to the viewer's predicted preferences. In the following, we outline how the insights from the study can help to choose a suitable adaptation strategy for videos and discuss issues that require further research in order for the system to be used in consumer applications.

5.5.1. Defining an adaptation strategy

Depending on the application, *eyeDirect* can have a variety of benefits for the user. Compared to watching a scene of a video that is not aligned with their preferences, viewers with positive adaptation experience higher *involvement*. Our observation of a significant positive effect even for short videos reveals the potential of personalization for fully fledged movies. This finding is particularly relevant for situations in which a generic scene is not appropriate, and thus branching cannot be avoided. Instead of centering the plot around an object that seeks to address the broadest possible target group, the director may offer multiple variants of a scene. One example are commercial breaks on streaming platforms such

as YouTube. In order to increase the relevance of advertisements, the platform currently uses – provided that permission is granted by the user – information from their Google Account including age, gender, interactions with an advertiser, and Google search queries (YouTube, 2023). Consequently, the commercials are more likely to target the users' general profiles, but may be irrelevant to them in their current situation. With *eyeDirect*, it is possible to identify momentary preferences and needs. Instead of showing context irrelevant commercials that the user will skip if possible, the advertisement can be chosen based on the gaze distribution to the content of previously watched videos or scenes of a movie.

The positive effect on *focused attention* when videos are aligned with the users' preferences has two important implications. First, in addition to providing a higher entertainment factor, presenting an object that the viewer is drawn towards instead of something they may not find interesting is relevant when sustaining the audience's attention is crucial. Second, attention to personalized videos is also higher compared to a generic display. Among other things, this can be used to make educational videos more effective. One example are safety instruction videos on airplanes. Airlines tend to spend large amount of money on making flight safety videos more appealing. Air New Zealand (2014), for instance, recreated the universe of 'The Hobbit': In 'The most epic safety video ever made', safety instructions are delivered by the inhabitants of Middle Earth. And yet, many passengers are not attentive to these videos. Given that most large aircrafts are equipped with an individual screen for each passenger, flight safety videos can easily be personalized. With *eyeDirect*, a brief overture into the video is sufficient to determine a frame story that is aligned with the passenger's interests, making them more likely to pay attention to the potentially life preserving instructions.

We found no evidence for higher *novelty* perception if the content of a video reflects a viewer's preferences. We do, however, expect long-term effects for users who watch a video multiple times. In particular, when a movie with several branches is watched repeatedly, the plot may change, leading to a completely new user experience. The plot thereby deliberately turns towards elements that were previously overlooked (Vesterby et al., 2005). Analyzing the long term effects of such movies is a promising direction for future research.

5.5.2. Considerations for consumer app design

While the adaptive videos created with *eyeDirect* were effective in the controlled experimental setting, special care must be taken with regard to both technical and ethical issues when using the system in consumer applications. Even though eye tracking is a mature technique that has been used in academic and non-academic contexts, some work is still needed to make it accessible to a greater audience. The high-quality Tobii Pro X3-120 eye tracker with which the experiment was performed requires special hardware which is impractical and too expensive for the common user. Less sophisticated devices produce gaze recordings of lower quality, thus making the preference predictions less accurate. In real-world settings, additional challenges such as poor ambient lighting or users frequently diverting their gaze away from the screen make the gaze data even more unreliable. However, open-source solutions like `OpenFace`[8], `GazeML`[9], and `MPIIFaceGaze`[10] that use built-in device cameras are catching up in their performance and are becoming more resilient to varying environmental conditions. While their accuracy is lower compared to specialized hardware, it is sufficient for assigning the gaze coordinates to an AOI (Zhang et al., 2019).

A related issue is the need for user-specific calibration – a primary prerequisite of state-of-the-art eye trackers to deliver accurate gaze estimates (Majaranta & Bulling, 2014). Especially when used on short videos, the traditional five-point calibration ruins a seamless viewing experience. Deep learning might help to create a calibration-free system, but the quality of the gaze estimates is still considerably higher after calibration (Krafka et al., 2016).

The ineffectiveness of the naïve preference predictions suggests that salient properties of the visual material bias the viewers' gaze. Consequently, the raw gaze distributions do not adequately represent preferences. Machine learning filters out saliency effects, but requires prior collection of scene-specific training data. In order to avoid repeating the time-consuming data collection for each new scene, Qvarfordt and Zhai (2005) define a set of visual properties that seek to capture the saliency bias and quantify their effect. Unbiased attention to an object is calculated by removing the effect of the visual properties as a function of their

[8] `OpenFace`: https://cmusatyalab.github.io/openface/ (Amos et al., 2016).
[9] `GazeML`: https://github.com/swook/GazeML (Park et al., 2018).
[10] `MPIIFaceGaze`: https://perceptualui.org/research/datasets/MPIIFaceGaze/ (Zhang et al., 2017).

saliency factor in the visual display. Cheng et al. (2010) compare new scenes to a video database on which machine learning models have been trained. The fitted model from a comparable video is used to predict preferences for a new scene.

From an ethical viewpoint, the confidentiality and responsible handling of the collected data has to be safeguarded. The effectiveness of the gaze to detect human preferences – while beneficial in many situations – implies that gaze data carries a wealth of information and might even reveal political, religious, or sexual tendencies. Users may therefore be reluctant to have their eye movements monitored. Among Internet of Things (IoT) applications, eye tracking provokes the strongest resistance, especially for preference detection tasks (Lee & Kobsa, 2017; Steil et al., 2018). However, users are less concerned when their data is not forwarded to others (Steil et al., 2018). They are more likely to accept an adaptive system such as *eyeDirect* where gaze data can be processed locally and discarded after preferences for an object have been determined. It is, however, the duty of the system designer to ensure responsible handling of the collected data.

5.6. Conclusion: Summary of Essay 3

This essay presents *eyeDirect*, a system for creating personalized videos based on gaze data. While simple majority voting applied to individual gaze features is not capable of adapting videos in a way that is meaningful to the user, personalization based on machine learning does have the desired impact on engagement.
The results of a comprehensive user study (N = 175) reveal that videos that are aligned with the viewer's preferences induce more *focused attention* compared to both adversely adapted content and videos with a generic storyline. They also lead to higher *involvement* compared to videos with adversely adapted content. *Novelty* perception, in contrast, does not increase. Videos created with *eyeDirect* can thus increase involvement and attention if the viewer's preferred object takes the center stage. In contrast, alternating the plot of a movie when played multiple times or prioritizing unattended elements may affect the perception of novelty.

Studying the long-term effects of full-fledged adaptive movies and plot variations when a movie is watched repeatedly are promising directions for future research. Possible extensions to the system include using additional physiological sensors to infer the viewer's preferences even if only a single object is displayed at a time.

6. General discussion and conclusion

The three research essays of this thesis formulate and evaluate adaptation rules that define how multimodal interfaces should change the presented content or modality in response to the user's dynamic state. Their potential for improving the effectiveness and/or satisfaction with multimodal interfaces is grounded in cognitive theories with a solid empirical foundation (Giles et al., 1987; Sweller et al., 1998; Wertheimer, 1938). Essay 1 and Essay 2 focus on modality adaptation, whereas Essay 3 pursues an adaptation strategy that modifies the content itself. Unlike traditional recommender systems, such as those used in online shops or social media, the content adaptation is specifically tailored towards multimodal interfaces that have the capability to detect preferences from non-invasive, dynamically collected gaze data.

In addition to shedding light on the effectiveness of different forms of adaptation, this thesis contributes to HCI research – and adaptive multimodal interfaces in particular – by identifying sensor input that reveals insightful information about the user. This information, in turn, can be used to identify whether the adaptation condition of a rule is met. The rule tested in Essay 1 relies on speech and manual cursor input from natural interactions with the application. Essay 2 and Essay 3 use vision-based data. Essay 2 extracts facial gestures from images captured with the integrated camera of a laptop. In Essay 3, gaze data is recorded with a commercial eye tracker, but could also be calculated from video frames recorded with a consumer-grade camera (Papoutsaki et al., 2016; Zhang et al., 2019). The key findings and contributions of each essay are summarized in Table 6.1.

Essay 1 shows that matching system output to the modality that the user chooses for the first inputs does not improve usability. The initial input is followed by a high number of modality switches, which is a clear indicator that input preferences at the beginning are unstable. Therefore, the typically short interaction for authentication is not a valid indicator for input preferences in later routine tasks. Whether modality preferences for authentication and routine tasks differ in general

TABLE 6.1: Summary of key findings and contributions of the three research essays. The insights from the essays have two major contributions: (1) They have practical implications for the design of multimodal systems; (2) they extend existing knowledge on cognitive theories.

	Key findings	Key contributions
ESSAY 1	• Output modality alignment to initial input does not improve usability • A major cause of the ineffectiveness of the adaptation is the gradual formation of modality preferences: 1. The first input does not indicate a general modality preference 2. Frequent modality switches • Preferences should be inferred from longer interactions ▷ **Violates robustness** ($\boldsymbol{R_{NF}3}$)	**Systems design**: • Formulation of an adaptation rule for *modality alignment* • Recommendation that user models should not update modality preferences based on a one-time snapshot of the first user input • Demonstration that modality alignment is no valid adaptation when user input is sparse **Theory**: Limited applicability of Communication Accommodation and Gestalt Theory in HCI when a person's own use of communication channels is volatile
ESSAY 2	• Multimodal presentation improves comprehension only for some users, and only if confusion is caused by poor audio or language deficiencies (versus contentual incongruity) ▷ **Violates universality** ($\boldsymbol{R_{NF}2}$) • The effect of multimodal presentation on the user cannot be predicted from observable factors, but instead depends on subjective perceptions of its usefulness • Confusion can be detected from automatically extracted facial gestures: 1. Unimodal (audio only) presentation: action units 2. Multimodal (audio & text) presentation: action units, emotions, and blink frequency	**Systems design**: • Formulation of an adaptation rule for *multimodal redundancy* • Demonstration of conditional effectiveness of multimodal redundancy 1. Specification of an additional requirement: Verification of adaptation effectiveness through continuous user state monitoring 2. Recommendation to reinstate the unimodal state if multimodal presentation does not improve comprehension • Identification of facial cues that reveal confusion during each unimodal (audio only) and multimodal (audio & text) presentation **Theory**: Interpersonal differences determine how strongly a user is affected by each of the two principles of Cognitive Load Theory (i.e., modality and redundancy)
ESSAY 3	• Content filtering in videos based on gaze-informed preferences improves attentiveness and involvement • Preference detection from gaze data is reliable and robust even in situations in which little or no active user input is available • The effect is universally observed across diverse gender and age groups	**Systems design**: • Formulation of an adaptation rule for *content filtering* • Demonstration that content filtering can be implemented in a way that improves usability, while also meeting the requirements of robustness and universality • Identification of gaze features that reliably predict preferences for elements in videos **Theory**: Engagement (mediated by intrinsic load) decreases with the number of visual elements when performing a purely observational task

is a topic that merits further investigation in future research. Moreover, Essay 1 does not make any claims about the effectiveness of modality alignment to the user's input once a solid preference has been established. Rather, the evidence from the experiment implies that, in order to accurately detect modality preferences, the user's behavior needs to be monitored throughout an extended interaction. Since this violates the robustness requirement $\boldsymbol{R_{NF}3}$, Essay 2 evaluates the effectiveness of non-adaptive multimodal presentation, which can be applied without having any data about the user. The results of the experimental investigation imply that, compared to a purely auditory presentation, bimodal audio-visual output reduces confusion caused by poor audio quality or language deficiencies – but not confusion from contextual incongruity – for some users. Assuming that the detected confusion is not just a transient state, but instead is likely to persist, Essay 2 proposes to dynamically activate bimodal output when a user is confused in order to proactively improve upcoming interactions. Since, however, bimodal presentation is only beneficial to those who effectively perceive it as useful, the effectiveness of bimodal presentation should be continuously monitored. If confusion levels do not decrease, the interface should revert to unimodal output to avoid negative side effects of redundant content presentation. Yet again, a selective application of the rule violates the requirement for universality $\boldsymbol{R_{NF}2}$. Essay 3 therefore tests the effectiveness of content filtering based on gaze-informed preferences. The gaze data is collected during a passive task (i.e., the users watch a video without providing any active input), thus making the adaptation robust to situations with no or little user interaction. A demographic subgroup analysis attests universal applicability of the rule across users of diverse age and gender.

6.1. Major contributions of this thesis

In the following, two major contributions of this thesis are outlined: (1) the practical implications for future designs of multimodal interfaces; (2) the theoretical contributions to the understanding of the cognitive effects that different adaptations have on the user.

6.1.1. Practical implications

This section seeks to outline to what extent each of the three adaptation rules meets the requirements formulated in Section 1.2 and discusses relevant design considerations that should be addressed when planning to implement an adaptation rule that violates a requirement.

The rules for modality alignment (Rule 1) and content filtering (Rule 3) were defined with a particular emphasis on meeting all five functional requirements. Yet, as it turned out, the one-time snapshot of the user's input – whether from authentication or from the first few interactions in subsequent routine tasks – in Rule 1 was not sufficiently dynamic ($\boldsymbol{R_F3}$) for an accurate and comprehensive characterization of the user. The snapshot may capture the user's preferred modality for the specific input at the precise moment of the interaction. However, it is not a valid indicator of the user's preferences for subsequent inputs, even if the external conditions remain unchanged. Even though empirical evidence suggests that modality preferences can be detected almost immediately (Oviatt et al., 2005), a longer monitoring phase might allow for a more accurate characterization of the user. In order to additionally capture state changes, the interface should continuously monitor the user's interaction behavior. Such a change can, for instance, happen when other people entering the room render auditory communication inconvenient.
In the multimodal redundancy strategy (Rule 2), the functional requirements for adaptation based on user-centric dynamic states ($\boldsymbol{R_F1}$ - $\boldsymbol{R_F3}$) were relaxed. This adjustment to the objectives was made to test an approach that, by not depending on the collection of user data, guarantees robustness.

From the summary of the requirements verification in Table 6.2 it becomes obvious that any relaxation of functional requirements has detrimental spill-over effects on the non-functional requirements. Whenever one of the functional requirements is not or only partly fulfilled, the adaptation rule does not meet the complete set of non-functional requirements either. Specifically, the one-time snapshots of the users' modality choices in Rule 1 turned out not to be an accurate reflection of their preferences. Consequently, the adaptation is not robust to situations where user input is sparse ($\boldsymbol{R_{NF}3}$) and leads to no improvement in usability ($\boldsymbol{R_{NF}1}$). Abandoning $\boldsymbol{R_F1}$ - $\boldsymbol{R_F3}$ in Rule 2 resulted in a non-adaptive approach

that does not take into account individual differences between users. While the rule does not improve overall usability ($\boldsymbol{R_{NF}1}$), we observed a beneficial effect for a subgroup of the users. Yet again, this violates the universality requirement ($\boldsymbol{R_{NF}2}$). Ultimately, only content filtering based on gaze-informed preference predictions meets all functional and non-functional requirements. Usability across all users of the adaptive interface was higher, irrespective of their demographic characteristics. Given that the movie streaming application receives no active input from the user, we expect a robust beneficial effect on usability for any application and context of use – provided that the ambient conditions deliver accurate gaze data (O'Brien, 2009; Zhang et al., 2019).

TABLE 6.2: Verification of requirements. Circles indicate that a rule fulfills (●) or partly fulfills (○) a requirement. Violated requirements are marked with a blue cross (✘). Compliance of suggested rule modifications in *italic* with non-functional requirements is based on statistical subgroup analyses of the original adaptation rule.

		REQUIREMENTS							
		functional					non-functional		
Adaptation rule		R_F1	R_F2	R_F3	R_F4	R_F5	$R_{NF}1$	$R_{NF}2$	$R_{NF}3$
Rule 1:	**Modality alignment**	●	●	○	●	●	✘	●	✘
Rule 2-1:	**Multimodal redundancy**	✘	✘	✘	●	●	✘	✘	●
Rule 2-2:	*Selective multimodal redundancy*	●	●	●	●	●	●	✘	●
Rule 3:	**Content filtering**	●	●	●	●	●	●	●	●

$\boldsymbol{R_F}$: R_F1 – adaptive presentation; R_F2 – user-centricity; R_F3 – dynamic states; R_F4 – non-invasiveness; R_F5 – ubiquitous deployment;
$\boldsymbol{R_{NF}}$: $R_{NF}1$ – usability; $R_{NF}2$ – universality; $R_{NF}3$ – robustness

While the non-compliance of Rule 1 and Rule 2 with some of the requirements compromises their unconditional applicability, they may still be beneficial to a subset of the target group, or in certain contexts of use. In particular, based on the insights from Essay 2, multimodal redundancy (Rule 2) promises to increase comprehension for some users if their confusion stems from poor audio quality or language deficiencies. Therefore, interfaces that are able to detect confusion by, for instance, analyzing facial expressions (Borges et al., 2019; D'Mello et al., 2009; Yasser et al., 2021) or gaze data (Kunze et al., 2013; Lallé et al., 2016), can apply the modified Rule 2-2. By activating bimodal presentation only when the continuously collected data suggests that it reduces confusion, users that are receptive to the adaptation are supported. At the same time, the selective adaptation rule avoids negative side effects that can occur from redundant presentation (Kalyuga, 2012). In addition to the virtual meeting

application evaluated in Essay 2, examples of applications where it is particularly desirable to minimize confusion are educational systems (Calvi et al., 2008) or videos (Fujii & Rekimoto, 2019).
Similarly, the findings from Essay 1 indicate that modality alignment is not a viable option when active input from the user is sparse. This is typically the case for applications that are only used for short interactions such as public self-service terminals (Gupta & Sharma, 2021) or provide little active input from the user. Examples of the latter are narrative applications like movies or digital books (Gilroy et al., 2012; Vesterby et al., 2005). Yet, a positive effect of modality alignment has been reported repeatedly (Schaeffner et al., 2016; Stelzel & Schubert, 2011; Stephan & Koch, 2011). We thus presume that insufficient data from the one-time snapshot of the user's interaction behavior accounts for the ineffectiveness of Rule 1. We further expect that the users' preferences for a modality will become clear after interacting with the interface for an extended period of time. Prior research has shown that users tend to switch modalities mostly to overcome issues related to performance and usability, or simply out of curiosity to try out other available communication channels (Gürkök et al., 2011). Matching the presentation during this initial phase to every modality change would undoubtedly induce high switching costs (Schaeffner et al., 2016; Stephan & Koch, 2010). However, since issues with a modality typically surge at the very beginning when using a new system, we expect modality preferences to stabilize after the first few interactions. Provided that the empirical evidence on the positive effect of modality alignment holds and that preferences can be accurately detected if sufficient interaction data is collected, applications that engage the user over an extended time may indeed benefit from the adaptation rule. Typical examples are search engines, communication tools, and entertainment applications such as games, comics, or social media (Kim et al., 2019; Tian et al., 2021).

6.1.2. Theoretical contributions

Each essay of this thesis discusses the implications of its theoretical contributions for future research. This section extends the discussions with a broader view on the theoretical insights by relating them to the cognitive theories that underlie the adaptation rules. The insights from the essays are once more evaluated through the lens of cognitive processes that are involved in the perception and processing

of information. Existing knowledge about the cognitive theories is extended by considering their implications for human-computer interaction and, in particular, assessing their applicability in multi-modal systems.

Essay 1 suggests that *Communication Accommodation* (Giles et al., 1987) does not immediately extrapolate to information systems. While humans perceive a communication partner who mimics their behavior as empathetic (Iacoboni, 2009), it appears that they do not expect the same behavior from a smart interface – at least not at the beginning of the interactions, when they have not yet established a dominant communication pattern. Rather, the frequent modality switches during the initial phase imply that the perceptual system is primed equally for all available modalities. Consequently, modality alignment does not make the communication more efficient.

The insights from Essay 1 furthermore have implications for the application of *Gestalt Theory* across different modalities (Giles et al., 1987). The usability gains from combining compatible input and output channels (Oviatt et al., 2003) do not emerge when the use of modalities within the same communication direction (i.e., input) is in itself inconsistent. This corroborates empirical evidence of the transition costs that arise when commands are issued by alternating manual and auditory input – even if the output is presented in a modality that is compatible with the most recent input (Schaeffner et al., 2016). The observations from Essay 1 imply that an alignment of the presentation to the user's initial input is not perceived as symmetrical if they afterwards use a different modality. Instead, the subsequent modality change offsets any previous perception of symmetry.

Essay 2 reveals interpersonal differences regarding the effectiveness of bimodal presentation. Given that the auditory material of the user study was presented by the same person using the same equipment throughout all experimental sessions, external factors such as the quality of the audio transcription provide no satisfactory explanation for the observed differences. Instead, Essay 2 implies that users differ in how strongly they are affected by each of the two conflicting principles (i.e., modality and redundancy) of *Cognitive Load Theory*. Some users – those for whom the expansion of the working memory (modality principle) outweighs the negative effects of redundancy – benefit from multimodal output. For others, the activation of multiple subsystems of the brain through the different modalities cannot compensate for the additional load from redundant presentation. Previous

work has attributed such individual differences in the susceptibility to redundancy and modality effects to working memory capacity (Batka & Peterson, 2005) and prior knowledge (Adesope & Nesbit, 2012).
Essay 3 adds to the understanding of how the amount of individual elements in a video that, while logically connected, are clearly separable affects engagement. Reducing the number of elements that users must simultaneously attend to increases their focus – even if no cognitive task is given. While the subjective evaluations of engagement are not synonymous to cognitive load, a causal relationship between the two cognitive constructs can be assumed (Bueno-Vesga et al., 2021) and has been validated on educational videos (Altinpulluk et al., 2020). Consequently, Essay 3 implies that intrinsic load during a purely observational task increases with the number of visual elements.

6.2. Limitations & future research

The specific limitations of each essay were discussed in detail in Chapters 3 - 5. In the following, we extend the individual discussions by outlining the overall limitations of this thesis and the implied potential for future research.

6.2.1. Experimental design

For the evaluations of Rule 1 and Rule 3, user models were created with sensor data from the user studies. We used this experimental approach to learn how effective the adaptation can realistically be, taking into account that it relies on imperfect information about the user. Essay 3 applies a simulation-based evaluation to assess different user modeling algorithms and determine how effective the adaptation can be at most, when the best performing algorithm is implemented. Alternatively, we could have used questionnaires to determine the users' true preferences and create perfect user models. This would have allowed us to isolate the effect of the interface change itself. However, it is unreasonable to assume the existence of a perfect sensor-informed user model (Chen et al., 2021). In contrast, our observations draw a more accurate picture of the effectiveness of the adaptation when the multimodal interface is used for real-world applications.

Nevertheless, it should be noted that all three adaptation rules are evaluated on experimental prototypes with simplified functions and limited adaptive behavior. The effectiveness of the adaptation when integrated into a full-fledged multimodal interface and deployed in the field may differ. Our main motivation for following a rapid prototyping approach (Tripp & Bichelmeyer, 1990) rests on the design process itself. Whenever an evaluation does not meet the specified objectives, the artifact needs to be refined (Peffers et al., 2007). Since the development of a full-fledged system is very time-consuming, prototypes have become an indispensable evaluation method for many software design projects (Tripp & Bichelmeyer, 1990). Having established the effectiveness of content filtering based on gaze-informed preferences, one of the major missions for future research is now to validate the adaptation rule with different applications and when deployed in the field.

6.2.2. Application specificity

This thesis follows a problem-centered design process (Peffers et al., 2007). Thus, we formulated generic rules with the aim to serve a broad range of applications and contexts. To evaluate their effectiveness, each rule was integrated into an exemplary application from a domain where the proposed adaptation promises to be particularly beneficial. While Essay 3 showcases the advantages of content filtering in the context of movie streaming, the potential of the adaptation to decrease cognitive load (van Merriënboer & Sweller, 2010) identifies it as a promising strategy for other applications as well. Yet, the specifics of the filtering approach need to be defined for each application individually. For instance, smart home displays should always provide easy access to all rooms in the house, independent of whether the user has previously adjusted its ambient conditions. Nevertheless, a leaner interface might, for example, group infrequently visited rooms in a submenu which the user can then expand if needed (Brusilovsky, 2012; Shneiderman, 2020).

In this sense, it should be a key concern of developers who wish to integrate the adaptation rules into their multimodal interface to identify an appropriate filtering mechanism that meets the particular requirements of their application.

6.2.3. Design process

The rules were evaluated using throw-away prototypes. After each iteration of the design process, a fundamentally different adaptation rule was defined. We applied this procedure after detecting violations of the non-functional requirements that cannot be removed with simple modifications. In particular, in order for modality alignment (Rule 1) to be robust ($\boldsymbol{R_{NF}3}$), interaction data must be collected throughout the entire interaction (instead of using only initial input) and the adaptation should be executed only after the user has formed a clear preference for one modality (Gürkök et al., 2011). This, in turn, requires defining a suitable adaptation trigger (Feigh et al., 2012). Moreover, interactions with a smart home display tend to be short, often even below 10 seconds (Castelli et al., 2017). Manipulating the experimental instructions in a way that produces the required amount of data would result in a highly unrealistic scenario. Therefore, while potentially meeting the usability requirement ($\boldsymbol{R_{NF}1}$), a modification of the adaptation rule to collect more interaction data violates the objective of defining a robust rule that can be applied to any use case ($\boldsymbol{R_{NF}3}$).
Multimodal redundancy (Rule 2), in turn, is only beneficial to a subgroup of users. We propose an extension to the rule so that multimodal presentation improves the usability for those who are susceptible to the modality principle (Gellevij et al., 2002; Mousavi et al., 1995), while not deteriorating the experience for users who are immune to its benefits (Adesope & Nesbit, 2012; Batka & Peterson, 2005). While this again complies with the usability requirement ($\boldsymbol{R_{NF}1}$), no modification can resolve the issues of universal applicability ($\boldsymbol{R_{NF}2}$).

Against this backdrop, the throw-away prototyping strategy allowed us to keep close track of our objectives. Then again, discarding an adaptation rule entirely entails that the effectiveness of the proposed modifications needs to be verified in future research.

6.2.4. Theoretical perspectives

The adaptation rules were formulated with the objective to improve information processing in the working memory by presenting content in a way that cognitive theories suggest will be beneficial (Giles et al., 1987; Sweller, 1988; Wertheimer,

1938). Multimodal redundancy and content filtering have the potential to reduce cognitive workload (*Cognitive Load Theory*). Modality alignment, in turn, promises synergy effects through modality priming (*Communication Accommodation Theory*) or emergent properties of symmetry (*Gestalt Theory*). The selected theories were chosen because of their conceptual fit with the configuration space of possible output adjustments as it applies to most multimodal interfaces – i.e., the manipulation of its content or modality (Jameson, 2007; Kong et al., 2011; Maybury & Wahlster, 1999). Alternative theoretical perspectives may lead to different adaptation rules that pursue other optimization goals and, thus, provide opportunities for future research.

One alternative perspective is provided by Affect Theory (Tomkins, 1984). Affective computing has a longstanding history of investigating the role of emotions and affect in HCI (Picard, 2000). For example, affective agents can improve trust through human-like qualities such as displaying emotions (Pelau et al., 2021). Cinematographic applications can use affect to shape the development of a narrative (Gilroy et al., 2012; Peng et al., 2018).

Accessible interfaces offer another perspective by striving to serve all users, independent of their physical or cognitive abilities (Peissner et al., 2012). For instance, adapting the presentation speed or complexity of visual elements can support users with motor or vision impairments (Stephanidis et al., 1998).

Future research can refer to these additional theoretical perspectives to develop adaptation rules that address the quality of interactions with multimodal interfaces on yet another dimension. Through a holistic consideration of the interaction quality on multiple dimensions, adaptive multimodal interfaces have the potential to improve usability on all three levels of effectiveness, efficiency, and satisfaction.

Bibliography

Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, *34*(11), 1617–1630.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*(2), 219. https://doi.org/10.1037/0033-2909.121.2.219

Adam, M. T., Gimpel, H., Maedche, A., & Riedl, R. (2017). Design blueprint for stress-sensitive adaptive enterprise systems. *Business & Information Systems Engineering*, *59*(4), 277–291. https://doi.org/10.1007/s12599-016-0451-3

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, *104*(1), 250. https://doi.org/10.1037/a0026147

Afzal, S., & Robinson, P. (2010). Modelling affect in learning environments - motivation and methods. *2010 10th IEEE International Conference on Advanced Learning Technologies*, 438–442. https://doi.org/10.1109/ICALT.2010.127

Ahmad, M. I., Keller, I., Robb, D. A., & Lohan, K. S. (2020). A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-020-01455-7

Air New Zealand. (2014). The most epic safety video ever made. Retrieved June 11, 2021, from https://www.youtube.com/watch?v=qOw44VFNk8Y&t=82s

Aldridge, L. C., & Lansdown, T. C. (1999). Driver preferences for speech based interaction with in-vehicle systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(18), 977–981. https://doi.org/10.1177/154193129904301807

Alibay, F., Kavakli, M., Chardonnet, J.-R., & Baig, M. Z. (2017). The usability of speech and/or gestures in multi-modal interface systems. *Proceedings of the 9th International Conference on Computer and Automation Engineering*, 73–77. https://doi.org/10.1145/3057039.3057089

Altinpulluk, H., Kilinc, H., Firat, M., & Yumurtaci, O. (2020). The influence of segmented and complete educational videos on the cognitive load, satisfaction, engagement, and academic achievement levels of learners. *Journal of Computers in Education*, *7*(2), 155–182. https://doi.org/10.1007/s40692-019-00151-7

Amazon. (2021). Amazon Alexa Voice AI. Retrieved June 10, 2021, from https://developer.amazon.com/en-US/alexa

Amazon. (2022). Amazon echo. Retrieved December 15, 2022, from https://www.amazon.de/echo-show/s?k=echo+show

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). *Openface: A general-purpose face recognition library with mobile applications* (tech. rep.). CMU-CS-16-118, CMU School of Computer Science.

Anderson, M., & Perrin, A. (2017). *Tech adoption climbs among older adults* (tech. rep.). Pew Research Center.

André, E. (2000). The generation of multimedia presentations. *Handbook of natural language processing*, *12*, 305.

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.004

Apoki, U. C., Al-Chalabi, H. K. M., & Crisan, G. C. (2020). From digital learning resources to adaptive learning objects: An overview. *Modelling and Development of Intelligent Systems*, 18–32. https://doi.org/10.1007/978-3-030-39237-6_2

Apple. (2021). Siri does more than ever. even before you ask. Retrieved June 10, 2021, from https://www.apple.com/siri/

Apple. (2022a). Apple watch. Retrieved December 15, 2022, from https://www.apple.com/watch/

Apple. (2022b). Iphone. Retrieved December 15, 2022, from https://www.apple.com/iphone/

Arens, Y., Hovy, E. H., & Vossers, M. (1991). On the knowledge underlying multimedia presentations. *Proceedings of the 1991 International Conference on Intelligent Multimedia Interfaces*, 280–306.

Arguel, A., Lockyer, L., Lipp, O. V., Lodge, J. M., & Kennedy, G. (2017). Inside out: Detecting learners' confusion to improve interactive digital learning environments. *Journal of Educational Computing Research*, *55*(4), 526–551. https://doi.org/10.1177/0735633116674732

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bafna, T., Hansen, J. P. P., & Baekgaard, P. (2020). Cognitive load during eye-typing. *ACM Symposium on Eye Tracking Research and Applications*. https://doi.org/10.1145/3379155.3391333

Baig, M. Z., & Kavakli, M. (2018). Qualitative analysis of a multimodal interface system using speech/gesture. *2018 13th IEEE Conference on Industrial Electronics and Applications*, 2811–2816. https://doi.org/10.1109/ICIEA.2018.8398188

Baig, M. Z., & Kavakli, M. (2019). A survey on psycho-physiological analysis & measurement methods in multimodal systems. *Multimodal Technologies and Interaction*, *3*(2). https://doi.org/10.3390/mti3020037

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, *16*(10), 814–819. https://doi.org/10.1111/j.1467-9280.2005.01619.x

Bakkes, S., Tan, C. T., & Pisan, Y. (2012). Personalised gaming: A motivation and overview of literature. *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System*. https://doi.org/10.1145/2336727.2336731

Batka, J. A., & Peterson, S. A. (2005). The effects of individual differences in working memory on multimedia learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(13), 1256–1260. https://doi.org/10.1177/154193120504901309

Bednarik, R. (2005). Potentials of eye-movement tracking in adaptive systems. *Fourth Workshop on the Evaluation of Adaptive Systems*.

Bennett, A. A., Campion, E. D., Keeler, K. R., & Keener, S. K. (2021). Videoconference fatigue? exploring changes in fatigue after videoconference meetings during covid-19. *Journal of Applied Psychology*, *106*(3), 330. https://doi.org/10.1037/apl0000906

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., & Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement Disorders*, *12*(6), 1028–1034. https://doi.org/10.1002/mds.870120629

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Best Of Wildlife. (2019). Brave moment cat fights a large python to protect its kittens. Retrieved September 19, 2019, from https://www.youtube.com/watch?v=2yY5gy0u4Ck

Bierschwale, J. M., Sampaio, C. E., Stuart, M. A., & Smith, R. L. (1989). Speech versus manual control of camera functions during a telerobotic task. *Proceedings of the Human Factors Society Annual Meeting*, *33*(2), 134–138. https://doi.org/10.1177/154193128903300229

Bigornia, A. (2015). IBM/Facebook partnership. making irrelevant ads a thing of the past. Retrieved November 12, 2018, from https://www.ibm.com/blogs/insights-on-business/consumer-products/ibmfacebook-partnership-making-irrelevant-ads-a-thing-of-the-past/

Boll, S., Klas, W., & Wandel, J. (1999). A cross-media adaptation strategy for multimedia presentations. *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, 37–46. https://doi.org/10.1145/319463.319468

Bolt, R. A. (1980). "Put-That-There": Voice and gesture at the graphics interface. *SIGGRAPH Computer Graphics*, *14*(3), 262–270. https://doi.org/10.1145/965105.807503

Bolt, R. A. (1981). Gaze-orchestrated dynamic windows. *SIGGRAPH Computer Graphics*, *15*(3), 109–119. https://doi.org/10.1145/965161.806796

Borges, N., Lindblom, L., Clarke, B., Gander, A., & Lowe, R. (2019). Classifying confusion: Autodetection of communicative misunderstandings using facial action units. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, 401–406. https://doi.org/10.1109/ACIIW.2019.8925037

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., Caltagirone, C., Kong, W., Wei, D., Zhou, Z., Polidori, L., Vitiello, S., & Babiloni, F. (2012). Assessment of mental fatigue during car driving by using high resolution eeg activity and neurophysiologic indices. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6442–6445. https://doi.org/10.1109/EMBC.2012.6347469

Breetzke, T., & Flowerday, S. V. (2016). The usability of ivrs for smart city crowdsourcing in developing cities. *The Electronic Journal of Information Systems in Developing Countries*, *73*(1), 1–14. https://doi.org/10.1002/j.1681-4835.2016.tb00527.x

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System*, *25*(1), 39–53. https://doi.org/10.1016/S0346-251X(96)00059-0

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control [Psychophysiology of Workload]. *Biological Psychology*, *42*, 361–377. https://doi.org/10.1016/0301-0511(95)05167-8

Brown, A., Jones, R., Crabb, M., Sandford, J., Brooks, M., Armstrong, M., & Jay, C. (2015). Dynamic subtitles: The user experience. *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, 103–112. https://doi.org/10.1145/2745197.2745204

Brun, Y., Marzo Serugendo, G., Gacek, C., Giese, H., Kienle, H., Litoiu, M., Müller, H., Pezzè, M., & Shaw, M. (2009). Engineering self-adaptive systems through feedback loops. In *Software*

engineering for self-adaptive systems (pp. 48–70). Springer-Verlag. https://doi.org/10.1007/978-3-642-02161-9_3

Brusilovskiy, P. L. (1994). The construction and application of student models in intelligent tutoring systems. *Journal of computer and systems sciences international*, *32*(1), 70–89.

Brusilovsky, P. (2003). Developing adaptive educational hypermedia systems: From design models to authoring tools. In *Authoring tools for advanced technology learning environments: Toward cost-effective adaptive, interactive and intelligent educational software* (pp. 377–409). Springer Netherlands. https://doi.org/10.1007/978-94-017-0819-7_13

Brusilovsky, P. (2012). Adaptive hypermedia for education and training. *Adaptive technologies for training and education*, *46*, 46–68.

Bubalo, N., Honold, F., Schüssel, F., Weber, M., & Huckauf, A. (2016). User expertise in multimodal HCI. *Proceedings of the European Conference on Cognitive Ergonomics*. https://doi.org/10.1145/2970930.2970941

Bueno-Vesga, J. A., Xu, X., & He, H. (2021). The effects of cognitive load on engagement in a virtual reality learning environment. *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 645–652. https://doi.org/10.1109/VR50410.2021.00090

Buisine, S., & Martin, J.-C. (2003a). Design principles for cooperation between modalities in bi-directional multimodal interfaces. *Proceedings of the CHI 2003 workshop on Principles for Multimodal User Interface Design*, 72–75.

Buisine, S., & Martin, J.-C. (2003b). Experimental evaluation of bi-directional multimodal interaction with conversational agents. *IFIP TC13 International Conference on Human-Computer Interaction*, 168–175.

Burke, R. R., & Leykin, A. (2014). Identifying the drivers of shopper attention, engagement, and purchase. *Review of Marketing Research*, *11*, 147–187. https://doi.org/10.1108/S1548-643520140000011006

Burke, R. (2007). Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 377–408). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_12

Buscher, G., Cutrell, E., & Morris, M. R. (2009). What do you see when you're surfing? using eye tracking to predict salient regions of web pages. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 21–30. https://doi.org/10.1145/1518701.1518705

Calvi, C., Porta, M., & Sacchi, D. (2008). E5learning, an e-learning environment based on eye tracking. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 376–380. https://doi.org/10.1109/ICALT.2008.35

Castelli, N., Ogonowski, C., Jakobi, T., Stein, M., Stevens, G., & Wulf, V. (2017). What happened in my home? an end-user development approach for smart home data visualization. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 853–866. https://doi.org/10.1145/3025453.3025485

Čech, J., & Soukupová, T. (2016). Real-time eye blink detection using facial landmarks. *21st Computer Vision Winter Workshop*.

Chang, C. T., & Chen, P. C. (2017). Cause-related marketing ads in the eye tracker: It depends on how you present, who sees the ad, and what you promote. *International Journal of Advertising*, *36*(2), 336–355. https://doi.org/10.1080/02650487.2015.1100698

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, *54*(4). https://doi.org/10.1145/3447744

Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine*, *110*(2), 111–124. https://doi.org/10.1016/j.cmpb.2012.10.021

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in hci. *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 315–318. https://doi.org/10.1145/1943403.1943454

Chen, Y., Yan, S., & Tran, C. C. (2019). Comprehensive evaluation method for user interface design in nuclear power plant based on mental workload. *Nuclear Engineering and Technology*, *51*(2), 453–462. https://doi.org/10.1016/j.net.2018.10.010

Cheng, S., Liu, X., Yan, P., Zhou, J., & Sun, S. (2010). Adaptive user interface of product recommendation based on eye-tracking. *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 94–101. https://doi.org/10.1145/2002333.2002348

Cheng, S., & Liu, Y. (2012). Eye-tracking based adaptive user interface: Implicit human-computer interaction for preference indication. *Journal on Multimodal User Interfaces*, *5*(1-2), 77–84. https://doi.org/10.1007/s12193-011-0064-6

Cherry, C. (1966). On human communication.

Cho, Y. (2021). Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445577

Chun, D. M., & Plass, J. L. (1997). Research on text comprehension in multimedia environments. *Language learning & technology*, *1*(1), 60–81.

Chung, E., Subramaniam, G., & Dass, L. C. (2020). Online learning readiness among university students in malaysia amidst covid-19. *Asian Journal of University Education*, *16*(2), 45–58. https://doi.org/10.24191/ajue.v16i2.10294

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. *Proceedings of the 6th International Conference on Intelligent User Interfaces*, 33–40. https://doi.org/10.1145/359784.359836

Clement, J., Kristensen, T., & Grønhaug, K. (2013). Understanding consumers' in-store visual perception: The influence of package design features on visual attention. *Journal of Retailing and Consumer Services*, *20*(2), 234–239. https://doi.org/10.1016/j.jretconser.2013.01.003

Coelho, J., & Duarte, C. (2011). The contribution of multimodal adaptation techniques to the GUIDE interface. *Proceedings of the 6th International Conference on Universal Access in Human-Computer Interaction: Design for All and EInclusion - Volume Part I*, 337–346.

Coelho, J., Duarte, C., Biswas, P., & Langdon, P. (2011). Developing accessible tv applications. *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, 131–138. https://doi.org/10.1145/2049536.2049561

Cohen, G., & Faulkner, D. (1983). Word recognition: Age differences in contextual facilitation effects. *British Journal of Psychology*, *74*(2), 239–251. https://doi.org/10.1111/j.2044-8295.1983.tb01860.x

Cohen, L. (2022). 9-year-old girl pens heartbreaking letter to her mom who died in ukraine: "we will meet in heaven". Retrieved July 6, 2022, from https://www.liverpoolecho.co.uk/news/showbiz-news/bbc-dad-delights-fans-update-23257826

Conati, C., Hoque, E., Toker, D., & Steichen, B. (2013). When to adapt: Detecting user's confusion during visualization processing. *User Modeling, Adaption, and Personalization: 21th International Conference*.

Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. *Cognitive Science*, *35*(4), 763–778. https://doi.org/10.1111/j.1551-6709.2010.01168.x

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., & Bandeira, N. (2017). "what can i help you with?": Infrequent users' experiences of intelligent personal assistants. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. https://doi.org/10.1145/3098279.3098539

Dalal, M., Feiner, S., McKeown, K., Pan, S., Zhou, M., Höllerer, T., Shaw, J., Feng, Y., & Fromer, J. (1997). Negotiation for automated generation of temporal multimedia presentations. *Proceedings of the Fourth ACM International Conference on Multimedia*, 55–64. https://doi.org/10.1145/244130.244147

Debie, E., Fernandez Rojas, R., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M., & Abbass, H. A. (2021). Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics*, *51*(3), 1542–1555. https://doi.org/10.1109/TCYB.2019.2939399

Dialpad. (2021). Closed captioning vs. live transcription: What's the difference? Retrieved May 24, 2022, from https://www.dialpad.com/blog/closed-captioning-vs-live-transcription/

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009). Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, *4*(3/4), 165–187. https://doi.org/10.1504/IJLT.2009.028805

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, *18*(1), 45–80.

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, *47*(3). https://doi.org/10.1145/2682899

Duarte, C., & Carriço, L. (2006). A conceptual framework for developing adaptive multimodal applications. *Proceedings of the 11th International Conference on Intelligent User Interfaces*, 132–139. https://doi.org/10.1145/1111449.1111481

Ducasse, J., Kljun, M., & Čopič Pucihar, K. (2020). Interactive web documentaries: A case study of audience reception and user engagement on iOtok. *International Journal of Human-Computer Interaction*, *36*(16), 1558–1584. https://doi.org/10.1080/10447318.2020.1757255

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, and Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human Machine Interaction*, *5440*, 1–25. https://doi.org/10.1007/978-3-642-00437-7_1

Durso, F. T., Geldbach, K. M., & Corballis, P. (2012). Detecting confusion using facial electromyography. *Human Factors*, *54*(1), 60–69. https://doi.org/10.1177/0018720811428450

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, *20*(3). https://doi.org/10.3390/s20030592

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: Manual*. Consulting Psychologists Press.

Elting, C., Zwickel, J., & Malaka, R. (2002). Device-dependant modality selection for user-interfaces: An empirical study. *Proceedings of the 7th International Conference on Intelligent User Interfaces*, 55–62. https://doi.org/10.1145/502716.502728

Engel, F. L. (1974). Visual conspicuity and selective background interference in eccentric vision. *Vision Research*, *14*(7), 459–471. https://doi.org/10.1016/0042-6989(74)90034-0

Engelke, U., & Le Callet, P. (2015). Perceived interest and overt visual attention in natural images. *Signal Processing: Image Communication*, *39*(B), 386–404. https://doi.org/10.1016/j.image.2015.03.004

Engen, K. J. V., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1908–1918. https://doi.org/10.1044/JSLHR-H-13-0076

Ennouamani, S., & Mahani, Z. (2017). An overview of adaptive e-learning systems. *2017 Eighth International Conference on Intelligent Computing and Information Systems*, 342–347. https://doi.org/10.1109/INTELCIS.2017.8260060

Erikson Institute. (2016). *Technology and young children in the digital age* (tech. rep.). Erikson Institute.

Etemadi, A. (2012). Effects of bimodal subtitling of english movies on content comprehension and vocabulary recognition. *International Journal of English Linguistics*, *2*(1), 239.

Fairclough, S. H., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, *56*, 171–184. https://doi.org/10.1016/j.ijpsycho.2004.11.003

Faure, V., Lobjois, R., & Benguigui, N. (2016). The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic Psychology and Behaviour*, *40*, 78–90. https://doi.org/10.1016/j.trf.2016.04.007

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024. https://doi.org/10.1177/0018720812443983

Fessl, A., Rivera-Pelayo, V., Pammer, V., & Braun, S. (2012). Mood tracking in virtual meetings. *European Conference on Technology Enhanced Learning*, 377–382. https://doi.org/10.1007/978-3-642-33263-0_30

Firmenich, S., Garrido, A., Paternò, F., & Rossi, G. (2019). User interface adaptation for accessibility. In Y. Yesilada & S. Harper (Eds.), *Web accessibility: A foundation for research* (pp. 547–568). Springer London. https://doi.org/10.1007/978-1-4471-7440-0_29

Fitbit. (2022). Trackers. Retrieved December 15, 2022, from https://www.fitbit.com/global/uk/products/trackers

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology*, *31*(2), 129–145. https://doi.org/10.1016/S0167-8760(98)00049-X

Fowler, A., Nesbitt, K., & Canossa, A. (2019). Identifying cognitive load in a computer game: An exploratory study of young children. *2019 IEEE Conference on Games*, 1–6. https://doi.org/10.1109/CIG.2019.8848064

Franke, N., Keinz, P., & Steger, C. J. (2009). Testing the value of customization: When do customers really prefer products tailored to their preferences? *Journal of Marketing*, *73*(5), 103–121. https://doi.org/10.1509/jmkg.73.5.103

Frisch, B., & Greene, C. (2020). What it takes to run a great virtual meeting. Retrieved August 16, 2022, from https://hbr.org/2020/03/what-it-takes-to-run-a-great-virtual-meeting

Fujii, K., & Rekimoto, J. (2019). Subme: An interactive subtitle system with english skill estimation using eye tracking. *Proceedings of the 10th Augmented Human International Conference 2019*. https://doi.org/10.1145/3311823.3311865

Fujiwara, K., Hoegen, R., Gratch, J., & Dunbar, N. E. (2022). Synchrony facilitates altruistic decision making for non-human avatars. *Computers in Human Behavior*, *128*, 107079. https://doi.org/10.1016/j.chb.2021.107079

Gal, D., & Simonson, I. (2021). Predicting consumers' choices in the age of the internet, ai, and almost perfect tracking: Some things change, the key challenges do not. *Consumer Psychology Review*, *4*(1), 135–152. https://doi.org/10.1002/arcp.1068

Galley, N., Betz, D., & Biniossek, C. (2015). Fixation durations - why are they so highly variable? In T. Heinen (Ed.), *Advances in visual perception research* (pp. 83–106). Nova Science Publishers, inc.

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, *56*, 1070–1085. https://doi.org/10.1080/00140139.2013.790483

Garcia Barrios, V., Gütl, C., Preis, A., Andrews, K., Pivec, M., Mödritscher, F., & Trummer, C. (2004). Adele: A framework for adaptive e-learning through eye tracking. *Proceedings of theInternational Conference on Knowledge Management and Knowledge Technologies*, 609–616.

Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 54–89). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_2

Gellevij, M., Meij, H. V. D., Jong, T. D., & Pieters, J. (2002). Multimodal versus unimodal instruction in a complex learning context. *The Journal of Experimental Education*, *70*(3), 215–239. https://doi.org/10.1080/00220970209599507

Genchi, W., Menoli, R., Parolo, D., Baggio, M., & Vecchia, A. D. (2019). Analysis of Lying through an Eye Blink Detection Algorithm. https://raw.githubusercontent.com/rmenoli/Eye-blinking-SVM/master/Report_final.pdf

Gifreu, A. (2013). The Interactive Documentary: My life, my passion, my playlist. Retrieved June 26, 2021, from https://docubase.mit.edu/playlist/the-interactive-documentary-my-life-my-passion-my-playlist/

Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. *Annals of the International Communication Association*, *10*(1), 13–48. https://doi.org/10.1080/23808985.1987.11678638

Gilroy, S., Porteous, J., Charles, F., & Cavazza, M. (2012). Exploring passive user interaction for adaptive narratives. *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, 119–128. https://doi.org/10.1145/2166966.2166990

Glassman, E. L., Kim, J., Monroy-Hernández, A., & Morris, M. R. (2015). Mudslide: A spatially anchored census of student confusion for online lecture videos. *Proceedings of the 33rd Annual*

ACM Conference on Human Factors in Computing Systems, 1555–1564. https://doi.org/10.1145/2702123.2702304

Google. (2022). Use captions in a meeting. Retrieved August 15, 2022, from https://support.google.com/meet/answer/9300310

Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2011). Predicting facial indicators of confusion with hidden markov models. *International Conference on Affective computing and intelligent interaction*, 97–106.

Graham, R., & Choo, J. (2022). Preliminary research on ai-generated caption accuracy rate by platforms and variables. *Journal on Technology and Persons with Disabilities*, *10*, 33–53.

Gupta, K., & Sharma, S. (2021). Kiosks as self-service technology in hotels: Opportunities and challenges. *Worldwide Hospitality and Tourism Themes*, *13*(2), 236–251. https://doi.org/10.1108/WHATT-10-2020-0125

Gürkök, H., Hakvoort, G., & Poel, M. (2011). Modality switching and performance in a thought and speech controlled computer game. *Proceedings of the 13th International Conference on Multimodal Interfaces*, 41–48. https://doi.org/10.1145/2070481.2070491

Gutierrez, M., Thalmann, D., & Vexo, F. (2005). Semantic virtual environments with adaptive multimodal interfaces. *11th International Multimedia Modelling Conference*, 277–283. https://doi.org/10.1109/MMMC.2005.65

Hakim, A., & Levy, D. J. (2019). A gateway to consumers' minds: Achievements, caveats, and prospects of electroencephalography-based prediction in neuromarketing. *WIREs Cognitive Science*, *10*(2), e1485. https://doi.org/10.1002/wcs.1485

Hansen, J., W. Andersen, A., & Roed, P. (1995). Eye-gaze control of multimedia systems. *Advances in Human Factors/Ergonomics*, *20*(100), 37–42. https://doi.org/10.1016/S0921-2647(06)80008-0

Hardoon, D. R., Shawe-Taylor, J., Ajanki, A., Puolamäki, K., & Kaski, S. (2007). Information retrieval by inferring implicit queries from eye movements. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, *2*, 179–186. https://proceedings.mlr.press/v2/hardoon07a.html

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183, Vol. 52). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, *28*(2), 202–223. https://doi.org/10.1287/mksc.1080.0459

Hayati, A., & Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of efl learners. *British Journal of Educational Technology*, *42*(1), 181–192. https://doi.org/10.1111/j.1467-8535.2009.01004.x

Heck, M., Edinger, J., & Becker, C. (2019). Gaze-based product filtering: A system for creating adaptive user interfaces to personalize stateless point-of-sale machines. *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 75–77. https://doi.org/10.1145/3332167.3357120

Heck, M., Edinger, J., Bünemann, J., & Becker, C. (2021). Exploring gaze-based prediction strategies for preference detection in dynamic interface elements. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 129–139. https://doi.org/10.1145/3406522.3446013

Ho, S. Y. (2006). The attraction of internet personalization to web users. *Electronic Markets*, *16*(1), 41–50. https://doi.org/10.1080/10196780500491162

Hoffmann, F., Tyroller, M.-I., Wende, F., & Henze, N. (2019). User-defined interaction for smart homes: Voice, touch, or mid-air gestures? *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. https://doi.org/10.1145/3365610.3365624

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 159–166. https://doi.org/10.1145/302979.303030

Hssina, B., & Erritali, M. (2019). A personalized pedagogical objectives based on a genetic algorithm in an adaptive learning system. *Procedia Computer Science*, *151*, 1152–1157. https://doi.org/10.1016/j.procs.2019.04.164

Hwang, Y. M., & Lee, K. C. (2018). Using an eye-tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human-Computer Interaction*, *34*(1), 15–24. https://doi.org/10.1080/10447318.2017.1314611

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*(1), 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Ibrahim, A., Lundberg, J., & Johansson, J. (2001). Speech enhanced remote control for media terminal. *7th European Conference on Speech Communication and Technology*, 2685–2688.

Iio, T., Shiomi, M., Shinozawa, K., Shimohara, K., Miki, M., & Hagita, N. (2015). Lexical entrainment in human robot interaction. *International Journal of Social Robotics*, *7*(2), 253–263. https://doi.org/10.1007/s12369-014-0255-x

Infineon. (2022). Infotainment. Retrieved December 15, 2022, from https://www.infineon.com/cms/de/applications/automotive/infotainment/

Ergonomics of human-system interaction - part 11: Usability: Definitions and concepts (Standard). (2018). International Organization for Standardization. Geneva, Switzerland.

Itti, L., Koch, C., & Niebur, E. (1988). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1017/S1042771600010292

Jacob, R. J. K. (1991). The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, *9*(2), 152–169. https://doi.org/10.1145/123078.128728

Jacobs, S. (1999). Section 255 of the telecommunications act of 1996: Fueling the creation of new electronic curbcuts. Retrieved August 30, 2022, from http://www.accessiblesociety.org/topics/technology/eleccurbcut.htm

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

Jameson, A. (2007). Adaptive interfaces and agents. In *The human-computer interaction handbook* (pp. 459–484). CRC press.

Jeng, T. (2009). Toward a ubiquitous smart space design framework. *Journal of Information Science & Engineering*, *25*(3). https://doi.org/10.6688/JISE.2009.25.3.1

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Communication*, *45*(4), 455–470. https://doi.org/10.1016/j.specom.2004.12.004

Jiménez, J., Iglesias, A. M., López, J. F., Hernández, J., & Ruiz, B. (2011). Tablet pc and head mounted display for live closed captioning in education. *2011 IEEE International Conference on Consumer Electronics*, 885–886. https://doi.org/10.1109/ICCE.2011.5722919

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2017a). Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, *51*(1), 4–11. https://doi.org/10.1145/3130332.3130334

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2017b). Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, *51*(1), 4–11. https://doi.org/10.1145/3130332.3130334

Jokinen, K., & Hurtig, T. (2006). User expectations and real experience on a multimodal interactive system. *INTERSPEECH 2006*, 1049–1052. https://doi.org/10.21437/Interspeech.2006-156

Jones, G. (2022). 'BBC dad' delights fans with update on kids five years after 'best TV moment ever'. Retrieved July 6, 2022, from https://www.liverpoolecho.co.uk/news/showbiz-news/bbc-dad-delights-fans-update-23257826

Josifovska, K., Yigitbas, E., & Engels, G. (2019). A digital twin-based multi-modal ui adaptation framework for assistance systems in industry 4.0. *Human-Computer Interaction. Design Practice in Contemporary Societies*, 398–409. https://doi.org/10.1007/978-3-030-22636-7_30

Jyotsna, C., & Amudha, J. (2018). Eye gaze as an indicator for stress level analysis in students. *2018 International Conference on Advances in Computing, Communications and Informatics*, 1588–1593. https://doi.org/10.1109/ICACCI.2018.8554715

Kalyuga, S. (2012). Instructional benefits of spoken words: A review of cognitive load factors. *Educational Research Review*, *7*(2), 145–159. https://doi.org/10.1016/j.edurev.2011.12.002

Kandemir, M., Saarinen, V.-M., & Kaski, S. (2010). Inferring object relevance from gaze in dynamic scenes. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 105–108. https://doi.org/10.1145/1743666.1743692

Karl, K. A., Peluchette, J. V., & Aghakhani, N. (2022). Virtual work meetings during the COVID-19 pandemic: The good, bad, and ugly. *Small Group Research*, *53*, 343–365. https://doi.org/10.1177/10464964211015286

Karpov, A., & Yusupov, R. (2018). Multimodal interfaces of human-computer interaction. *Herald of the Russian Academy of Sciences*, *88*(1), 67–74.

Karray, F., Alemzadeh, M., Abou Saleh, J., & Arab, M. N. (2008). Human-computer interaction: Overview on state of the art. *International journal on smart sensing and intelligent systems*, *1*(1), 137. https://doi.org/10.21307/ijssis-2017-283

Kazienko, P., & Adamski, M. (2007). AdROSA - adaptive personalization of web advertising. *Information Sciences*, *177*(11), 2269–2295. https://doi.org/10.1016/j.ins.2007.01.002

Kephart, J., & Chess, D. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41–50. https://doi.org/10.1109/MC.2003.1160055

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *International Journal of Engineering Research and Application*, *7*(03), 20–24. https://doi.org/10.9790/9622-0703022024

Kerpedjiev, S., Carenini, G., Roth, S. F., & Moore, J. D. (1997). Integrating planning and task-based design for multimedia presentation. *Proceedings of the 2nd international conference on Intelligent user interfaces*, 145–152.

Kim, D., Gluck, J., Hall, M., & Agarwal, Y. (2019). Real world longitudinal ios app usage study at scale. *arXiv preprint arXiv:1912.12526*. https://doi.org/10.48550/arXiv.1912.12526

Kim, J. W., Lee, B. H., Shaw, M. J., Chang, H. L., & Nelson, M. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *International Journal of Electronic Commerce*, *5*(3), 45–62. https://doi.org/10.1080/10864415.2001.11044215

Kim, J. C., Laine, T. H., & Åhlund, C. (2021). Multimodal interaction systems based on internet of things and augmented reality: A systematic literature review. *Applied Sciences*, *11*(4). https://doi.org/10.3390/app11041738

Kimura, A. (2020). Pysaliencymap. Retrieved January 14, 2023, from https://github.com/akisatok/pySaliencyMap/

Kimura, A. K. (2018). Defining, evaluating, and achieving accessible library resources: A review of theories and methods. *Reference Services Review*, *46*(3), 425–438. https://doi.org/10.1108/RSR-03-2018-0040

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kiyota, A. (2022). Problematizing fluent speakers' unintentional exclusion of emergent bilinguals: A case study of an english-medium instruction classroom in japan. *International Journal of Literacy, Culture, and Language Education*, *2*(Special Issue: Language Weaponization in Society and Education), 6–19. https://doi.org/10.14434/ijlcle.v2iMay.34385

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, *48*(1), 45–77. https://doi.org/10.1007/s11251-020-09502-9

Klitmøller, A., & Lauring, J. (2013). When global virtual teams share knowledge: Media richness, cultural difference and language commonality. *Journal of World Business*, *48*(3), 398–406. https://doi.org/10.1016/j.jwb.2012.07.023

Koh, J. I., Ray, S., Cherian, J., Taele, P., & Hammond, T. (2022). Show of hands: Leveraging hand gestural cues in virtual meetings for intelligent impromptu polling interactions. *27th International Conference on Intelligent User Interfaces*, 292–309. https://doi.org/10.1145/3490099.3511153

Kong, J., Zhang, W., Yu, N., & Xia, X. (2011). Design of human-centric adaptive multimodal interfaces. *International Journal of Human-Computer Studies*, *69*(12), 854–869. https://doi.org/10.1016/j.ijhcs.2011.07.006

Kozma, L., Klami, A., & Kaski, S. (2009). GaZIR: Gaze-based zooming interface for image retrieval. *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 305–312. https://doi.org/10.1145/1647314.1647379

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2176–2184. https://doi.org/10.1109/CVPR.2016.239

Krupitzer, C., Roth, F. M., VanSyckel, S., Schiele, G., & Becker, C. (2015). A survey on engineering approaches for self-adaptive systems. *Pervasive and Mobile Computing*, *17*, 184–206. https://doi.org/10.1016/j.pmcj.2014.09.009

Kulkarni, M. (2019). Digital accessibility: Challenges and opportunities. *IIMB Management Review*, *31*(1), 91–98. https://doi.org/10.1016/j.iimb.2018.05.009

Kumari, P., Mathew, L., & Syal, P. (2017). Increasing trend of wearables and multimodal interface for human activity monitoring: A review. *Biosensors and Bioelectronics*, *90*, 298–307. https://doi.org/10.1016/j.bios.2016.12.001

Kunze, K., Kawaichi, H., Yoshimura, K., & Kise, K. (2013). Towards inferring language expertise using eye tracking. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 217–222. https://doi.org/10.1145/2468356.2468396

Lai, T. Y., & Hwang, G. H. (2010). Application of half-life theory and fuzzy theory to a selection and recommendation system for web advertisement delivery in consideration of the time effect. *WSEAS Transactions on Information Science and Applications*, *7*(1), 70–80.

Lalanne, D., Nigay, L., Palanque, p., Robinson, P., Vanderdonckt, J., & Ladry, J.-F. (2009). Fusion engines for multimodal input: A survey. *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 153–160. https://doi.org/10.1145/1647314.1647343

Lallé, S., Conati, C., & Carenini, G. (2016). Predicting confusion in information visualization from eye tracking and interaction data. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2529–2535.

Lange, E. B., Thiele, D., & Kuijpers, M. M. (2022). Narrative aesthetic absorption in audiobooks is predicted by blink rate and acoustic features. *Psychology of Aesthetics, Creativity, and the Arts*, *16*(1), 110–124.

Langheinrich, M., Nakamura, A., Abe, N., Kamba, T., & Koseki, Y. (1999). Unintrusive customization techniques for web advertising. *Computer Networks*, *31*(11-16), 1259–1272. https://doi.org/10.1016/S1389-1286(99)00033-X

Langley, P. (1999). User modeling in adaptive interface. *CISM International Centre for Mechanical Sciences*, 357–370. https://doi.org/10.1007/978-3-7091-2490-1_48

Laput, G. P., Dontcheva, M., Wilensky, G., Chang, W., Agarwala, A., Linder, J., & Adar, E. (2013). Pixeltone: A multimodal interface for image editing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2185–2194. https://doi.org/10.1145/2470654.2481301

Larkin, T. (2021). Zoom's auto-generated captions available to all free users. Retrieved August 15, 2022, from https://blog.zoom.us/zoom-auto-generated-captions/

Lee, H., & Kobsa, A. (2017). Privacy preference modeling and prediction in a simulated campuswide iot environment. *2017 IEEE International Conference on Pervasive Computing and Communications*, 276–285. https://doi.org/10.1109/PERCOM.2017.7917874

Lee, H.-K., Suh, K.-S., & Benbasat, I. (2001). Effects of task-modality fit on user performance [Decision-making and E-Commerce Systems]. *Decision Support Systems*, *32*(1), 27–40. https://doi.org/10.1016/S0167-9236(01)00098-7

Lee, M., Billinghurst, M., Baek, W., Green, R., & Woo, W. (2013). A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, *17*(4), 293–305.

Lenskiy, A., & Paprocki, R. (2016). Blink rate variability during resting and reading sessions. *2016 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 1–6. https://doi.org/10.1109/NORBERT.2016.7547466

Lewis, J. R. (1995). Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Li, J., Ngai, G., Va Leong, H., & Chan, S. (2016). Multimodal human attention detection for reading. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 187–192. https://doi.org/10.1145/2851613.2851681

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing (Early Access*. https://doi.org/10.1109/TAFFC.2020.2981446

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (cert). *2011 IEEE International Conference on*

Automatic Face & Gesture Recognition (FG), 298–305. https://doi.org/10.1109/FG.2011.5771414

Loijens, L., & Krips, O. (2021). *Facereader: Methodology note* (tech. rep.). Noldus Information Technology. Wageningen, Netherlands.

Luger, E., & Sellen, A. (2016). "like having a really bad PA": The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Maat, L., & Pantic, M. (2007). Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. *Artifical Intelligence for Human Computing, LNCS*, *4451*, 251–271. https://doi.org/10.1007/978-3-540-72348-6_13

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects information details within picutres. *Perception & Psychophysics*, *2*(11), 547–552. https://doi.org/10.3758/BF03210264

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human-computer interaction. In *Advances in physiological computing* (pp. 39–65). Springer London. https://doi.org/10.1007/978-1-4471-6392-3_3

Mallick, R., Slayback, D., Touryan, J., Ries, A. J., & Lance, B. J. (2016). The use of eye metrics to index cognitive workload in video games. *2016 IEEE Second Workshop on Eye Tracking and Visualization*, 60–64.

Martin, J.-C. (1998). Tycoon: Theoretical framework and software tools for multimodal interfaces. *Intelligence and Multijodality in Multimedia Interfaces*, *9*.

Matz, S. C., & Netzer, O. (2017). Using Big Data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences*, *18*, 7–12. https://doi.org/10.1016/j.cobeha.2017.05.009

Maybury, M. (1994). Intelligent multimedia interfaces. *Conference Companion on Human Factors in Computing Systems*, 423–424.

Maybury, M. T., & Wahlster, W. (1999). Intelligent user interfaces: An introduction. *Readings in Intelligent User Interfaces*, 1–13.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Meißner, M., Pfeiffer, J., Pfeiffer, T., & Oppewal, H. (2019). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. *Journal of Business Research*, *100*, 445–458. https://doi.org/10.1016/j.jbusres.2017.09.028

Melichar, M., & Cenek, P. (2006). From vocal to multimodal dialogue management. *Proceedings of the 8th International Conference on Multimodal Interfaces*, 59–67. https://doi.org/10.1145/1180995.1181008

Microsoft. (2012). Microsoft Speech API (SAPI) 5.3. Retrieved June 28, 2021, from https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ms723627(v=vs.85)

Microsoft. (2021). Cortana - your personal productivity assistant in Microsoft 365. Retrieved June 10, 2021, from https://www.microsoft.com/en-us/cortana

Microsoft. (2022). View live transcription in a teams meeting. Retrieved August 15, 2022, from https://support.microsoft.com/en-us/office/view-live-transcription-in-a-teams-meeting-dc1a8f23-2e20-4684-885e-2152e06a4a8b

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142–151. https://doi.org/10.1145/345124.345169

Moniri, M. M., Feld, M., & Müller, C. (2012). Personalized in-vehicle information systems: Building an application infrastructure for smart cars in smart spaces. *2012 Eighth International Conference on Intelligent Environments*, 379–382. https://doi.org/10.1109/IE.2012.40

Morris, K. K., Frechette, C., Dukes III, L., Stowell, N., Topping, N. E., & Brodosi, D. (2016). Closed captioning matters: Examining the value of closed captions for "all" students. *Journal of Postsecondary Education and Disability*, *29*(3), 231–238.

Morris, M. R. (2012). Web on the wall: Insights from a multimodal interaction elicitation study. *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, 95–104. https://doi.org/10.1145/2396636.2396651

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, *87*(2), 319. https://doi.org/10.1037/0022-0663.87.2.319

Mousavinasab, E., Zarifsanaiey, N., Kalhori, S. R. N., Rakhshan, M., Keikha, L., & Saeedi, M. G. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the mcgurk effect. *Perception & psychophysics*, *58*(3), 351–362. https://doi.org/10.3758/BF03206811

Muñoz-Leiva, F., Hernández-Méndez, J., & Gómez-Carmona, D. (2019). Measuring advertising effectiveness in Travel 2.0 websites through eye-tracking technology. *Physiology and Behavior*, *200*, 83–95. https://doi.org/10.1016/j.physbeh.2018.03.002

Murali, P. K., Kaboli, M., & Dahiya, R. (2022). Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, *4*(2), 2100122. https://doi.org/10.1002/aisy.202100122

Murali, P., Hernandez, J., McDuff, D., Rowan, K., Suh, J., & Czerwinski, M. (2021). AffectiveSpotlight: Facilitating the communication of affective responses from audience members during online presentations. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445235

Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education*, *141*, 103618. https://doi.org/10.1016/j.compedu.2019.103618

Nakano, T., Yamamoto, Y., Kitajo, K., Takahashi, T., & Kitazawa, S. (2009). Synchronization of spontaneous eyeblinks while viewing video stories. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1673), 3635–3644. https://doi.org/10.1098/rspb.2009.0828

Nasoz, F., Lisetti, C. L., & Vasilakos, A. V. (2010). Affectively intelligent and adaptive car interfaces. *Information Sciences*, *180*(20), 3817–3836. https://doi.org/10.1016/j.ins.2010.06.034

Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1973–1976. https://doi.org/10.1145/1056808.1057070

Netflix. (2017). Interactive storyteling on Netflix: Choose what happens next. Retrieved June 11, 2021, from https://about.netflix.com/en/news/interactive-storytelling-on-netflix-choose-what-happens-next

Netflix. (2018). Black Mirror: Bandersnatch. Retrieved September 19, 2019, from https://media.netflix.com/en/only-on-netflix/80988062

Netflix. (2021a). Animals on the loose: A you vs. wild movie. Retrieved June 26, 2021, from https://media.netflix.com/en/only-on-netflix/81205737

Netflix. (2021b). You vs. wild. Retrieved June 26, 2021, from https://media.netflix.com/en/only-on-netflix/80227574

Ng, H.-W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 443–449. https://doi.org/10.1145/2818346.2830593

Nicholson, A.-M., & Ed Sheeran. (2018). 2002. Retrieved September 19, 2019, from https://www.youtube.com/watch?v=u3ePPA0yzSU

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. https://doi.org/10.1002/asi

O'Brien, S. (2009). Eye tracking in translation process research: Methodological challenges and solutions. *Methodology, technology and innovation in translation process research*, *38*, 251–266.

OECD. (2022). Ict access and usage by household and individuals. Retrieved December 16, 2022, from https://stats.oecd.org/Index.aspx?DataSetCode=ICT_HH2

Office of Communications. (2006). *Television access services* (tech. rep.). Ofcom. https://www.ofcom.org.uk/___data/assets/pdf_file/0016/42442/access.pdf

Olsen, A. (2012). *The Tobii I-VT fixation filter* (tech. rep.). Tobii Technology.

Oppermann, R. (1994). *Adaptive user support: Ergonomic design of manually and automatically adaptable software*. L. Erlbaum Associates Inc.

Orden, K. F. V., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, *43*(1), 111–121. https://doi.org/10.1518/001872001775992570

Otterbring, T., Wästlund, E., Gustafsson, A., & Shams, P. (2014). Vision (im)possible? the effects of in-store signage on customers' visual attention. *Journal of Retailing and Consumer Services*, *21*(5), 676–684. https://doi.org/10.1016/j.jretconser.2014.05.002

Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., & Huang, D.-Y. (2017). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 577–582. https://doi.org/10.1145/3136755.3143012

Oviatt, S. (1997). Mulitmodal interactive maps: Designing for human performance. *Human–Computer Interaction*, *12*(1-2), 93–129. https://doi.org/10.1080/07370024.1997.9667241

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, *42*(11), 74–81. https://doi.org/10.1145/319382.319398

Oviatt, S. (2003a). Advances in robust multimodal interface design. *IEEE Comput. Graph. Appl.*, *23*(5), 62–68. https://doi.org/10.1109/MCG.2003.1231179

Oviatt, S. (2003b). Flexible and robust multimodal interfaces for universal access. *Universal access in the information society*, *2*(2), 91–95. https://doi.org/10.1007/s10209-002-0041-7

Oviatt, S. (2006). Human-centered design meets cognitive load theory: Designing interfaces that help people think. *Proceedings of the 14th ACM International Conference on Multimedia*, 871–880. https://doi.org/10.1145/1180639.1180831

Oviatt, S. (2017). Theoretical foundations of multimodal interfaces and systems. In *The handbook of multimodal-multisensor interfaces: Foundations, user modeling, and common modality combinations - volume 1* (pp. 19–50). ACM; Morgan & Claypool. https://doi.org/10.1145/3015783.3015786

Oviatt, S., Bernard, J., & Levow, G.-A. (1998). Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech*, *41*(3-4), 419–442. https://doi.org/10.1177/002383099804100409

Oviatt, S., Cohen, P., Wu, L., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., & Ferro, D. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human–Computer Interaction*, *15*(4), 263–322. https://doi.org/10.1207/S15327051HCI1504_1

Oviatt, S., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? cognitive load and multimodal communication patterns. *Proceedings of the 6th International Conference on Multimodal Interfaces*, 129–136. https://doi.org/10.1145/1027933.1027957

Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., & Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 44–51. https://doi.org/10.1145/958432.958443

Oviatt, S., Lunsford, R., & Coulston, R. (2005). Individual differences in multimodal integration patterns: What are they and why do they exist? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 241–249. https://doi.org/10.1145/1054972.1055006

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, *32*(1/2), 1–8.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3839–3845.

Park, S., Zhang, X., Bulling, A., & Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. https://doi.org/10.1145/3204493.3204545

Partan, S. R. (2013). Ten unanswered questions in multimodal communication. *Behavioral Ecology and Sociobiology*, *67*(9), 1523–1539. https://doi.org/10.1007/s00265-013-1565-y

Patel, S. (2016). 85 percent of facebook video is watched without sound. *Digiday*. Retrieved December 19, 2022, from https://digiday.com/media/silent-world-facebook-video/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peissner, M., Häbe, D., Janssen, D., & Sellner, T. (2012). MyUI: Generating accessible user interfaces from multimodal design patterns. *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 81–90. https://doi.org/10.1145/2305484.2305500

Peitek, N., Siegmund, J., Parnin, C., Apel, S., & Brechmann, A. (2018). Beyond gaze: Preliminary analysis of pupil dilation and blink rates in an fmri study of program comprehension. *Proceedings of the Workshop on Eye Movements in Programming*. https://doi.org/10.1145/3216723.3216726

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, *122*, 106855. https://doi.org/10.1016/j.chb.2021.106855

Peng, F., LaBelle, V. C., Yue, E. C., & Picard, R. W. (2018). A trip to the moon: Personalized animated movies for self-reflection. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3173574.3173827

Perego, E., Missier, F. D., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, *13*(3), 243–272. https://doi.org/10.1080/15213269.2010.502873

Picard, R. W. (2000). *Affective computing* (tech. rep.). MIT Media Laboratory.

Plutchik, R. (1991). *The emotions*. University Press of America.

Poehnl, S., & Bogner, F. X. (2013). Cognitive load and alternative conceptions in learning genetics: Effects from provoking confusion. *The Journal of Educational Research*, *106*(3), 183–196. https://doi.org/10.1080/00220671.2012.687790

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Poulton, G. (2020). Why trust and autonomy are essential factors when working from home. Retrieved September 9, 2021, from https://www.rolandberger.com/en/Insights/Publications/The-home-office-becomes-the-new-normal.html

Prabhakaran, B. (2000). Adaptive multimedia presentation strategies. *Multimedia Tools and Applications*, *12*(2), 281–298. https://doi.org/10.1023/A:1009627926302

Prange, S., George, C., & Alt, F. (2021). Design considerations for usable authentication in smart homes. *Mensch Und Computer 2021*, 311–324. https://doi.org/10.1145/3473856.3473878

Profanter, S., Perzylo, A., Somani, N., Rickert, M., & Knoll, A. (2015). Analysis and semantic modeling of modality preferences in industrial human-robot interaction. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1812–1818. https://doi.org/10.1109/IROS.2015.7353613

Python Software Foundation. (2021). Graphical user interfaces with Tk. Retrieved June 11, 2021, from https://docs.python.org/3/library/tkinter.html

Qvarfordt, P., & Zhai, S. (2005). Conversing with the user based on eye-gaze patterns. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 221–230. https://doi.org/10.1145/1054972.1055004

Radlinski, F., & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 239–248. https://doi.org/10.1145/1081870.1081899

Raja, D. S. (2016). *Bridging the disability divide through digital technologies* (tech. rep.). World Bank Group.

Rebman, C., Reithel, B., & Cegielski, C. (2001). An exploratory study of speech recognition technology and its implications for current electronic meeting support applications. *7th Americas Conference on Information Systems*, 298–306. https://aisel.aisnet.org/amcis2001/61

Reeves, L. M., Lai, J., Larson, J. A., Oviatt, S., Balaji, T. S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.-C., McTear, M., Raman, T., Stanney, K. M., Su, H., & Wang, Q. Y. (2004). Guidelines for multimodal user interface design. *Commun. ACM*, *47*(1), 57–59. https://doi.org/10.1145/962081.962106

Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., & Tschernomas, V. (2003). Smartkom: Adaptive and flexible multimodal access to multiple applications. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 101–108. https://doi.org/10.1145/958432.958454

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58. https://doi.org/10.1145/245108.245121

Revadekar, A., Oak, S., Gadekar, A., & Bide, P. (2020). Gauging attention of students in an e-learning environment. *2020 IEEE 4th Conference on Information & Communication Technology*, 1–6. https://doi.org/10.1109/CICT51604.2020.9312048

Ribeiro-Neto, B., Cristo, M., Golgher, P. B., & Silva de Moura, E. (2005). Impedance coupling in content-targeted advertising. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 496–503. https://doi.org/10.1145/1076034.1076119

Rickel, E., Harris, K., Mandile, E., Pagliari, A., Derby, J. L., & Chaparro, B. S. (2022). Typing in mid air: Assessing one- and two-handed text input methods of the microsoft hololens 2. *Virtual, Augmented and Mixed Reality: Design and Development*, 357–368. https://doi.org/10.1007/978-3-031-05939-1_24

Robison, K. (2019). Player control tests. Retrieved August 16, 2022, from https://about.netflix.com/en/news/player-control-tests

Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, *12*(2), 479–522. https://doi.org/10.1007/s40685-018-0069-z

Rothrock, L., Koubek, R., Fuchs, F., Haas, M., & Salvendy, G. (2002). Review and reappraisal of adaptive interfaces: Toward biologically inspired paradigms. *Theoretical Issues in Ergonomics Science*, *3*(1), 47–84. https://doi.org/10.1080/14639220110110342

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, *35*, 991–1009. https://doi.org/10.1016/j.ergon.2005.04.005

Sae-Bae, N., Wu, J., Memon, N., Konrad, J., & Ishwar, P. (2019). Emerging nui-based methods for user authentication: A new taxonomy and survey. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *1*(1), 5–31. https://doi.org/10.1109/TBIOM.2019.2893297

Saha, D., & Mukherjee, A. (2003). Pervasive computing: A paradigm for the 21st century. *Computer*, *36*(3), 25–31. https://doi.org/10.1109/MC.2003.1185214

Salojärvi, J., Puolamäki, K., & Kaski, S. (2004). *Relevance feedback from eye movements for proactive information retrieval* (tech. rep.). Helsinki University of Technology.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *The American Journal of Psychology*, *93*(2), 207–234. https://www.jstor.org/stable/1422228

Sandhu, R., & Dyson, B. J. (2012). Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses. *Acta Psychologica*, *140*(2), 111–118. https://doi.org/10.1016/j.actpsy.2012.04.003

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics*, *36*(5), 439–445. https://doi.org/10.1016/j.ergon.2006.01.007

Sathik, M., & Jonathan, S. G. (2013). Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus*, *2*, 455. https://doi.org/10.1186/2193-1801-2-455

Schaeffner, S., Koch, I., & Philipp, A. M. (2016). The role of sensory-motor modality compatibility in language processing. *Psychological research*, *80*(2), 212–223. https://doi.org/10.1007/s00426-015-0661-1

Schaffer, S., Schleicher, R., & Möller, S. (2011). Measuring cognitive load for different input modalities. *9. Berliner Werkstatt Mensch-Maschine-Systeme*, 287–292.

Schaffer, S., Schleicher, R., & Möller, S. (2015). Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human-Computer Studies*, *75*, 21–34. https://doi.org/10.1016/j.ijhcs.2014.11.004

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *1994 First Workshop on Mobile Computing Systems and Applications*, 85–90. https://doi.org/10.1109/WMCSA.1994.16

Schmidt, A., Beigl, M., & Gellersen, H.-W. (1999). There is more to context than location. *Computers & Graphics*, *23*(6), 893–901. https://doi.org/10.1016/S0097-8493(99)00120-X

Schneegass, C., Kosch, A., Thomas and Schmidt, & Hussmann, H. (2019). Investigating the potential of eeg for implicit detection of unknown words for foreign language learning. *Human-Computer Interaction*, 293–313.

Schweikert, C., Gobin, L., Xie, S., Shimojo, S., & Hsu, D. F. (2018). Preference prediction based on eye movement using multi-layer combinatorial fusion. *International Conference on Brain Informatics*, 282–293. https://doi.org/10.1007/978-3-030-05587-5_27

Sebe, N. (2009). Multimodal interfaces: Challenges and perspectives. *Journal of Ambient Intelligence and smart environments*, *1*(1), 23–30.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The mcgurk effect in chinese subjects. *Perception & Psychophysics*, *59*(1), 73–80. https://doi.org/10.3758/BF03206849

Sekiyama, K., & Tohkura, Y. (1991). Mcgurk effect in non-english listeners: Few visual effects for japanese subjects hearing japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797–1805. https://doi.org/10.1121/1.401660

Seligman, M., & Dillinger, M. (2006). Usability issues in an interactive speech-to-speech translation system for healthcare. *Proceedings of the Workshop on Medical Speech Translation*, 1–4.

Sezer, O. B., Dogdu, E., & Ozbayoglu, A. M. (2018). Context-aware computing, learning, and big data in internet of things: A survey. *IEEE Internet of Things Journal*, *5*(1), 1–27. https://doi.org/10.1109/JIOT.2017.2773600

Shatnawi, M., & Mohamed, N. (2012). Statistical techniques for online personalized advertising: A survey. *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 680–687. https://doi.org/10.1145/2245276.2245406

Shenk, J., G., A. C., Arriaga, O., & Owlwasrowk. (2021). Justinshenk/fer: Zenodo. *Zenodo*. https://doi.org/10.5281/zenodo.5362356

Shi, Z., Zhang, Y., Bian, C., & Lu, W. (2019). Automatic academic confusion recognition in online learning based on facial expressions. *2019 14th International Conference on Computer Science & Education*, 528–532. https://doi.org/10.1109/ICCSE.2019.8845348

Shneiderman, B. (1997). Direct manipulation for comprehensible, predictable and controllable user interfaces. *Proceedings of the 2nd international conference on Intelligent user interfaces*, 33–39.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504.

Shockley, K. M., Gabriel, A. S., Robertson, D., Rosen, C. C., Chawla, N., Ganster, M. L., & Ezerins, M. E. (2021). The fatiguing effects of camera use in virtual meetings: A within-person field experiment. *Journal of Applied Psychology*, *106*, 1137–1155. https://doi.org/10.1037/apl0000948

Shokeen, J., & Rana, C. (2018). A review on the dynamics of social recommender systems. *International Journal of Web Engineering and Technology*, *13*(3). https://doi.org/10.1504/IJWET.2018.10016164

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*, 679–687. https://doi.org/10.1111/j.1469-8986.2008.00681.x

Sikander, G., & Anwar, S. (2019). Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, *20*(6), 2339–2352. https://doi.org/10.1109/TITS.2018.2868499

Simon, H. A. (1988). The science of design: Creating the artificial. *Design Issues*, *4*(1/2), 67–82. https://doi.org//10.2307/1511391

Skulmowski, A., & Xu, K. M. (2021). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 1–26. https://doi.org/10.1007/s10648-021-09624-7

Slanzi, G., Balazs, J. A., & Velásquez, J. D. (2017). Combining eye tracking, pupil dilation and eeg analysis for predicting web users click intention. *Information Fusion*, *35*, 51–57. https://doi.org/10.1016/j.inffus.2016.09.003

Slessor, G., Laird, G., Phillips, L. H., Bull, R., & Filippou, D. (2010). Age-related differences in gaze following: Does the age of the face matter? *The Journals of Gerontology: Series B*, *65B*(5), 536–541. https://doi.org/10.1093/geronb/gbq038

Smith, A. L., & Chaparro, B. S. (2015). Smartphone text input method performance, usability, and preference with younger and older adults. *Human Factors*, *57*(6), 1015–1028. https://doi.org/10.1177/0018720815575644

Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S. H., AlGhamdi, M. A., Nagra, A. A., & Masood, K. (2022). Racial identity-aware facial expression recognition using deep convolutional neural networks. *Applied Sciences*, *12*(1). https://doi.org/10.3390/app12010088

Spina, C. (2021). Captions. *Library Technology Reports*, *57*(3), 7–12.

Starker, I., & Bolt, R. A. (1990). A gaze-responsive self-disclosing display. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3–10. https://doi.org/10.1145/97243.97245

Steil, J., Hagestedt, I., Huang, M. X., & Bulling, A. (2018). Privacy-aware eye tracking using differential privacy. *CoRR*, *abs/1812.0*, 1–9. https://doi.org/10.1145/3314111.3319915

Stelzel, C., & Schubert, T. (2011). Interference effects of stimulus–response modality pairings in dual tasks and their robustness. *Psychological Research*, *75*(6), 476–490. https://doi.org/10.1007/s00426-011-0368-x

Stephan, D. N., & Koch, I. (2010). Central cross-talk in task switching: Evidence from manipulating input–output modality compatibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1075. https://doi.org/10.1037/a0019695

Stephan, D. N., & Koch, I. (2011). The role of input–output modality compatibility in task switching. *Psychological Research*, *75*(6), 491–498. https://doi.org/10.1007/s00426-011-0353-4

Stephanidis, C., Paramythis, A., Sfyrakis, M., Stergiou, A., Maou, N., Leventis, A., Paparoulis, G., & Karagiannidis, C. (1998). Adaptable and adaptive user interfaces for disabled users in the AVANTI project. *Intelligence in Services and Networks: Technology for Ubiquitous Telecom Services*, 153–166. https://doi.org/10.1007/BFb0056962

STMicroelectronics. (2022). In-vehicle infotainment (ivi). Retrieved December 15, 2022, from https://www.st.com/en/applications/in-vehicle-infotainment-ivi.html

Stokel-Walker, C. (2021). Microsoft teams ai couldt ell you who is most enjoying your video call. Retrieved April 6, 2022, from https://www.newscientist.com/article/2267147-microsoft-teams-ai-could-tell-you-who-is-most-enjoying-your-video-call/

Subaidi bin Abdul Samat, M., & Aziz, A. A. (2020). The effectiveness of multimedia learning in enhancing reading comprehension among indigenous pupils. *Arab World English Journal*, *11*(2). https://doi.org/10.2139/ssrn.3649324

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699.

Sullivan, S., Campbell, A., Hutton, S. B., & Ruffman, T. (2015). What's good for the goose is not good for the gander: Age and gender differences in scanning emotion faces. *The Journals of Gerontology: Series B*, *72*(3), 441–447. https://doi.org/10.1093/geronb/gbv033

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, *8*(4), 351–362. https://doi.org/10.1207/s1532690xci0804_5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Taib, R., Berkovsky, S., Koprinska, I., Wang, E., Zeng, Y., & Li, J. (2020). Personality sensing: Detection of personality traits using physiological responses to image and video stimuli. *ACM Trans. Interact. Intell. Syst.*, *10*(3). https://doi.org/10.1145/3357459

Tam, K. Y., & Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research*, *16*(3), 271–291. https://doi.org/10.1287/isre.1050.0058

Tasty. (2016). 4 burgers around the world. Retrieved September 19, 2019, from https://www.youtube.com/watch?v=Rix%7B%5C_%7D3b9ThLI

Tasty. (2019). 4 mind blowing ice cream tasty desserts. Retrieved September 19, 2019, from https://www.youtube.com/watch?v=VzHcBimEjh8%7B%5C&%7Dt=76s

Tchankue, P., Wesson, J., & Vogts, D. (2011). The impact of an adaptive user interface on reducing driver distraction. *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 87–94. https://doi.org/10.1145/2381416.2381430

TechViral. (2022). How to play youtube audio only on desktop to save bandwidth. Retrieved December 19, 2022, from https://techviral.net/play-youtube-audio-only-desktop/

Tian, Y.-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In *Handbook of face recognition* (pp. 247–275). Springer New York. https://doi.org/10.1007/0-387-27257-7_12

Tian, Y., Zhou, K., & Pelleg, D. (2021). What and how long: Prediction of mobile app engagement. *ACM Transactions on Information Systems*, *40*(1). https://doi.org/10.1145/3464301

Toledano, D. T., Fernández Pozo, R., Hernández Trapote, Á., & Hernández Gómez, L. (2006). Usability evaluation of multi-modal biometric verification systems. *Interacting with Computers*, *18*(5), 1101–1122.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Tomkins, S. S. (1984). Affect theory. *Approaches to emotion*, *163*(163-195), 31–65.

Tran, T. P. (2017). Personalized ads on facebook: An effective marketing tool for online marketers. *Journal of Retailing and Consumer Services*, *39*, 230–242. https://doi.org/10.1016/j.jretconser.2017.06.010

Trewin, S., Swart, C., Koved, L., Martino, J., Singh, K., & Ben-David, S. (2012). Biometric authentication on a mobile device: A study of user effort, error and task disruption. *Proceedings of the 28th Annual Computer Security Applications Conference*, 159–168. https://doi.org/10.1145/2420950.2420976

Tripp, S. D., & Bichelmeyer, B. (1990). Rapid prototyping: An alternative instructional design strategy. *Educational technology research and development*, *38*(1), 31–44. https://doi.org/10.1007/BF02298246

Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, *35*(3), 405–426. https://doi.org/10.1287/mksc.2015.0956

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, *78*(5), B176–B185.

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*, 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

Uellenbeck, S., Dürmuth, M., Wolf, C., & Holz, T. (2013). Quantifying the security of graphical passwords: The case of android unlock patterns. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, 161–172. https://doi.org/10.1145/2508859.2516700

van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education*, *44*(1), 85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Veltman, J., & Gaillard, A. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*, 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. http://www.jstor.org/stable/30036540

Verma, G., Nalamada, T., Harpavat, K., Goel, P., Mishra, A., & Srinivasan, B. V. (2021). Non-linear consumption of videos using a sequence of personalized multimodal fragments. *26th International Conference on Intelligent User Interfaces*, 249–259. https://doi.org/10.1145/3397481.3450672

Vertegaal, R., Weevers, I., Sohn, C., & Cheung, C. (2003). Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 521–528. https://doi.org/10.1145/642611.642702

Vesterby, T., Voss, J. C., Hansen, J. P., Glenstrup, A. J., Hansen, D. W., & Rudolph, M. (2005). Gaze-guided viewing of interactive movies. *Digital Creativity*, *16*(4), 193–204. https://doi.org/10.1080/14626260500476523

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu koray, k., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion*, *20*(4), 625. https://doi.org/10.1037/emo0000578

W3C. (2003). W3c multimodal interaction framework. Retrieved January 26, 2023, from https://www.w3.org/TR/mmi-framework/

Wahlster, W. (2003). SmartKom: Symmetric multimodality in an adaptive and reusable dialogue shell. *Proceedings of the Human Computer Interaction Status Conference*, 47–62.

Wang, N., & Yeung, D.-Y. (2013). Learning a deep compact image representation for visual tracking. *26*, 809–817.

Wang, Q., Yang, J., Ren, M., & Zheng, Y. (2006). Driver fatigue detection: A survey. *2006 6th World Congress on Intelligent Control and Automation*, *2*, 8587–8591. https://doi.org/10.1109/WCICA.2006.1713656

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*(3). https://doi.org/10.1145/3386252

Want, R., Hopper, A., Falcão, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems*, *10*(1), 91–102. https://doi.org/10.1145/128756.128759

Wedel, M., & Pieters, R. (2006). Eye tracking for visual marketing. *Foundations and Trends in Marketing*, *1*(4). https://doi.org/10.1561/1700000011

Weiser, M. (1991). The computer for the 21 st century. *Scientific American*, *265*(3), 94–105. Retrieved December 15, 2022, from http://www.jstor.org/stable/24938718

Wertheimer, M. (1938). Gestalt theory. *A source book of Gestalt psychology*, 1–11. https://doi.org/10.1037/11496-001

Weyns, D., Iftikhar, M. U., de la Iglesia, D. G., & Ahmad, T. (2012). A survey of formal methods in self-adaptive systems. *Proceedings of the Fifth International C* Conference on Computer Science and Software Engineering*, 67–79. https://doi.org/10.1145/2347583.2347592

Whitehill, J., Bartlett, M., & Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–6. https://doi.org/10.1109/CVPRW.2008.4563182

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Wilding, M. (2016). Everyone's confused in meetings but no one asks questions - here's what needs to change. Retrieved August 11, 2022, from https://www.forbes.com/sites/melodywilding/2016/09/26/everyones-confused-in-meetings/?sh=6dab9ede567d

Williams, A. S., Garcia, J., & Ortega, F. (2020). Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, *26*(12), 3479–3489. https://doi.org/10.1109/TVCG.2020.3023566

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning and Technology*, *14*.

Xiao, B., Girand, C., & Oviatt, S. (2002). Multimodal integration patterns in children. *7th International Conference on Spoken Language Processing*, 629–632. https://www.isca-speech.org/archive_v0/icslp_2002/i02_0629.html

Xu, S., Jiang, H., & Lau, F. C. (2008). Personalized online document, image and video recommendation via commodity eye-tracking. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 83–90. https://doi.org/10.1145/1454008.1454023

Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, 121–130. https://doi.org/10.1145/2724660.2724677

Yasser, F. I., Abd, B. H., & Abbas, S. (2021). Detection of confusion behavior using a facial expression based on different classification algorithms. *Engineering and Technology Journal*, *39*(2A), 316–325. https://doi.org/10.30684/etj.v39i2A.1750

YouTube. (2023). Ad settings. Retrieved January 14, 2023, from https://www.youtube.com/intl/en_us/howyoutubeworks/user-settings/ad-settings/

Yusupov, R., & Ronzhin, A. (2010). From smart devices to smart space. *Herald of the Russian Academy of Sciences*, *80*(1), 63–68. https://doi.org/10.1134/S1019331610010089

Zagermann, J., Pfeil, U., & Reiterer, H. (2018). Studying eye movements as a basis for measuring cognitive load. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3170427.3188628

Zareian, G., Adel, S. M. R., & Noghani, F. A. (2015). The effect of multimodal presentation on efl learners' listening comprehension and self-efficacy. *Academic Research International*, *6*(1), 263.

Zhang, F., Xu, M., & Xu, C. (2022). Weakly-supervised facial expression recognition in the wild with noisy data. *IEEE Transactions on Multimedia*, *24*, 1800–1814. https://doi.org/10.1109/TMM.2021.3072786

Zhang, X., Sugano, Y., & Bulling, A. (2019). Evaluation of appearance-based methods and implications for gaze-based applications. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300646

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2017). It's written all over your face: Full-face appearance-based gaze estimation. *CVPR '17*, 51–60. https://doi.org/10.1109/CVPRW.2017.284

Zheng, Y., Ye, X., & Hsiao, J. H. (2022). Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learning and Instruction*, *77*, 101542. https://doi.org/10.1016/j.learninstruc.2021.101542

Zue, V., & Glass, J. (2000). Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, *88*(8), 1166–1180. https://doi.org/10.1109/5.880078

Appendix

A. Essay 2: Main and interaction effects on facial gestures

Two-way ANOVA tests with *Comprehension* and *Article Topic* or *Subtitle* as fixed factors, respectively, were performed for all facial gestures. Post-hoc pairwise comparisons were performed using paired t-tests with Šidák correction for all factors with significant group differences form the ANOVAs. Tables A.1-A.4 present the complete results.

TABLE A.1: ANOVA and post-hoc pairwise comparisons for blinkRate. Group differences were tested for the baseline, confusion, and reconciliation intervals. Significance was determined with two-way mixed ANOVAs using *Comprehension* (C) and *Subtitle* (S), and two-way repeated-measures ANOVAs with *Comprehension* and *Article Topic* (T) as fixed factors. For the interaction effects C*S and C*T, group differences are reported if they are significant within the specified interaction group. Pairwise t-tests with Šidák correction were used for post-hoc comparisons.

	ANOVA			**groups**								**pairwise t-test**		
Factor	F	sig.	η^2	interaction	A		B	[A] μ	(σ)	[B] μ	(σ)	t	sig.	η^2
C	5.285	.007	.107		baseline	–	confusion	.401	(.272)	.333	(.255)	2.952	.015	.016
					baseline	–	reconcile	.401	(.171)	.404	(.282)	0.121	.999	.000
					confusion	–	reconcile	.333	(.255)	.404	(.282)	2.753	.025	.017
S	0.220	.642	.004											
C * S	2.778	.068	.008	[captions]:	baseline	–	confusion	.411	(.296)	.288	(.227)	3.832	.005	.052
				[captions]:	baseline	–	reconcile	.411	(.296)	.387	(.309)	0.598	.993	.002
				[captions]:	confusion	–	reconcile	.288	(.227)	.387	(.309)	2.586	.097	.032
				[no caption]:	baseline	–	confusion	.390	(.251)	.380	(.279)	0.332	1.00	.000
				[no caption]:	baseline	–	reconcile	.390	(.251)	.422	(.256)	1.178	.825	.004
				[no caption]:	confusion	–	reconcile	.380	(.279)	.422	(.256)	1.221	.236	.006
T	0.040	.842	.000											
C * T	0.583	.534	.003											

TABLE A.2: ANOVA and post-hoc pairwise comparisons for Emotion. Group differences were tested for the baseline, confusion, and reconciliation intervals. Significance was determined with two-way mixed ANOVAs using *Comprehension* (C) and *Subtitle* (S), and two-way repeated-measures ANOVAs with *Comprehension* and *Article Topic* (T) as fixed factors. For the interaction effects C*S and C*T, group differences are reported if they are significant within the specified interaction group. Pairwise t-tests with Šidák correction were used for post-hoc comparisons.

		ANOVA			**groups**							**pairwise t-test**		
	Factor	F	sig.	η^2	interaction	A	B	[A] μ	(σ)	[B] μ	(σ)	t	sig.	η^2
positiveEmotion	C	0.673	.513	.015										
	S	0.228	.636	.005										
	C * S	0.475	.623	.000										
	T	**6.378**	**.015**	**.114**		funny	– sad	.107	(.231)	.033	(.147)	2.526	.015	.036
	C * T	0.376	.631	.001										
negativeEmotion	C	0.156	.813	.004										
	S	1.810	.186	.040										
	C * S	**4.090**	**.020**	**.002**	[captions]:	baseline	– confusion	.305	(.340)	.316	(.362)	0.751	.975	.000
					[captions]:	baseline	– reconcile	.305	(.340)	.347	(.362)	2.323	.166	.004
					[captions]:	confusion	– reconcile	.316	(.362)	.347	(.362)	2.342	.160	.002
					[no caption]:	baseline	– confusion	.209	(.294)	.196	(.275)	0.936	.932	.001
					[no caption]:	baseline	– reconcile	.209	(.294)	.177	(.266)	1.282	.765	.003
					[no caption]:	confusion	– reconcile	.196	(.275)	.177	(.266)	0.779	.971	.001
	T	**4.307**	**.044**	**.072**		funny	– sad	.220	(.324)	.302	(.366)	2.075	.044	.014
	C * T	0.586	.493	.001										
neutralEmotion	C	0.176	.813	.004										
	S	0.072	.790	.002										
	C * S	**5.559**	**.005**	**.002**	[captions]:	baseline	– confusion	.581	(.359)	.570	(.377)	0.741	.977	.000
					[captions]:	**baseline**	**– reconcile**	**.581**	**(.359)**	**.540**	**(.363)**	**2.641**	**.086**	**.003**
					[captions]:	confusion	– reconcile	.570	(.377)	.540	(.363)	2.263	.187	.002
					[no caption]:	baseline	– confusion	.509	(.395)	.535	(.399)	1.620	.538	.001
					[no caption]:	baseline	– reconcile	.509	(.395)	.554	(.413)	1.853	.388	.003
					[no caption]:	confusion	– reconcile	.535	(.399)	.554	(.413)	0.768	.973	.001
	T	0.234	.631	.004										
	C * T	0.755	.436	.002										
emotionFluctuation	C	.628	.514	.014										
	S	1.088	.303	.023										
	C * S	0.057	.945	.000										
	T	0.004	.947	.000										
	C * T	0.934	.378	.004										

TABLE A.3: ANOVA results on activation intensity of each action unit. Activation intensities are calculated as the average intensity across all frames as identified by `OpenFace`. Fixed factors were *Comprehension* (C) in combination with *Article Topic* (T) and *Subtitle* (S), respectively. P-values for groups with unequal variance are Greenhouse-Geisser corrected.

	C			**S**			**C * S**			**T**			**C * T**		
AU	F	sig.	η^2	F	sig.	η^2	F	sig.	η^2	F	sig.	η^2	F	sig.	η^2
1	0.159	.739	.004	0.031	.864	.001	0.516	.599	.003	0.012	.914	.000	0.048	.953	.000
2	0.703	.498	.016	0.043	.836	.001	2.031	.137	.019	1.788	.188	.013	1.409	.250	.008
4	3.083	.066	.065	0.187	.667	.004	1.590	.210	.001	0.853	.361	.017	1.325	.271	.002
5	1.209	.303	.027	0.291	.593	.004	0.074	.929	.001	2.466	.123	.014	0.664	.520	.005
6	3.041	.058	.065	0.514	.477	.012	0.710	.494	.000	1.951	.169	.037	1.510	.227	.002
7	3.830	.033	.080	0.154	.697	.004	1.014	.367	.000	0.207	.651	.004	0.427	.654	.001
9	3.397	.038	.071	0.791	.379	.014	0.136	.873	.001	0.041	.840	.000	1.677	.193	.012
10	3.914	.034	.784	0.260	.613	.006	0.105	.900	.000	0.586	.448	.012	0.716	.491	.001
12	1.627	.208	.036	1.196	.280	.026	0.353	.704	.000	8.696	.005	.150	0.428	.653	.000
14	0.155	.827	.003	0.349	.558	.008	0.746	.477	.000	9.065	.004	.155	0.052	.949	.000
15	0.878	.419	.020	3.302	.076	.060	1.621	.204	.006	0.000	.983	.000	0.363	.697	.001
17	5.094	.020	.104	1.586	.215	.034	1.571	.214	.002	0.103	.750	.001	3.370	.039	.017
20	1.950	.148	.042	0.011	.916	.000	0.671	.514	.004	2.092	.155	.027	0.334	.717	.002
23	6.983	.002	.137	1.994	.165	.035	0.956	.388	.004	0.154	.696	.001	2.959	.057	.033
25	0.792	.411	.018	3.105	.085	.058	0.577	.564	.002	3.441	.070	.050	0.905	.408	.002
26	1.090	.317	.024	0.298	.588	.006	0.056	.946	.000	0.358	.552	.005	1.090	.341	.003
28	0.912	.405	.020	1.456	.234	.016	0.351	.705	.004	1.767	.191	.013	0.289	.749	.002
45	2.796	.067	.060	0.500	.483	.008	0.049	.952	.000	0.004	.950	.000	2.115	.127	.007

TABLE A.4: Post-hoc pairwise comparisons for action units. Pairwise t-tests with Šidák correction were used for significance testing. Group differences in *Comprehension* (C) were tested for the baseline, confusion, and reconciliation intervals. Additionally, the effect of *Article Topic* (T) was assessed on the funny and sad article. The effect of *Subtitle* (S) was tested for sessions with either auto-generated captions, or no caption. For the interaction effects C*S and C*T, group differences are reported if they are significant within the specified interaction group.

			groups						pairwise t-test		
AU	**Factor**	interaction	A	B	[A] μ	(σ)	[B] μ	(σ)	t	sig.	η^2
4	C		baseline	– confusion	.519	(.520)	.533	(.531)	1.200	.555	.000
			baseline	– reconcile	.519	(.520)	.563	(.542)	2.195	.097	.002
			confusion	– reconcile	.533	(.531)	.563	(.542)	1.422	.412	.001
6	C		baseline	– confusion	.246	(.336)	.263	(.344)	1.246	.524	.001
			baseline	– reconcile	.246	(.336)	.278	(.346)	2.201	.096	.002
			confusion	– reconcile	.263	(.344)	.278	(.346)	1.425	.410	.001
7	C		baseline	– confusion	.628	(.696)	.639	(.681)	0.696	.868	.000
			baseline	– reconcile	.628	(.696)	.684	(.712)	2.354	.068	.002
			confusion	– reconcile	.636	(.681)	.684	(.712)	1.909	.177	.001
9	C		baseline	– confusion	.042	(.057)	.045	(.057)	1.046	.659	.001
			baseline	– reconcile	.042	(.057)	.057	(.039)	2.075	.126	.025
			confusion	– reconcile	.045	(.057)	.057	(.039)	1.689	.267	.015
10	C		baseline	– confusion	.298	(.396)	.312	(.391)	1.528	.350	.000
			baseline	– reconcile	.298	(.396)	.333	(.388)	2.327	.072	.002
			confusion	– reconcile	.312	(.391)	.333	(.388)	1.632	.295	.001
12	T		funny	– sad	.420	(.512)	.227	(.319)	2.949	.005	.059
14	T		funny	– sad	.597	(.666)	.386	(.533)	3.011	.004	.030
15	S		funny	– sad	.151	(.164)	.085	(.051)	1.851	.075	.068
17	C		baseline	– confusion	.291	(.329)	.333	(.357)	3.899	.001	.004
			baseline	– reconcile	.291	(.329)	.356	(.439)	2.590	.038	.007
			confusion	– reconcile	.333	(.357)	.356	(.439)	1.019	.677	.001
	C * T	[sad]:	baseline	– confusion	.283	(.361)	.344	(.459)	2.799	.045	.006
			baseline	– reconcile	.274	(.339)	.328	(.440)	2.281	.154	.005
			confusion	– reconcile	.274	(.339)	.399	(.589)	2.510	.091	.016
		[funny]:	baseline	– confusion	.328	(.440)	.399	(.589)	1.648	.491	.005
			baseline	– reconcile	.307	(.375)	.337	(.356)	1.629	.505	.002
			confusion	– reconcile	.307	(.375)	.314	(.336)	0.290	1.00	.000
23	**C**		baseline	– confusion	.063	(.074)	.096	(.121)	3.335	.005	.026
			baseline	– reconcile	.063	(.074)	.118	(.184)	2.945	.015	.037
			confusion	– reconcile	.096	(.121)	.118	(.184)	1.515	.357	.005
	C * T	[sad]:	baseline	– confusion	.049	(.059)	.086	(.152)	1.878	.341	.025
			baseline	– reconcile	.049	(.059)	.086	(.152)	2.308	.145	.050
			confusion	– reconcile	.086	(.152)	.151	(.307)	1.866	.348	.018
	C * T	[funny]:	baseline	– confusion	.077	(.123)	.106	(.144)	2.758	.050	.012
			baseline	– reconcile	.077	(.123)	.085	(.093)	0.643	.988	.001
			confusion	– reconcile	.106	(.144)	.085	(.093)	1.401	.669	.007
25	S		funny	– sad	.241	(.198)	.159	(.098)	1.787	.083	.065
	T		funny	– sad	.233	(.245)	.168	(.140)	1.855	.070	.026
45	C		baseline	– confusion	.220	(.119)	.197	(.116)	1.679	.272	.011
			baseline	– reconcile	.222	(.232)	.232	(.124)	0.740	.463	.002
			confusion	– reconcile	.197	(.116)	.232	(.124)	2.083	.124	.022

B. Essay 3: Gaze feature extraction

B.1. Fixation calculation

The calculation of fixations is based on the algorithms employed by the eye tracker manufacturer Tobii (Olsen, 2012). Applying a strict average methodology, gaze points are set to the mean value of the left and right eye. In case only one eye has been detected by the eye tracker, both values are discarded. All thresholds are set to the Tobii Studio default values. These have been found to produce the most robust results in experiments with a similar setup to the one used in the study that is presented in Essay 3 (Olsen, 2012). The pseudocode in Algorithm A.1 describes a simplified version of the implementation for extracting fixations.

ALGORITHM A.1: Fixation extraction.

```
Input: gazeList
Output: fixationList

/**
Group consecutive samples based on their velocity
*/
for all g in gazeList do
  /** Calculate visual angle per second between two gaze points (Note 1) */
  call visualAngle(eye, g.previous, g.next)
  velocity ← visualAngle / (g.next.time - g.previous.time)

  /** Classify samples as gap, fixation, or saccade */
  if g.validity = false then
    gazeType ← gap
  else if velocity < 30 then
    gazeType ← fixation
  else
    gazeType ← saccade
  end if

  /** Check if the sample has the same type as the previous */
  if g.previous.gazeType = gazeType then
    /** Put into a temporary clusterList */
    append gaze to clusterList
  /** If type changes, create a new fixation, saccade or gap object */
  else
    init Feature
    Feature.type ← gazeType
    Feature.coordinates ← call mean(clusterlist)
    Feature.firstSample ← clusterList[0]  /** first cluster sample */
    Feature.lastSample ← clusterList[-1]  /** last cluster sample */
    Feature.start ← clusterList[0].time   /** start with first cluster sample */
    Feature.stop ← clusterList[-1].time   /** stop with last cluster sample */

```

```
35      /** If the feature is a fixation, put into fixationList */
36      if Feature.type = fixation then
37        append fixation to fixationList
38      end if
39
40      /** Clear clusterList and initialize with next sample from gazeList */
41      clusterList ← g
42    end if
43  end for
44
45  /**
46  Merge adjacent fixations that were split due to noise
47  */
48  for all f in fixationList do
49    /** Check if the gap between two adjacent fixations is less than 75 ms */
50    if f.next.start - f.stop < 75 then
51      g1 ← f.lastSample
52      g2 ← f.next.firstSample
53      call visualAngle(mean(g1.eye, g2.eye), g1.coordinates, g2.coordinates)
54      /** Merge fixations with visual angle below .5 degrees/second */
55      if visualAngle < 0.5 then
56        f.coordinates ← call mean(f.next.coordinates, f.coordinates)
57        f.stop ← f.next.stop
58        /** Delete appended fixation from fixationList */
59        call delete(f.next)
60      end if
61    end if
62  end for
63
64  /**
65  Discard short fixations below 60 ms (not relevant for information absorption)
66  */
67  for all f in fixationList do
68    if f.duration < 60 then
69      call delete(f)
70    end if
71  end for
```

Note 1: The recommended window for velocity calculation is 20 milliseconds. The Tobii Pro X3-120 eye tracker has a sampling rate of 120 Hz (i.e., samples are recorded in intervals of 8.33 milliseconds). To approximate the recommended window size, the visual angle is calculated for the movement $M_{g1,g2}$ between two gaze samples $g1$ at time $t-1$ and $g2$ at $t+1$ with respect to their tangential distance $D_{e,\overline{g1g2}}$ from the eye e at time t:

$$2 * atan(\frac{M_{g1,g2}}{2 * D_{e,\overline{g1g2}}}) \tag{1}$$

Figure B.1 shows the spacial relations between the coordinates of the eye and the two gaze samples that are used for the calculation.

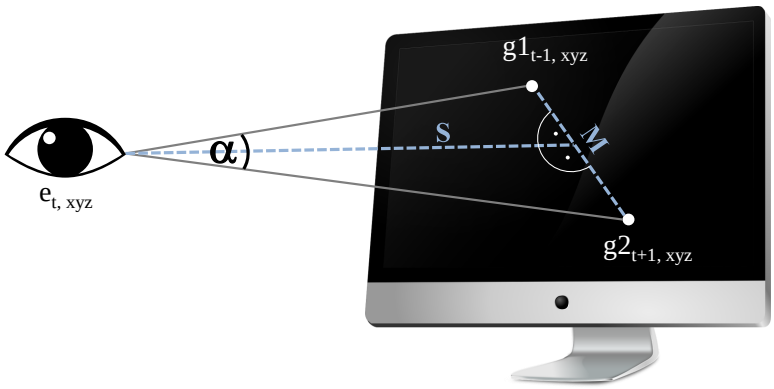


FIGURE B.1: Visual angle calculation. The visual angle is calculated from the movement M of the gaze between two samples $g_{t,xyz}$ from time $t-1$ and $t+1$ with respect to the distance S of the eye e_{xyz} to the screen at time t.

B.2. Feature definition

Using raw gaze data (A.2 - A.5), or the extracted fixations (A.6 - A.10) as input, the gaze feature extractor produces eight duration, frequency, and sequential features. The procedure for their calculation is summarized in Algorithms A.2 - A.10.

ALGORITHM A.2: Calculate gazeCount.

```
1  Input: gazeList, aoiList
2  Output: endGazeList
3
4  /** Initialize a list with a separate fixation counter for each AOI */
5  for all a in aoiList do
6    init counter
7    append counter to endGazeList
8  end for
9
10 for all g in gazeList do
11   /** If gaze is in AOI, increment its counter */
12   for all a in aoiList do
13     inAoi ← call checkCoordinates(a, g)
14     if inAoi = true then
15       call increment(endGazeList[a])
16     end if
17   end for
18 end for
```

ALGORITHM A.3: Calculate endCount.

```
Input: gazeList, aoiList
Output: endGazeList

/** Filter gaze points from last 400 ms */
gazeList → gazeList[-400:]

/** Initialize a list with a separate fixation counter for each AOI */
for all a in aoiList do
  init counter
  append counter to endGazeList
end for

for all g in gazeList do
  /** If gaze is in AOI, increment its counter */
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, g)
    if inAoi = true then
      call increment(endGazeList[a])
    end if
  end for
end for
```

ALGORITHM A.4: Calculate maxDwell.

```
Input: gazeList, aoiList
Output: maxDwellList

/** Initialize a list with the current longest dwell on each AOI */
for all a in aoiList do
  init maximum
  append maximum to maxDwellList
end for

/** Initialize a list with a separate fixation counter for each AOI */
init Focus: aoi, start

for all g in gazeList do
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, g)
    if inAoi = true then
      /** If a new AOI is fixated, calculate dwell on the last AOI */
      if Focus.aoi not a then
        dwell ← g.time - Focus.start
        /** If dwell is longer than current maximum, update value for AOI */
        if dwell > maxDwellList[a] then
          maxDwellList[a] ← dwell
        end if
        /** Update start of next dwell with time of current gaze sample */
        Focus.start ← g.time
        Focus.aoi ← a
      end if
    end if
  end for
end for
```

ALGORITHM A.5: Calculate meanDwell.

```
Input: gazeList, aoiList
Output: meanDwellList

/** Initialize list of dwell objects with number + total duration for each AOI */
for all a in aoiList do
  init Dwell: count, duration
  append Dwell to dwellList
end for

/** Initialize a list with a separate fixation counter for each AOI */
init Focus: aoi, start

for all g in gazeList do
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, g)
    if inAoi = true then
      /** If new AOI is fixated, update dwell count and duration on last AOI */
      if Focus.aoi not a then
        dwell ← g.time - Focus.start
        add dwell to dwellList[a].duration
        call increment(dwellList[a].count)

        /** Update start of next dwell with time of current gaze sample */
        Focus.start ← g.time
        Focus.aoi ← a
      end if
    end if
  end for
end for

for all a in aoiList do
  meanDwell ← dwellList[a].duration / dwellList[a].count
  append meanDwell to meanDwellList
end for
```

ALGORITHM A.6: Calculate fixCount.

```
Input: fixationList, aoiList
Output: fixCountList

/** Initialize a list with a separate fixation counter for each AOI */
for all a in aoiList do
  init counter
  append counter to fixCountList
end for

for all f in fixationList do
  /** If fixation is in AOI, increment its counter */
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, f)
    if inAoi = true then
      call increment(fixCountList[a])
    end if
  end for
end for
```

ALGORITHM A.7: Calculate meanFixDur.

```
Input: fixationList, aoiList, fixCountList
Output: meanDurationList

/** Initialize a list with a separate duration counter for each AOI */
for all a in aoiList do
  init duration
  append duration to meanDurationList
end for

for all f in fixationList do
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, f)
    /** If fixation is in AOI, update its total duration */
    if inAoi = true then
      add duration to meanDurationList[a]
    end if
  end for
end for

/** Divide the total duration of each AOI by its fixation number */
for all d in meanDurationList do
  d ← d / fixCountList[a]
end for
```

ALGORITHM A.8: Calculate maxFix.

```
Input: fixationList, aoiList
Output: maxFixList

/** Initialize list with the current longest fixation on each AOI */
for all a in aoiList do
  init maximum
  append maximum to maxFixList
end for

for all f in fixationList do
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, f)
    if inAoi = true then
      /** If fixation is longer than current maximum, update the value */
      if f.duration > maxFixList[a] then
        maxFixList[a] ← f.duration
      end if
    end if
  end for
end for
```

ALGORITHM A.9: Calculate longFix.

```
  Input: fixationList, aoiList
  Output: longFixList

  /** Initialize a list with a separate fixation counter for each AOI */
  for all a in aoiList do
    init counter
    init counter in longFixList
  end for

  for all f in fixationList do
    /** If fixation is longer 400 ms, increment the counter of its AOI */
    if f > 400 do
      for all a in aoiList do
        inAoi ← call checkCoordinates(a, f)
        if inAoi = true then
          call increment(fixCountList[a])
        end if
      end for
    end if
  end for
```

ALGORITHM A.10: Calculate maxContFix.

```
Input: fixationList, aoiList
Output: contFixList

/** Initialize a list with the current longest dwell on each AOI */
for all a in aoiList do
  init maximum
  append maximum to contFixList
end for

/** Initialize a list with a separate fixation counter for each AOI */
init Focus: aoi, count

for all f in fixationList do
  for all a in aoiList do
    inAoi ← call checkCoordinates(a, f)
    if inAoi = true then
      if Focus.aoi = a then
        call increment(f.count)
      /** If a new AOI is fixated, calculate number of fixations on the last AOI */
      else
        count ← f.start - Focus.start
        /** If count is longer than current maximum, update value for AOI */
        if count > contFixList[a] then
          contFixList[a] ← count
        end if

      /** Reset counter for next AOI dwell */
      Focus.count ← 0
      Focus.aoi ← a
      end if
    end if
  end for
end for
```

C. Publications contained in the thesis

Heck, M., Edinger, J., & Becker, C. (2019). Gaze-based product filtering: A system for creating adaptive user interfaces to personalize stateless point-of-sale machines. *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 75–77.

» *Categorization of a set of personalization techniques for visual user interfaces. The content filtering technique in Essay 3 was chosen from the set with regard to the use case constraints (i.e., same-element content adaptation).*

Heck, M., Edinger, J., Bünemann, J., & Becker, C. (2021). Exploring gaze-based prediction strategies for preference detection in dynamic interface elements. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 129–139.

» *Experimental evaluation to identify robust gaze features and machine learning classifiers for preference predictions. The results were used to narrow down the feature set and classifiers that were tested in Essay 3.*

Heck, M., Edinger, J., Bünemann, J., & Becker, C. (2021). The subconscious director: Dynamically personalizing videos using gaze data. ***26th International Conference on Intelligent User Interfaces*****, 98-108.**

» ***Essay 3***

Heck, M., Sonntag, P., & Becker, C. (2021). Is this really relevant? A guide to best practice gaze-based relevance prediction research. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 220-228.

» *Literature review on content filtering based on gaze-informed user preferences. The insights from the article were used to define the preference prediction strategies in Essay 3.*

Heck, M., Edinger, J., & Becker, C. (2022). Lessons learned from an eye tracking study for targeted advertising in the wild. *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*, 539-544.

» *Pilot study testing the effectiveness of using gaze-informed preferences to personalize advertisements. The insights from the field study were used to optimize the experimental design of the study that is presented in Essay 3.*

Heck, M., Shon, S.H., & Becker, C. (2022). Does using voice authentication in multimodal systems correlate with increased speech interaction during non-critical routine tasks? ***27th International Conference on Intelligent User Interfaces*****, 868-877.**

» ***Essay 1***

Short CV

Since 04/2022	**Researcher** Institute for Parallel and Distributed Systems University of Stuttgart *Supervisor: Prof. Dr. Christian Becker*
03/2019 – 03/2022	**Researcher** Information Systems II: Distributed Systems University of Mannheim *Supervisor: Prof. Dr. Christian Becker*
09/2016 – 02/2019	**Master of Science Mannheim Master in Management** University of Mannheim
10/2012 – 09/2015	**Bachelor of Arts International Management** University of Flensburg