Does Using Voice Authentication in Multimodal Systems Correlate With Increased Speech Interaction During Non-critical Routine Tasks?

Melanie Heck
University of Mannheim
Mannheim, Germany
melanie.heck@uni-mannheim.de

Seong Hyun Shon
University of Mannheim
Mannheim, Germany
sshon@mail.uni-mannheim.de

Christian Becker
University of Mannheim
Mannheim, Germany
christian.becker@uni-mannheim.de

ABSTRACT

Multimodal systems offer their functionalities through multiple communication channels. A messenger application may take either keyboard or voice input, and present incoming messages as text or audio output. This allows the users to communicate with their devices using the modality that best suits their context and personal preference. Authentication is often the first interaction with an application. The users' login behavior can thus be used to immediately adapt the communication channel to their preferences. Yet given the sensitive nature of authentication, this interaction may not be representative for the user's inclination to use speech input in non-critical routine tasks. In this paper, we test whether the interactions during authentication differ from non-critical routine tasks in a smart home application. Our findings indicate that, even in such a private space, the authentication behavior does not correlate with the use, nor with the perceived usability of speech input during non-critical task. We further find that short interactions with the system are not indicative of the user's attitude towards audio output, independent of whether authentication or non-critical tasks are performed. Since security concerns are minmized in the secure environment of private spaces, our findings can be generalized to other contexts where security threats are even more apparent.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; Contextual design.

KEYWORDS

multimodal interfaces, modality arbitration, smart home application

ACM Reference Format:
Melanie Heck, Seong Hyun Shon, and Christian Becker. 2022. Does Using Voice Authentication in Multimodal Systems Correlate With Increased Speech Interaction During Non-critical Routine Tasks?. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3490099.3511129

IUI '22, March 22–25, 2022, Helsinki, Finland
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9144-3/22/03.
https://doi.org/10.1145/3490099.3511129

1 INTRODUCTION

The widespread availability of voice control with software components such as Apple's Siri [4], Microsoft's Cortana [31], or Amazon's Alexa [3] have introduced another convenient interaction mechanism, complementing the traditional touch input on personal devices. With the redundant implementation of all available functionalities in each modality [34], users can easily switch between multiple interaction options. Multimodal interaction thus allows users to interact with their devices in a modality that suits the immediate context. Given the dynamics of most people's busy lifestyles, this context may sometimes change almost instantly [35]. Yet despite a long-standing history of research on multimodal systems [16, 21], their design is still a challenging task, as each user's interaction with a multimodal system is different [37]. While it is possible to activate multiple communication channels simultaneously, this may not be desirable. For instance, audio output can be disturbing in public places, or even pose a threat to security when sensitive information is read aloud by the system [14]. In an attempt to overcome these issues, adaptive multimodal systems use cues from the user's chosen input modality to infer the most appropriate output channel in a given situation [11, 28].

In most applications, the user's first point of interaction is authentication. However, given its sensitivity to security issues, the users may log in using a different modality than they prefer for non-critical routine tasks. Most people have reservations about speaking a passphrase out of fear of being overheard [49]. We therefore investigate whether the users' inputs for authentication differ from their interaction preferences during non-critical tasks. We study this question in the context of a smart home application that can be controlled using touch our 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 during the remainder of the interaction, and whether users who use voice authentication find speech input more usable. We further test whether it is beneficial to the user if the system output is displayed in the same sensory modality. Thus, depending on the input, system instructions and responses are presented as written text or audio output.

2 RELATED WORK

Previous work that attempts to understand user behavior when interacting with *multimodal systems* has mainly focused studying under which contexts users prefer multimodal interaction over unimodal input. A consensus exists that the usability of multimodal interaction is primarily driven by the activity [32, 33, 36, 54]. However, the views on how task complexity relates to the suitability of multimodal interaction are conflicting. Oviatt et al. [33, 36] found that users prefer unimodal touch or speech input when executing simple tasks, but revert to multimodal input as the tasks become more cognitively complex. Consistent with these findings, Morris and Ringel [32] showed that unimodal interaction is preferred for issuing simple commands to a speech and gesture controlled TV. When interacting with the considerably more complex image editing application PixelTone [25], in contrast, the users preferred multimodal interaction over using only speech or touch input. Conversely, elderly users seem to prefer multimodal input for controlling a simple TV application [12], and a recent study by Williams et al. [54] suggests that the combined use of speech and gesture input is more strenuous during complex tasks. Apart from external task related variables, user specific variables influence the usability of multimodal interaction [8, 37]. The more experienced the users are, the more often they use multiple input modalities [8]. Once a user has developed a preference for unimodal or multimodal interaction, their interaction behavior remains consistent [37]. When asked explicitly to use both speech and touch 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.

When given the choice between multiple *interaction modalities*, manual input has been found to be preferred over speech in Human-Robot [40] and Human-Computer Interaction, including a computer-based drawing application [2] and an application to document electronic health records in hospitals [46]. 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 [7]. In a study with a smart television [20], the traditional remote control was preferred over speech interaction. Gestures were the modality of choice for selecting virtual objects in an AR application [26]. While this may be attributed to the higher cognitive load of speech input [44], speech can be preferred if the benefits in terms of efficiency and effectiveness are sufficiently high. Schaffer et al. [45] report that users of a restaurant booking application for mobile phones preferred speech input 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 [47] 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 opted for manual input for simple location specifications, but preferred speech for more extensive and not clearly defined object descriptions [33]. Similarly, user interaction with the meeting documentation system Archivus [29] suggest that the mouse is used for simple navigation tasks, whereas speech input is preferred for free text entry. Since efficiency and effectiveness depend on individual factors like demographics and prior knowledge, modality preferences can differ from person to person [23]. Dynamic context variables such as the user's environment and cognitive load influence the effectiveness of an interaction modality [32]. Ideally, multimodal systems should therefore adapt to the individual user and current situation.

Similarly, preferences for *output modalities* depend on the interaction context [12]. Adaptive multimodal systems thus dynamically infer the user's context from multimodal input, and present the output in the modality that is most appropriate for the situation [11, 28]. While increasing usability for inexperienced users, output adaptation has not been found to benefit domain experts [28].

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

3 THE SMART HOME APPLICATION

The desktop application "Smart Home Display" is a simple prototype for an adaptive smart home application. Telework has become the norm during the Covid-19 pandemic, and about 75% of office workers now wish to at least partly work from home [38]. Assuming that the users prefer to control their home appliances through the same device that they are already using during their office hours, we decided to conduct the study with a desktop application. The application can be operated using touch or speech input. Adaptive functionalities were integrated following the FAME [15] development guidelines for adaptive multimodal applications:

(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 Turk's taxonomy of sensory modalities [50], 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 [19]. However, speech-controlled intelligent personal assistants such as Alexa or Siri have penetrated the consumer market, so that most users are now comfortable with using speech input. In a study investigating the usability of different input modalities for smart home appliances, users were found to be most experienced with touch and speech input, and find these modalities the easiest and most enjoyable to use [19]. The system adapts exclusively to the input modality that was chosen by the user for authentication. Environmental variables such as background noise do not influence the adaptation.

(2) **Identify adaptable variables**: *What variable system components should respond to the 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 stores whether speech or touch input was used during the login task. The system continuously listens for potential speech input throughout the entire interaction. However, the user model is static, i.e., it is not updated if the user subsequently communicates with the system using a different input 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 graphical user interface (GUI) of the chosen input modality during authentication.
(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` [17] which provides convenient functionalities for rapid prototyping. TTS conversion was realized with Window's native `Microsoft Speech API (SAPI 5.3)` [30]. 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 [24]. The speech recognition engine runs in the background and continuously listens for speech input.

3.1 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 authentication page. Authentication is a critical security mechanism for Smart Homes in order to protect services such as paid TV channels from unauthorized access [39]. 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.

Two alternative *authentication tasks* that are commonly used on consumer devices were implemented. Both tasks use a secret (i.e., user-specific knowledge) for authentication, independent of the chosen input modality. Note that this text-dependent authentication method differs from biometric voice authentication, where the user's identity is verified based on a set of vocal parameters [42]. We chose this approach so that authentication would be based on the same type of credential, independent of whether touch or speech was used. In the study, one authentication was chosen by the system at random. This allowed us to investigate whether the input mechanism that was chosen for authentication was biased by the authentication task. *Login 1* uses identical tasks for touch and speech input. The user has to enter the PIN "5678", using either the number pad displayed on the screen, or saying the number sequence out loud. *Login 2* uses a different task for each input modality. If choosing touch, the user is requested to draw a simple pattern onto the display. This authentication procedure is widely used on Android smartphones [51], and therefore provides both familiarity and high usability. Alternatively, the user can speak the sentence "I wish to enter". In both tasks, the cognitive effort if using touch is comparable to speech input (cf. Section 5.1). 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 methods (*Login 1* vs. *Login 2*), we account for the legacy bias that might be associated with the authentication tasks. 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 input for PIN authentication on the smart home application, even though in the safety of their own home, speech input does not reveal the secret PIN to unauthorized individuals.

Since the speech recognition rate for numerical digits is comparatively low (about 75%) [5], we anticipated performance issues during speech-based PIN authentication. In order to not discourage participants from using speech authentication 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 provided spoken input during the login was correct. In contrast, spoken input after the authentication needed to be recognized correctly in order for the system to respond.

3.2 Smart home functionalities

In our user study, we aimed to obtain a balanced sample of participants using either of the available input mechanisms for the noncritical 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 input modality. The chosen form of interaction should be entirely the result of a personal preference. We therefore selected a scenario in which both modalities would naturally be equally convenient to use. Given that the users are already seated in front of their computer in the orchestrated telework scenario, the mouse lies within easy reach. Speech interaction is typically only preferred under specific circumstances [12, 29, 32, 33]. It has been found to be especially useful when executing simple tasks that require only minimal input [1, 27]. The system thus provides control panels for three smart home functionalities 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 using touch or speech input. Speech commands correspond to the button labels. Instructions and system responses to user actions are provided in audio or text format, depending on the modality that was used for authentication.

4 USER STUDY

The objective of the user study was to test whether users who favor voice over touch authentication in a multimodal system also find speech more useful during non-critical routine tasks:

> **H1:** 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 modality selection criterion in authentication, even more so than confidence and privacy [48]. Since this is also the case for non-critical tasks [13], **H1** assumes that the same modality is preferred for authentication and non-critical tasks.

Modality switches induce additional cognitive load [13, 43]. Buisine et al. [9] 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 [53], 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:

> **H2:** Users prefer system output in the same modality they use for input commands.

To test our hypotheses, a between-subjects experimental design with one experimental group and one control group was adopted. Participants could authenticate themselves using either touch or speech. In the experimental group, system output was presented in the modality that is most logically associated with the sensory input from the login task. Participants who authenticated themselves using 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 provided with audio output exclusively. In the control group, in contrast, all system outputs were presented in the modality that was not consistent with the chosen authentication modality.

4.1 Apparatus

Participants were seated in front of an HP laptop with a 1.6GHz i5-8250U processor and 16GB RAM. The GUI was projected onto the laptop's display (1920x1080 pixels). It was ensured that a stable wifi connection persisted throughout each experimental trial in order 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 application. To compensate for muffled voice input, a Jabra Evolve 40 headset was connected to the laptop.

4.2 Participants

We recruited 41 campus residents and their personal contacts (23 female, mean age = 26.1 ± 3.4) for the experiment. 11 participants were enrolled in an undergraduate program at our university. 27 were currently pursuing or had already completed a graduate degree. Participants were from a total of 14 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.

4.3 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 application. They were told that their mouse and speech inputs were being recorded. After the trial, they received a detailed explanation about the objective of the study.

Before using the system, the participants were given the login data. 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.

4.4 Metrics

The **perceived workload of authentication** was used as a control variable to test whether the modality choice was biased by the intrinsic cognitive load of the task itself. It was measured with items from the NASA Task Load Index [18] on a 7-point Likert scale:

- **A1:** How mentally demanding was the login?
- **A2:** How physically demanding was the login?
- **A3:** How hurried or rushed was the pace of the login?
- **A4:** How successful were you in accomplishing the login?
- **A5:** How hard did you have to work (mentally and physically) to accomplish the login?
- **A6:** How insecure, discouraged, irritated, stressed and annoyed did you feel during the login?

Usability of speech input was used as a subjective measure of the users' attitude towards speech interaction. It was evaluated on a 7-point Likert scale with *effort* and *performance expectancy* measures adapted from the UTAUT [52]:

- **U1:** Using speech interaction enables me to accomplish tasks more quickly than with traditional mouse interaction.
- **U2:** My interaction with the smart home application is clear and understandable.
- **U3:** Using speech gives me greater control over the system than using mouse-based inputs.
- **U4:** Using speech makes it easier to interact with the smart home application than using the mouse.
- **U5:** Using speech to interact with the system is cumbersome.
- **U6:** The smart home application responded to my speech input in a timely manner.
- **U7:** I wished that the system better recognized my speech.

The **interaction behavior** was used as an objective measure for their 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. We recorded all predicted speech alternatives and their confidence values. Confidence scores (ranging from 0 to 1) indicate how reliable the STT conversion is [22]. From the raw log data, we defined three metrics related to the user's *modality choice*, and three additional indicators for the *speech recognition quality*. The metrics and their calculations are summarized in Table 1.

Subjective preferences for output modalities were measured with six UTAUT items. The items were adapted to assess the **usability of text and audio output** of the smart home application:

Construct	Metric	Definition
modality choice	`click_count`	Total number of clicks during the entire interaction
	`speech_count`	Total number of recognized speech inputs from the entire interaction
	`speech_ratio`	Ratio of speech input to total input: $\frac{\texttt{speech\_count}}{\texttt{speech\_count} + \texttt{click\_count}}$
speech recognition quality	`avg_speech_confidence`	Mean speech recognition confidence for a user, calculated from the confidence value for the final speech alternative
	`executable_command_count`	Number of speech inputs that result in the successful execution of the associated command (excluding input where some speech was recognized, but could not be associated with a command, either because a wrong keyword was used, or the words were not recognized correctly)
	`executable_command_ratio`	Ratio of executable to total speech input: $\frac{\texttt{executable\_command\_count}}{\texttt{speech\_count}}$

Table 1: Metrics used to analyze the log data.

- **O1:** I find it useful to receive spoken instructions.
- **O2:** Receiving spoken instruction enables me to accomplish tasks more quickly than with text instructions.
- **O3:** Spoken instructions make using the system more interesting.
- **O4:** It scares me to think that I could miss important information if the instructions were only displayed as spoken messages.
- **O5:** I would have wished to receive more spoken audio instructions.
- **O6:** I would have wished to receive more text instructions.

5 RESULTS

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 authenticated themselves using speech, and 24 subjects opted for touch input (cf. Table 2).

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

Table 2: Input modality usage per authentication task

5.1 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 differently than those using touch. A two-sided t-test showed that cognitive workload does not significantly differ between the two input modalities (p-value = .351). Looking at each authentication task individually, we did not find 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' choice of input modality.

5.2 H1: Voice authenticators are more inclined to use speech input in non-critical tasks

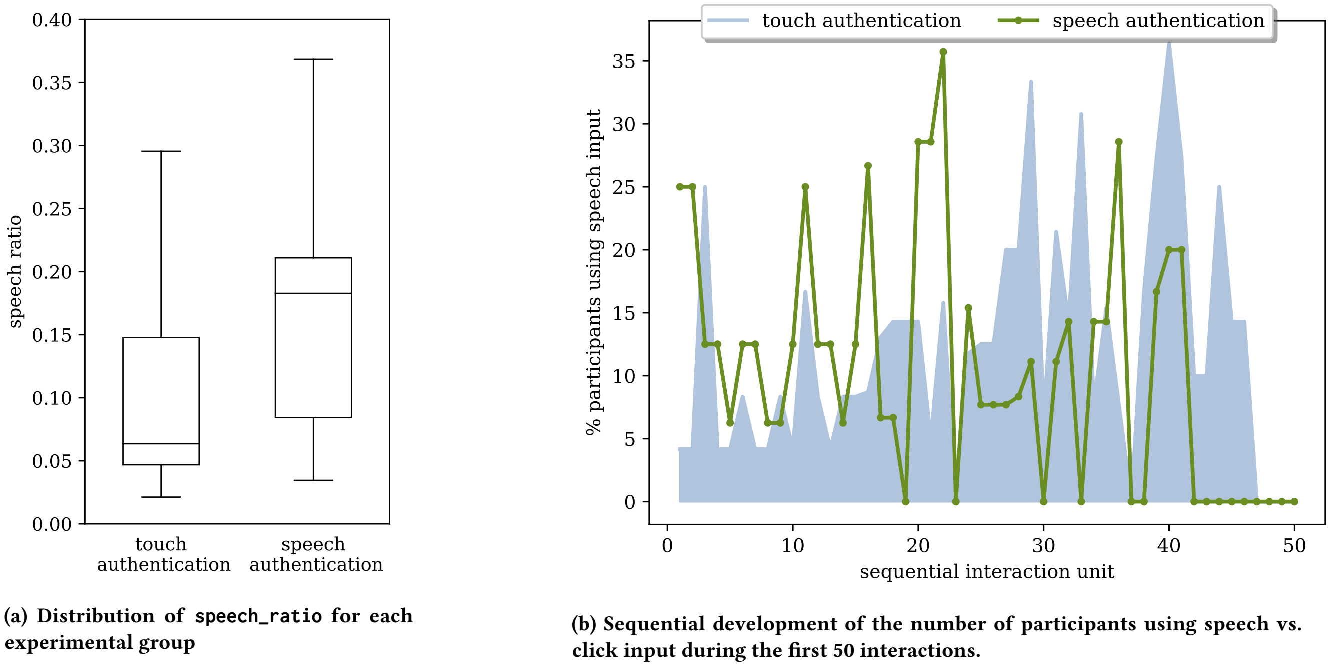
In our quest to answer **H1**, we investigated the link between the participants' choice of login modality and their attitude towards voice input as a control mechanism for the smart home application. We tested the link for two different measures of attitudes towards voice input. Responses from the follow-up questionnaire were used as a subjective measure of the usability of speech input. Additionally, we used the interaction behavior as an objective measure.

Correlation with perceived usability: To determine whether the authentication modality affects the perceived usability of speech input during later interactions, we summarized the usability metrics U1-U7 into three constructs representing *Ease of Use* (U1, U4, U5), *Confidence* (U2, U3), and *Perceived Success* (U6, U7) of speech input. Table 3 shows that speech authenticators assign slightly higher usability scores to *Ease of Use* and *Confidence*, while evaluating *Perceived Success* lower than touch authenticators. Yet the effect is not significant for any of the usability constructs. Using two-sided t-tests, we also found no significant effect for any of the individual usability metrics. We conclude that speech authenticators did not ascribe a higher usability to speech input than touch authenticators.

modality		Ease of Use	Confidence	Perceived Success
		(U1, U4, U5)	(U2, U3)	(U6, U7)
Touch	μ	3.26	3.77	2.29
	σ	(1.58)	(1.35)	(1.59)
Speech	μ	3.73	4.06	2.06
	σ	(1.07)	(1.06)	(1.39)
	p-value	.2854	.4627	.6424

Table 3: Usability evaluation of speech input per chosen authentication modality. Statistical significance of inter-group differences are calculated with a two-sided t-test.

Correlation with interaction behavior during routine tasks: We analyzed whether the participants' input choice during authentication is representative for later input choices. Table 4 summarizes the



(a) Distribution of speech_ratio for each experimental group

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

Figure 1: Usage of input modalities after authentication

average values of the interaction metrics for the two experimental groups (i.e., touch vs. speech authentication).

modality		click _count	speech _count	speech _ratio	modality _switches
Touch	μ	43.17	5.08	12.85%	8.75
	σ	(44.08)	(5.13)	(14.26%)	(5.76)
Speech	μ	32.06	5.56	16.56%	8.25
	σ	(16.80)	(2.83)	(9.11%)	(4.02)
	p-value	.2828	.7138	.3338	.7564

Table 4: Interactions per chosen authentication modality. Significance of inter-group differences is calculated with a two-sided t-test.

Figure 1a shows that the median speech_ratio of voice authenticators is higher than for participants who logged into the system using touch input. Yet a two-sided t-test shows that the participants' choice of input modality after authentication does not significantly differ between the experimental groups. From the modality usage over time (cf. Figure 1b) it can be seen that the proportion of speech input from voice authenticators 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, independent of the chosen authentication modality. Yet most participants frequently revert to speech input multiple times throughout the interaction, with on average 8.6 modality switches.

Interaction effects: Since the task complexity did not change throughout the interaction, other factors must be responsible for the participants' frequent switching behavior. Contrary to observations that have been reported in the literature, the participants in our study used speech input more often for directional commands such as "up"/"down" (56.6%) than for semantic commands, e.g., for selecting of the "Weather Forecast" control panel. For touch input, we observed the reverse: 57.4% of clicks were attributed to semantic commands. A preference for touch input has typically been related to directional commands for controlling continuous functions [6], whereas the directional commands in our smart home application are used for singular adjustments (e.g., switching on the light). Speech input is preferred for semantic commands that are considerably longer [1] or not clearly defined [10]. In contrast, we used very short commands throughout the experiment. With the experimental setup of our user study and with the configurations of our smart home application, 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 observed interaction behavior of the participants. The logged speech input data shows that the speech recognition quality is highly dispersed across the participants, with the executable_command_ratio ranging from .2 to 1. (mean = .687, stdv. = .265). The executable_command_ratio 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 avg_speech_confidence ranges from .538 to .933 (mean = .847, stdv. = .070). Yet Pearson's correlation coefficient provides no evidence that a higher speech recognition quality increased the use of speech input. As can be seen in Figure 2, a higher ratio of successfully executed speech commands is even negatively correlated with `speech_ratio` (corr. = -.11, p-value = .52).

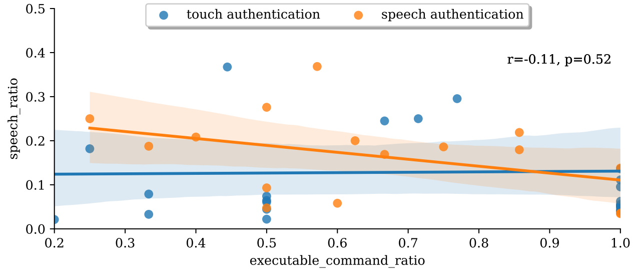


Figure 2: Correlation of the number of speech inputs with the number of executable commands. Significance is calculated for the combined data of both experimental groups. No positive relationship exists between the ratio of successfully executed spoken commands and the use of speech input.

However, as can be seen from the data in Figure 3, 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 metric `executable_command_ratio`. The relationship is significant on a 95% confidence level. 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.

Based on our findings, we reject the hypothesis that the chosen mode of authentication reveals a stronger inclination to use speech input in non-critical tasks (**H1**). Instead, attitudes towards speech interaction are formed gradually as the users perform the routing task, and are mainly driven by the quality of the speech recognition.

5.3 H2: Users prefer system output in the same sensory modality they use for input commands

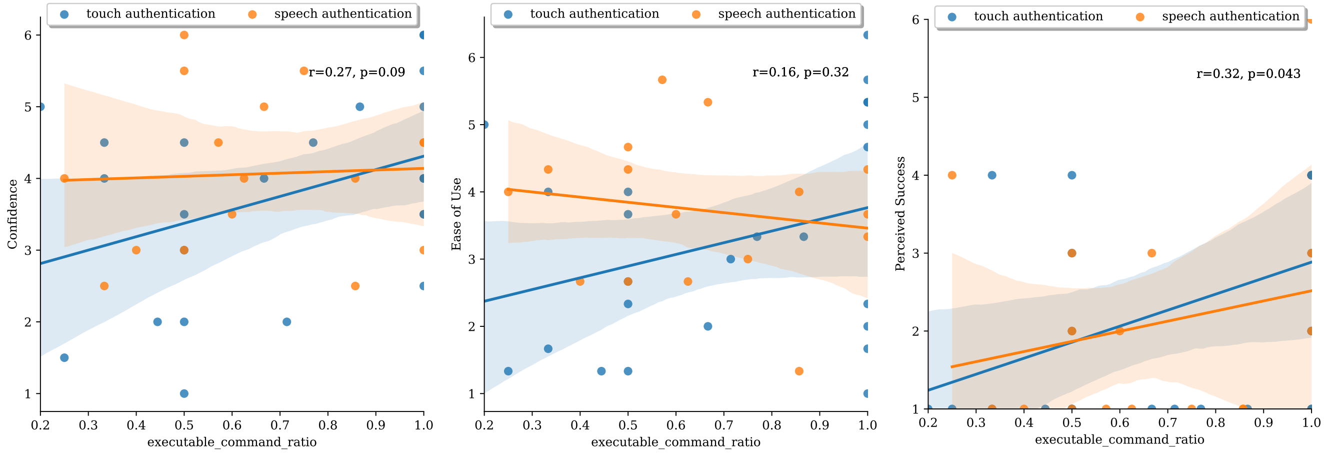
Our previous analyses revealed that the users do not necessarily continue using the communication channel they used during the authentication procedure. We therefore first tested **H2** for the chosen authentication modality. We then repeated the analysis for the dominant input modality during the subsequent non-critical routine tasks.

Correlation of authentication input with usability of speech output: We tested whether participants of the treatment group (i.e., system outputs match the authentication modality) evaluated the system usability different than participants of the control group (i.e., system outputs do not match the authentication modality).

Figure 4 shows the participants' *Perceived Usability* (measured as the mean value of **O1-O4**), as well as the degree to which they felt that *Insufficient Audio* (**O5**), or *Insufficient Text* (**O6**) output was available.

From the distributions in Figures 4a-4b it can be seen 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 shows that, for both **O5** and **O6**, the difference between the experimental groups is not statistically significant.

Paradoxically, *Perceived Usability* is evaluated slightly higher by participants of the control group, where the output modality did not match their chosen authentication modality (cf. Figure 4c). Yet the effect is not significant. The fine-grained analysis of the individual

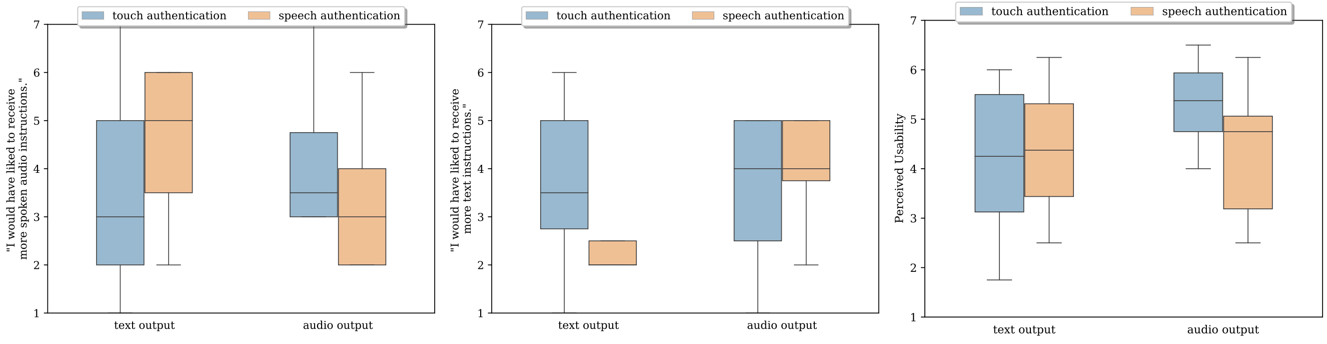


(a) Confidence

(b) Ease of Use

(c) Perceived Success

Figure 3: Usability evaluation of speech input in relation to speech recognition quality. Speech recognition quality is measured by the metric `executable_command_ratio`.



(a) Insufficient Audio **(b) Insufficient Text** **(c) Perceived Usability**

Figure 4: Usability evaluation of text and audio output in dependence of the chosen authentication modality.

usability measures (**O1**-**O4**) in Table 5 shows that, even when examined in isolation, the difference between the experimental groups is not significant for any of the measures.

Correlation of routine task input with usability of speech output: Given that we did not find a clear link between the authentication modality and the frequency of speech interactions in subsequent non-critical routine tasks, we additionally tested whether users who more frequently used speech input throughout the entire interaction had a more positive perception of the usability of speech output. While a small positive correlation exists (cf. Figure 5), the relationship is not significant (corr. = .11, p-value = .52).

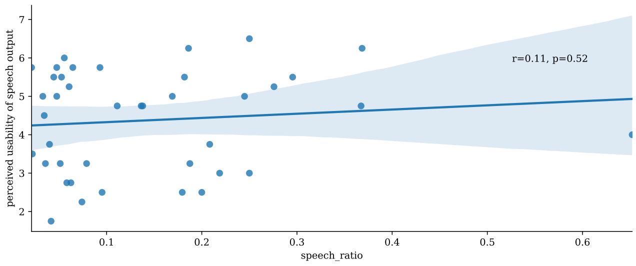


Figure 5: Usability evaluation of speech output in relation to the frequency of speech interaction during non-critical routine tasks (measured by `speech_ratio`).

Our data therefore provide no support for the hypothesis that users prefer system output in a format that matches their sensory input modality (**H2**). This finding is consistent across all evaluated task types, and thus applies to security sensitive authentication as well as non-critical routine tasks.

6 DISCUSSION

Based on the data from our 40 study participants interacting with the smart home application, we found indications that users who authenticate themselves using 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 routine tasks than touch authenticators. We therefore found no support for the hypothesis that the interaction behavior during login is representative for the user's inclination to use speech input in non-critical tasks (**H1**). 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 their perceived usability of speech input. Similar behaviors have been observed in a study by Rebman et al. [41], where users of a meeting support applications were dissatisfied with the speech recognition quality, but would still use the technology for future interactions. In contrast, Schaffer et al. [45] report opposing results from a study in which the users of a restaurant booking

output	authentication			Perceived Usability				Insufficient Audio	Insufficient Text
				O1	O2	O3	O4	O5	O6
Text	touch	N = 20	μ	4.60	3.75	4.70	4.35	3.40	3.55
			σ	(1.91)	(1.97)	(1.97)	(1.74)	(1.77)	(1.60)
	speech	N = 4	μ	5.50	4.25	5.25	5.50	4.50	2.50
			σ	(1.66)	(1.92)	(1.48)	(0.87)	(1.66)	(0.87)
Audio	touch	N = 4	μ	6.00	5.25	6.00	4.00	4.25	3.50
			σ	(0.00)	(1.48)	(1.00)	(1.41)	(1.64)	(1.66)
	speech	N = 12	μ	5.00	4.50	4.67	4.58	3.25	4.00
			σ	(1.78)	(2.06)	(1.93)	(1.75)	(1.23)	(0.91)

Table 5: Usability evaluation of audio output per authentication and activated output modality.

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 (**H2**).

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 [16] 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 [12, 32]. 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 [32]. 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.

7 CONCLUSION

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 vs. audio) to the communication channel that was used to issue the first few commands (i.e., touch vs. 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.

REFERENCES

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

[2] Farzana Alibay, Manolya Kavakli, Jean-Rémy Chardonnet, and Muhammad Zeeshan Baig. 2017. The usability of speech and/or gestures in multi-modal interface systems. In *Proceedings of the 9th International Conference on Computer and Automation Engineering* (Sydney, Australia) *(ICCAE '17)*. ACM, New York, NY, USA, 73–77. https://doi.org/10.1145/3057039.3057089

[3] Amazon.com Inc. 2021. Amazon Alexa Voice AI. https://developer.amazon.com/en-US/alexa

[4] Apple Inc. 2021. Siri does more than ever. Even before you ask. https://www.apple.com/siri/

[5] Muhammad Zeeshan Baig and Manolya Kavakli. 2018. Qualitative analysis of a multimodal interface system using speech/gesture. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA '18)*. IEEE, Wuhan, China, 2811–2816. https://doi.org/10.1109/ICIEA.2018.8398188

[6] John M. Bierschwale, Carlos E. Sampaio, Mark A. Stuart, and Randy L. Smith. 1989. Speech versus manual control of camera functions during a telerobotic task. *Proceedings of the Human Factors Society Annual Meeting* 33, 2 (1989), 134–138. https://doi.org/10.1177/154193128903300229

[7] Thayne Breetzke and Stephen V. Flowerday. 2016. The usability of IVRs for smart city crowdsourcing in developing cities. *Electron. J. Inf. Syst. Dev. Ctries.* 73, 1 (2016), 1–14. https://doi.org/10.1002/j.1681-4835.2016.tb00527.x

[8] Nikola Bubalo, Frank Honold, Felix Schüssel, Michael Weber, and Anke Huckauf. 2016. User expertise in multimodal HCI. In *Proceedings of the European Conference on Cognitive Ergonomics* (Nottingham, UK) *(ECCE '16)*. ACM, New York, NY, USA, Article 10, 6 pages. https://doi.org/10.1145/2970930.2970941

[9] Stéphanie Buisine and Jean-Claude Martin. 2003. Design principles for cooperation between modalities in bi-directional multimodal interfaces. In *Proceedings of the CHI 2003 workshop on Principles for multimodal user interface design*. Ft. Lauderdale, FL, USA, 72–75.

[10] Stéphanie Buisine and Jean-Claude Martin. 2003. Experimental evaluation of bi-directional multimodal interaction with conversational agents. In *IFIP TC13 International Conference on Human-Computer Interaction (INTERACT '03)*. IOS Press, (c) IFIP, Zurich, Switzerland, 168–175.

[11] José Coelho and Carlos Duarte. 2011. The contribution of multimodal adaptation techniques to the GUIDE interface. In *International Conference on Universal Access in Human-Computer Interaction. Design for All and eInclusion. LNCS, vol. 6765* (Orlando, FL, USA) *(UAHCI '11)*. Springer, Berlin, Heidelberg, 337–346. https://doi.org/10.1007/978-3-642-21672-5_37

[12] José Coelho, Carlos Duarte, Pradipta Biswas, and Patrick Langdon. 2011. Developing accessible TV applications. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (Dundee, UK) *(ASSETS '11)*.

ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/2049536.2049561

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

[14] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can I help you with?": Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[15] Carlos Duarte and Luís Carriço. 2006. A conceptual framework for developing adaptive multimodal applications. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (Sydney, Australia) *(IUI '06)*. ACM, New York, NY, USA, 132–139. https://doi.org/10.1145/1111449.1111481

[16] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. Multimodal interfaces: A survey of principles, models and frameworks. *Human Machine Interaction* 5440 (2009), 1–25. https://doi.org/10.1007/978-3-642-00437-7_1

[17] Python Software Foundation. 2021. Graphical User Interfaces with Tk. https://docs.python.org/3/library/tkinter.html

[18] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[19] Fabian Hoffmann, Miriam-Ida Tyroller, Felix Wende, and Niels Henze. 2019. User-defined interaction for smart homes: Voice, touch, or mid-air gestures?. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia* (Pisa, Italy) *(MUM '19)*. ACM, New York, NY, USA, Article 33, 7 pages. https://doi.org/10.1145/3365610.3365624

[20] Aseel Ibrahim, Jonas Lundberg, and Jenny Johansson. 2001. Speech enhanced remote control for media terminal. In *7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*. ISCA, Aalborg, Denmark, 2685–2688.

[21] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* 108, 1 (2007), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

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

[23] Kristiina Jokinen and Topi Hurtig. 2006. User expectations and real experience on a multimodal interactive system. In *Proc. Interspeech 2006*. ISCA, Pittsburgh, PA, USA, 1049–1052. https://doi.org/10.21437/Interspeech.2006-156

[24] Veton Këpuska and Gamal Bohouta. 2017. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl* 7, 03 (2017), 20–24.

[25] Gierad P. Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 2185–2194. https://doi.org/10.1145/2470654.2481301

[26] Minkyung Lee, Mark Billinghurst, Woonhyuk Baek, Richard Green, and Woontack Woo. 2013. A usability study of multimodal input in an augmented reality environment. *Virtual Reality* 17, 4 (2013), 293–305.

[27] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[28] Ludo Maat and Maja Pantic. 2007. Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing, LNCS*, Vol. 4451. Springer, Berlin, Heidelberg, 251–271. https://doi.org/10.1007/978-3-540-72348-6_13

[29] Miroslav Melichar and Pavel Cenek. 2006. From vocal to multimodal dialogue management. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (Banff, Alberta, Canada) *(ICMI '06)*. ACM, New York, NY, USA, 59–67. https://doi.org/10.1145/1180995.1181008

[30] Microsoft. 2012. Microsoft Speech API (SAPI) 5.3. https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ms723627(v=vs.85)

[31] Microsoft. 2021. Cortana - Your personal productivity assistant in Microsoft 365. https://www.microsoft.com/en-us/cortana

[32] Meredith Ringel Morris. 2012. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces* (Cambridge, MA, USA) *(ITS '12)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2396636.2396651

[33] Sharon Oviatt. 1997. Mulitmodal interactive maps: Designing for human performance. *Hum. Comput. Interact.* 12, 1-2 (1997), 93–129. https://doi.org/10.1080/07370024.1997.9667241

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

[35] Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. 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 (2000), 263–322. https://doi.org/10.1207/S15327051HCI1504_1

[36] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (State College, PA, USA) *(ICMI '04)*. ACM, New York, NY, USA, 129–136. https://doi.org/10.1145/1027933.1027957

[37] Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson, and Lesley Carmichael. 2003. Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, Canada) *(ICMI '03)*. ACM, New York, NY, USA, 44–51. https://doi.org/10.1145/958432.958443

[38] Geoff Poulton. 2020. Why trust and autonomy are essential factors when working from home. https://www.rolandberger.com/en/Insights/Publications/The-home-office-becomes-the-new-normal.html

[39] Sarah Prange, Ceenu George, and Florian Alt. 2021. Design considerations for usable authentication in smart homes. In *Mensch Und Computer 2021* (Ingolstadt, Germany) *(MuC '21)*. ACM, New York, NY, USA, 311–324. https://doi.org/10.1145/3473856.3473878

[40] Stefan Profanter, Alexander Perzylo, Nikhil Somani, Markus Rickert, and Alois Knoll. 2015. Analysis and semantic modeling of modality preferences in industrial human-robot interaction. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Hamburg, Germany, 1812–1818. https://doi.org/10.1109/IROS.2015.7353613

[41] Carl Rebman, Brian Reithel, and Casey Cegielski. 2001. An exploratory study of speech recognition technology and its implications for current electronic meeting support applications. In *7th Americas Conference on Information Systems (AMCIS '01)*. AISeL, Boston, MA, USA, 298–306. https://aisel.aisnet.org/amcis2001/61

[42] Napa Sae-Bae, Jonathan Wu, Nasir Memon, Janusz Konrad, and Prakash Ishwar. 2019. Emerging NUI-based methods for user authentication: A new taxonomy and survey. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1, 1 (2019), 5–31. https://doi.org/10.1109/TBIOM.2019.2893297

[43] Rajwant Sandhu and Benjamin J. Dyson. 2012. Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses. *Acta Psychologica* 140, 2 (2012), 111–118. https://doi.org/10.1016/j.actpsy.2012.04.003

[44] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2011. Measuring cognitive load for different input modalities. In *9. Berliner Werkstatt Mensch-Maschine-Systeme*. Berlin, Germany, 287–292.

[45] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2015. Modeling input modality choice in mobile graphical and speech interfaces. *Int. J. Hum. Comput. Stud.* 75 (2015), 21–34. https://doi.org/10.1016/j.ijhcs.2014.11.004

[46] Mark Seligman and Mike Dillinger. 2006. Usability issues in an interactive speech-to-speech translation system for healthcare. In *Proceedings of the Workshop on Medical Speech Translation* (New York, New York) *(MST '06)*. ACL, USA, 1–4.

[47] Amanda L. Smith and Barbara S. Chaparro. 2015. Smartphone text input method performance, usability, and preference with younger and older adults. *Human Factors* 57, 6 (2015), 1015–1028. https://doi.org/10.1177/0018720815575644

[48] Doroteo T. Toledano, Rubén Fernández Pozo, Álvaro Hernández Trapote, and Luis Hernández Gómez. 2006. Usability evaluation of multi-modal biometric verification systems. *Interacting with Computers* 18, 5 (03 2006), 1101–1122. https://doi.org/10.1016/j.intcom.2006.01.004

[49] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. 2012. Biometric authentication on a mobile device: A study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference* (Orlando, FL, USA) *(ACSAC '12)*. ACM, New York, NY, USA, 159–168. https://doi.org/10.1145/2420950.2420976

[50] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

[51] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the security of graphical passwords: The case of Android unlock patterns. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (Berlin, Germany) *(CCS '13)*. ACM, New York, NY, USA, 161–172. https://doi.org/10.1145/2508859.2516700

[52] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly* 27, 3 (2003), 425–478. http://www.jstor.org/stable/30036540

[53] Wolfgang Wahlster. 2003. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In *Proceedings of the Human Computer Interaction Status Conference*. Berlin, Germany, 47–62.

[54] Adam S. Williams, Jason Garcia, and Francisco Ortega. 2020. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Trans. Vis. Comput. Graph.* 26, 12 (2020), 3479–3489. https://doi.org/10.1109/TVCG.2020.3023566