Enhancing the Accessibility of Voice Assistants for Individuals with Dysarthria through Non-Verbal Voice Cue Interaction

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Aisha Jaddoh

2024

Cardiff University
School of Computer Science & Informatics

Abstract

This thesis presents a comprehensive exploration into the interaction between individuals with dysarthria and Smart Voice Assistants (SVAs), focusing particularly on the utilization of a novel interaction method: non-verbal voice cues, and assessing the efficacy of different interaction methods. Dysarthria, a speech disorder resulting from neurological injury, significantly impairs speech clarity, challenging those attempting to interact with SVAs that typically rely on clear verbal commands.

The thesis examines the needs of users with dysarthria, firstly through a review of related literature to situate our thesis. This research examines current gaps in voice recognition technology for individuals with dysarthria and limitations in interaction. We follow this by conducting three studies to investigate the current use of SVAs by individuals with dysarthria and to evaluate the efficacy of alternative interaction methods in enhancing accessibility.

Through user-centered and participatory design, we develop 'Daria,'; a bespoke system designed to use non-verbal cues, enabling users to leverage sounds they can articulate comfortably. We thereafter use Daria to further test our hypotheses. This method effectively bypasses the need for intermediary devices, thus reducing interaction fatigue and streamlining communication. We conducted comprehensive research involving three sequential studies to identifying user challenges, evaluating interaction effectiveness and command mapping, and comparing usability and user experience across different design options against alternative methods.

The findings conclude that non-verbal voice cues are viable and preferred by individuals with moderate to severe dysarthria due to their simplicity and ease of use, highlighting the effectiveness of this interaction method in facilitating more reliable and efficient communication for users. The mapping approach within the system proved to be successful and memorable for users, enhancing overall system usability. Consequently, non-verbal voice cues have been identified as a highly usable technique for individuals with dysarthria, offering them an accessible and efficient user experience.

Contents

List of Tables

List of Figures

List of Acronyms

ASRs automatic speech recognition systems

ASR automatic speech recognition system

AAC Augmentative and Alternative Communication systems

AAC Augmentative and alternative communication

ALS Amyotrophic lateral sclerosis

AMQP Advanced message queuing protocol

AT Assistive technology

CNN Convolutional neural network

COVID-19 Coronavirus disease

CP cerebral palsy

HCI Human–computer interaction

ISO International Standards Organization

MOM Message-oriented middleware technology

MS multiple sclerosis

NASA-TLX NASA task load index

PD Parkinson Disease

PRISMA Preferred reporting items for systematic review and meta analysis

SASSI Subjective assessment of speech system interfaces

SBAHC Sultan Bin Abdulaziz Humanitarian City

SDGs Sustainable Development Goals

SUS System usability scale

SVAs Smart voice assistants

TBI Traumatic brain injury

WER Word error rate

VUI voice user interface

List of Publications

The work introduced in this thesis is based on the following peer-reviewed publications, in which my contributions are briefly described below:

[1] Jaddoh, A., Loizides, F., Rana, O. and Syed, Y.A., 2024. Interacting with Smart Virtual Assistants for Individuals with Dysarthria: A Comparative Study on Usability and User Preferences. Applied Sciences, 14(4), p.1409.

- In this paper, I designed and conducted the study, analyzed and interpreted the collected data,and wrote the draft of the manuscript.

[2] Jaddoh, A., Loizides, F., Lee, J. and Rana, O., 2023. An interaction framework for designing systems for virtual home assistants and people with dysarthria. Universal Access in the Information Society, pp.1-13.

- In this paper, I designed and conducted the study, analyzed and interpreted the collected data, developed the theoretical framework, and wrote the draft of the manuscript.

[3] Jaddoh, A., Loizides, F. and Rana, O., 2023. Interaction between people with dysarthria and speech recognition systems: A review. Assistive Technology, 35(4), pp.330-338.

- In this paper I conducted the extensive literature review, analyzed the selected studies, and wrote the first draft of the manuscript.

[4] Jaddoh, A., Loizides, F. and Rana, O., 2021. Non-verbal interaction with virtual home assistants for people with dysarthria. The Journal on Technology and Persons with Disabilities, p.71.

- In this paper, I wrote the draft of the manuscript.

- Paper under review at ACM Transactions on Accessible Computing peer reviewed journal. Jaddoh, A., Loizides, F., Alrefai, K. and Rana, O.. Overcoming Speech Barriers: Non-Verbal Voice Cue Interaction Technique for Enhancing Smart Voice Assistant Accessibility for Individuals with Dysarthria.

  - In this paper, I designed and conducted the study, analyzed and interpreted the collected data,and wrote the draft of the manuscript.

Acknowledgements

I am profoundly grateful for the guidance, support and wisdom provided by my supervisor Fernando Loizides. His mentorship has been a beacon of light throughout this journey, offering the perfect degree of availability and guidance while granting me the freedom to explore independently. This unique approach has not only nurtured my academic skills but has also instilled in me the confidence to venture into uncharted territories of research with curiosity and vigour. Moreover, there are lessons and advice I have learned from him that will stay with me forever, shaping my professional ethos and personal philosophy. I am also thankful that he cheered me on every step of the way. Equally, I extend my deep appreciation to my second supervisor, Omer. The opportunities he has provided me have been transformative, opening doors to new avenues of research and collaboration within the academic community.

I extend my heartfelt gratitude to my country, Saudi Arabia, and my sponsors, the Royal Commission for Jubail and Yanbu and Yanbu Industrial College, for their generous support and for providing me with the opportunity to pursue my PhD. My sincere appreciation goes to Sultan Bin Abdulaziz Humanitarian City for its invaluable assistance and for granting me permission to conduct my study on its premises and with its patients. The support, cooperation and resources from its therapists and research centre have been instrumental for the successful completion of my work.

I cannot express enough thanks to my parents, whose love, guidance and sacrifices have shaped me into the person I am today. I owe immense gratitude to my mother, who dedicated significant time to comprehending my work and provided unwavering support at every stage. Her commitment to understanding the intricacies of my research and her encouragement throughout the process has been profoundly inspiring and motivational.

To my husband and children, your boundless patience has been the quiet force behind my perseverance. Your understanding during the long hours of dedication to my work, the countless times you have embraced my absences with grace and the way you have shouldered additional responsibilities to give me the space to focus are gestures of love and support that I never took for granted. Your unwavering encouragement has been a constant source of strength and comfort that allowed me to chase my dreams without hesitation.

To my sisters, brothers and friends, your unwavering faith in me and your boundless enthusiasm for my work have been the bedrock of my resilience. Your ability to listen, offer constructive criticism and provide moral support during times of doubt and stress has been invaluable. Your genuine interest in my research, your readiness to celebrate my smallest victories and your constant reminders of the importance of balance and self-care have enriched my PhD journey in ways I can never fully express. Your companionship and loyalty have not only brightened my days but have also given me the strength to navigate the challenges of this rigorous academic endeavor.

This thesis is not just a reflection of my work but also a testament to the collective support and belief of all those mentioned above. Thank you all for being my pillars of strength and for contributing in countless ways to the completion of this significant chapter in my life.

To those who walked beside me,
your presence was my light and strength.

Chapter 1

Introduction

Smart voice assistants (SVAs), such as Google Home and Amazon Alexa, are voice-activated devices that have been widely adopted in the consumer industry. The number of households using smart home products is projected to reach nearly 380 million by the end of 2024 [5]. These voice-activated devices use automatic speech recognition systems (ASRs) and are controlled through voice commands to perform tasks. The voice user interface (VUI) underlying these SVAs allows for hands-free and eyes-free interaction. This facilitates greater accessibility to various devices and services, especially when multitasking [6]. Thus, VUIs require less visual demand and are less distracting than other interfaces [7, 8]. An example of a task that benefits from this voice interface occurs while driving a car and the user can play a text message or change the music without using their hands and eyes. Users can control not only SVAs but also various smart devices, such as doorbells, or access services, such as playing the news, through voice commands. The advantage of SVAs and VUIs is not only that it is hand and eye free but also this technology could help to perform tasks faster [9]. For example, according to Pearl [10], speaking a text message is faster than typing the message.

These devices also offer valuable benefits for individuals who have disabilities [11,12], serving as assistive tools that provide enhanced access for those who may struggle with conventional interfaces. For example, users who are visually impaired do not need to visually navigate screens to request a service; they can simply utter the commands. Moreover, the response is also provided through voice, eliminating the need to struggle with reading text. In addition, individuals who have physical disabilities or limited hand dexterity can, for example, turn on a light without physical movement, allowing for greater independence.

Despite their widespread use, there are some challenges that affect users' experiences. One of the main challenges that users face is unstable speech recognition, leading to misinterpretations [11, 13], for instance, people who have heavy accents or use slang language [14, 15]. This extends to some individuals, especially those who have speech disabilities, who encounter challenges when interacting with SVAs [12, 16, 17]. Given that SVAs depend primarily on verbal communication, those who have speech impairments might struggle to use these devices as effectively as individuals without such issues. The challenges can intensify depending on the severity of the impairment [18]. One key issue is that these devices are not designed to recognise and understand their speech style. In other words, these devices are not typically trained on nonstandard or impaired speech patterns [19]. A primary obstacle to such training is the difficulty in collecting a large number of recordings from those who have speech impairments [20, 21], especially from individuals with dysarthria , which is considered one of the most severe communication disorders [16, 21].

As an alternative to directly interacting with SVAs, users who have speech impairments use intermediary devices. For example, they use tablets, phones, eye gaze systems or a mouse and keyboard to perform various tasks or send commands to SVAs [22]. Thus, users employ various input methods and modalities to interact with voice interfaces. However, having a physical disability that is usually accompanied by dysarthria can limit this interaction [21, 23]. All of this leads to users who have speech impairments having a different experience than users without the disability: they may have to use intermediary devices, which can result in slower interactions, or sometimes users cannot access potentially life-changing devices.

Various studies have been conducted in the field. Prior studies have focused on improving speech recognition models to understand people who have dysarthria [23–28].

However, overall, these studies highlight the need for better models. Other studies have centred around assistive technology devices entering the market. These devices are designed for individuals with dysarthria ; however, as previously explained, users may still need to rely on intermediary devices. In addition, various techniques have been tested for interacting with SVAs. One example is brain–computer interactions [29], in which brain signals are interpreted as voice commands to interact with SVAs. However, this approach is currently an intrusive method that is not yet mature enough for commercial use.

In addressing these challenges, this study centres on dysarthria, which is considered one of the most severe communication impairments [30]. This condition results in slow, weak, imprecise or uncoordinated movements of the speech muscles, making it challenging for individuals to control them effectively. Taking into account these distinctive features of dysarthric speech, we introduce a novel interaction technique that empowers people who have dysarthria to communicate with SVAs within their capabilities, effectively overcoming their communication obstacles. This method involves engaging with SVAs through nonverbal voice cues, which are sounds uttered by users that are not words, such as ‘aaaah’. For example, instead of saying, ‘Hey Google, turn on the light’, the user simply utters, ‘aaah’. In this, we are embracing an ability-based design approach [31], which fundamentally shifts the focus from compensating for user limitations to harnessing their existing abilities.

By aligning with the principles of ability-based design [31], this approach stands as a testament to the potential of inclusive technology. It not only enhances accessibility for those who have speech impairments but also ensures that the technology adapts to the user, rather than the other way around. This approach aligns with not only our aim to create a more accessible digital environment but also a broader commitment to universal design, specifically, a commitment to creating systems that are usable by all people to the greatest extent possible without the need for adaptation or specialised design [31].

This approach has several advantages that enhance the interaction experience for individuals with dysarthria , aligning it more closely with the experience of users without speech impairments: it allows for direct interaction with SVAs without the need for intermediary devices, because using intermediary devices causes fatigue for users [32]; it uses voice as an input method, because some people who have dysarthria prefer to use their voices to the maximum extent [32, 33]; and, importantly, it remains within users' speech capabilities.

A system was implemented to utilise this technique, and the system is called 'Daria'. The name is extracted from the word 'DysARthrIA' and uses all the letters in the same order. Our study makes a significant contribution to narrowing the accessibility gap. It not only enables individuals with dysarthria to seamlessly engage with these devices but also enriches their lives. By providing them with the means to access various services and devices, we offer greater independence and the potential to accomplish tasks more efficiently.

In this thesis, I explore the challenges faced by individuals with dysarthria . I delve into the design of an alternative interaction technique that uses nonverbal voice cues and develop a framework for designing nonverbal interaction systems. The evaluation of these techniques includes a comparison of various design options of Daria. In addition, I compare interactions using Daria with verbal interactions through off-the-shelf SVAs, as well as other interaction modalities, such as eye gaze. The thesis concludes with recommendations for interaction system designers and suggestions for future research.

1.1 Research Questions

This work aims to answer the following research questions:

RQ1: How do individuals with dysarthria currently utilize smart voice assistants and what are their present experiences with these devices?

This question will help us understand the usage of SVAs by individuals with dysarthria, specifically the purpose and frequency of their use. Moreover, it will aid in understanding the challenges they face and the effectiveness of these devices in understanding their speech.

RQ2: Can a standardized vocabulary be developed for individuals with dysarthria that aligns with their unique speech capabilities and the range of sounds they can produce?

This question aims to help us develop a vocabulary that aligns with the speech capabilities of individuals with dysarthria. By doing so, we can improve the accuracy of speech recognition systems for this group, which in turn can reduce communication barriers and alleviate frustration. Furthermore, this tailored vocabulary could serve as a guide for future developers in creating more inclusive and accessible designs

RQ3: How does the use of non-verbal voice cue interaction techniques affect the user experience and usability of smart voice assistants?

RQ3.1: How memorable will the non-verbal voice cues be for users?

RQ3.2: How does the usability, user experience, and workload differ between the proposed interaction technique and using verbal interaction?

This question is asked to explore the feasibility of non-verbal voice cue interaction techniques. It is not enough for the solution to be technically feasible; we also need to investigate other aspects, such as usability and user experience. Additionally, the workload involved is a critical factor, especially since individuals with dysarthria can experience fatigue from speaking. As for memorability, this aspect is essential in designing an interaction system that users can adopt and use effectively over time. Ensuring that users do not forget how to interact with the system is crucial for its long-term success

RQ4: What is the impact of allowing customization rather than standardization on the interaction?

Currently, off-the-shelf SVAs do not offer customization options. Various studies have examined the role of customization in SVAs and have found that users generally prefer devices that can be customized. It has been shown that customization can enhance user satisfaction and improve performance. In this research, the focus is on understanding customization because it is a crucial aspect, especially for people with disabilities, including those with dysarthria in particular. Additionally, this exploration can provide valuable insights into understanding user needs, leading to the design of more effective and accessible systems.

1.2 Contributions

Towards this thesis, our work has produced the following peer-reviewed publications:

[1] Jaddoh, A., Loizides, F., Rana, O. and Syed, Y.A., 2024. Interacting with Smart Virtual Assistants for Individuals with Dysarthria: A Comparative Study on Usability and User Preferences. Applied Sciences, 14(4), p.1409.

[2] Jaddoh, A., Loizides, F. and Rana, O., 2023. Interaction between people with dysarthria and speech recognition systems: A review. Assistive Technology, 35(4), pp.330-338.

[3] Jaddoh, A., Loizides, F., Lee, J. and Rana, O., 2023. An interaction framework for designing systems for virtual home assistants and people with dysarthria. Universal Access in the Information Society, pp.1-13.

[4] Jaddoh, A., Loizides, F. and Rana, O., 2021. Non-verbal interaction with virtual home assistants for people with dysarthria. The Journal on Technology and Persons with Disabilities, p.71.

Contributions

C1 A systematic literature review that presents the limitations and gaps in the area of people who have dysarthria and automatic speech recognition systems and smart voice assistants.

C2 Analysis of interviews conducted with people who have dysarthria, exploring the use of smart voice assistants and revealing the challenges faced.

C3 An interaction framework for designing systems for smart voice assistants and people who have dysarthria.

C4 Creating a 'language' allowing users to keep the verbal modality, known as nonverbal voice cues.

C5 Creating a (currently non-existing) specialised utterances data set.

C6 Analysis of smart voice assistants, revealing the effectiveness of nonverbal voice cue systems in accommodating diverse dysarthria severities, which is a significant advancement in voice interaction technology.

C7 Validating the effectiveness of our bespoke system Daria.

Contribution C1 is relevant to the work in [2]. C2, C3 and C4 are relevant to [3]. Contribution C6 and C7 are relevant to [1]

1.3 Thesis structure

The research presented in this thesis is structured as follows, including a brief overview of each chapter.

Chapter 2 – Literature review: This chapter explores ASR systems and smart voice assistants for individuals with dysarthria . It discusses dysarthria, its impact on lives and assistive device usage and the current use of VUIs by individuals with dysarthria and presents a systematic review on this topic. The chapter concludes by identifying gaps in the literature, focusing on nonverbal voice cue interactions as an alternative method.

Chapter 3 – Methodology: This chapter addresses the methodology of the research, focusing on a user-centred design approach. It highlights the low accuracy of SVAs for individuals with dysarthria and the need for user involvement in SVA research and outlines the research methods used to develop more user-attuned solutions.

Chapter 4 – First Study – Understanding Challenges and Feedback: The primary aim of this chapter was to explore the challenges faced by individuals with dysarthria when using SVAs and gather feedback on nonverbal voice cues. Interviews with 19 participants who had dysarthria provided insights into the design components that are crucial for developing voice assistants for this demographic. This study corresponded to the first research question (RQ1) of the thesis.

Chapter 5 – Designing Nonverbal Voice Cue Interaction Systems: Focusing on understanding user requirements, this chapter introduces a standardised approach for designing nonverbal voice cue interactive systems. It discusses key design components, aligns them with user-centred design principles and elaborates on creating intuitive and memorable mappings between actions and sounds.

Chapter 6 – System Design and Preliminary Study – Daria: This chapter introduces 'Daria', the nonverbal voice cue interaction system designed for individuals with dysarthria . It details the system's design and evaluates its efficacy through a preliminary study. The chapter links theoretical concepts from previous chapters with practical application, focusing on the design and evaluation phases of the user-centred design process.

The subsequent chapters 7, 8, and 9 detail the main study of this thesis, with each chapter dedicated to a distinct aspect of this study to allow for a comprehensive analysis.

Chapter 7 – In-Depth Evaluation of Daria with Pre-mapped List: Building on the foundational work of Chapter 6, this chapter delves into a detailed evaluation phase of Daria using a pre-mapped list of commands. It presents a between-subject study to assess usability and user experience, providing comparative insights into various user interactions with the Daria system.

Chapter 8 – Customised List Interaction Study: This chapter continues the evaluation of Daria, focusing on a second group of participants interacting with a customised list. The chapter aimed to understand the impact of customisation on usability, memorability and overall user experience for individuals with dysarthria .

Chapter 9 – Comparing Interaction Modalities: Expanding the research scope, this chapter explores eye gaze interaction as an alternative modality and compares it with verbal and nonverbal voice interactions. The chapter aimed to assess the usability of each interaction method for individuals with dysarthria , contributing to a more comprehensive understanding of accessible communication technologies.

Chapter 10 – Conclusion and Future Directions: The final chapter revisits the research questions, combines and discusses the results from previous chapters and addresses the research gaps identified earlier. It also discusses recommendations for future designers, the limitations of this study and potential future research directions.

Figure 1.1 provides a visual representation of the thesis structure, mapping each chapter to the corresponding research questions it addresses.

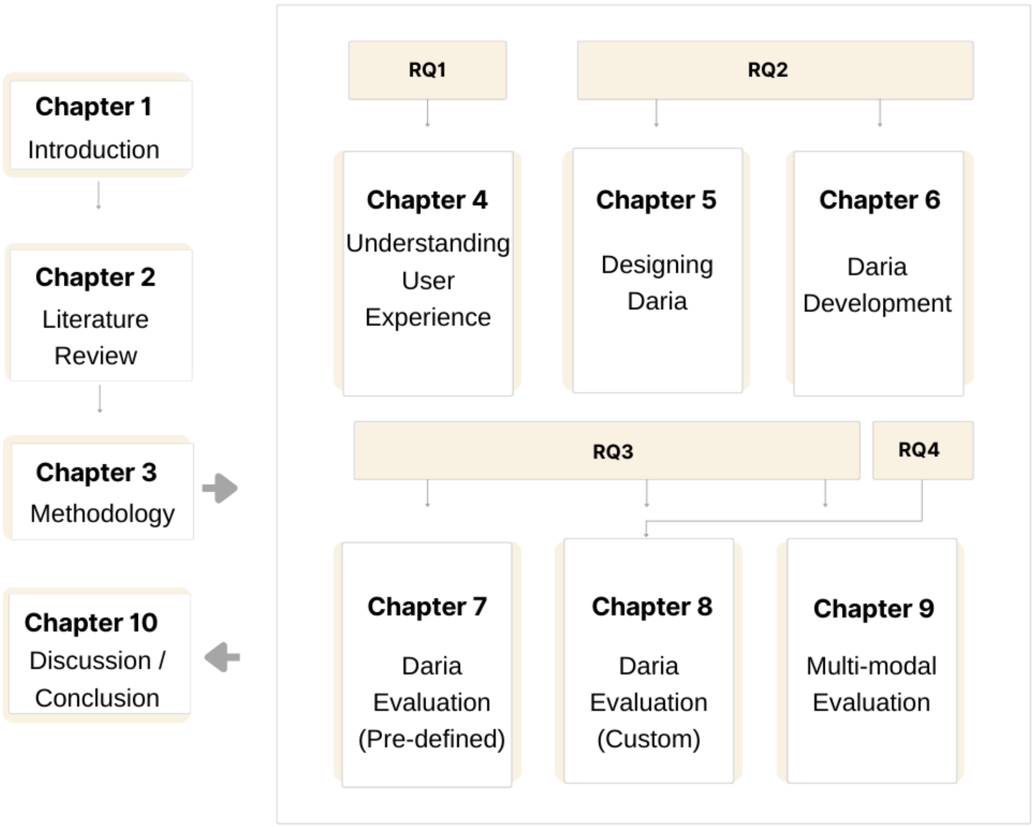


Figure 1.1: Thesis structure

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter explores the literature in the area of ASR systems and smart voice assistance when used by individuals with dysarthria. This frames the work we conducted in this thesis and identifies the limitations and gap in the literature, which contributes to answering the research questions.

The chapter starts with a general discussion about dysarthria, followed by an examination of the research about how dysarthria affects individuals' lives and the assistive devices they use. This helps in understanding the case of dysarthria and the effect. After that, research about VUIs and how they are used by individuals with dysarthria is covered. This section helps to explain the current uses of assistive devices and the options they have. In Sections 2.6.1 and 2.6.2, we describe the systematic review we conducted that was published in [2], including later research that was published after publication. This covers the status of current research. The results of these sections lead to Section 2.6, which discusses nonverbal voice cue interactions as an alternative method of interaction. Section 2.7 covers the methodologies in current research regarding non-verbal voice cue interaction for SVAs, specifically addressing dysarthria and the gap in this area.

The work in this chapter has been previously published in Assistive Technology Journal, the Official Journal of RESNA [2].

2.2 Dysarthria

Dysarthria is a neurological motor speech impairment caused by damage to the central or peripheral nervous system [34, 35]. This condition results in slow, weak, imprecise or uncoordinated movements of the speech muscles [34]. Dysarthria can be congenital or acquired. One of the most well-known congenital causes of dysarthria is cerebral palsy (CP), which typically begins in childhood. Estimates suggest that dysarthria appears in up to 90% of individuals who have CP [35]. Acquired dysarthria typically presents in adults and can be nonprogressive. Common causes include stroke and Traumatic brain injury (TBI). Dysarthria occurs in approximately one-third of people who have TBI [35]. Gradual onset dysarthria is associated with conditions such as Amyotrophic lateral sclerosis (ALS), in which it may be the initial symptom in 25% of cases, as well as multiple sclerosis (MS), Parkinson Disease (PD) and other conditions [35].

Dysarthric speech has various characteristic that are not necessarily exhibited by all individuals who have the condition; however, it varies from case to case [34]. These characteristics include a low rate of speech, poor articulation, speech muscle fatigue and low levels of speech intelligibility. One prominent characteristic is a slower rate of speech, which is slower than that of a normal speaker [36–38], that could lead to prolonged words and syllables and varied vowel spacing. Another aspect of dysarthria is poor articulation or pronunciation.

Prior studies have explored the relationship between dysarthria and consonants and vowels [39–41]. Consonants have been found to be challenging to pronounce for people who have dysarthria. Unvoiced consonants, for instance, are pronounced incorrectly [39, 40]. Similar to this, difficulties in pronouncing vowels are primarily related to front rounded vowels and mid-vowels, in which the tongue is positioned halfway to the roof of the mouth [41]. This consideration is crucial for the system design described in Chapters 5 and 6.

Another characteristic is speech muscle fatigue. Weak facial muscle endurance is commonly linked to dysarthria [42] and can be caused by problems in various parts of the nerve-muscle process. Studies have shown that a lack of strength and endurance in the lips and tongue are associated with conditions related to dysarthria, such as Parkinsonism [43], stroke and TBI [44]. This means that speech is tiring for individuals with dysarthria; the longer the speech, the more difficult it becomes.

Moving to the intelligibility characteristic, this refers to the degree to which speech by an individual who has dysarthria is understood by a listener [35, 45]. Given that dysarthria is caused by 'disturbances in muscular control over the speech mechanism' [34], this leads to reduced speech intelligibility. The intelligibility is reduced when the severity increases. The severity of dysarthria, which is measured by a speech and language pathologist or other rehabilitation professional, is classified into mild, moderate or severe. According to Duffy [35], mild dysarthria is when there are detectable disturbances in speech but it does not affect intelligibility. Moderate dysarthria is when there are detectable speech disturbances that affect intelligibility. Finally, severe dysarthria is when intelligibility is very low and the use of alternative communication is needed. Further, there is another cases where the classification between moderate and sever becomes challenging due to the overlapping characteristics in speech patterns and the variability of symptoms among individuals. Various factors can affect the severity, including intelligibility, speech rate and other factors [46, 47].

2.3 Dysarthria's effects on individuals

Dysarthria can have a severe impact on an individual's quality of life, particularly affecting their communication skills and independence [48,49]. Several researchers have explored the multifaceted impact of dysarthria [48–51]. Although the impairment directly impairs speech, its effects also extend to emotional and physiological wellbeing and relationships [49]. Hindered communication can also lead to a host of secondary issues, such as frustration, social isolation and emotional distress. According to Light's [52] definition, communication is not merely about exchanging information; it also encompasses expressing needs, engaging in social interactions and even influencing moods and feelings.

One of the areas of life that dysarthria affects is relationships [48–51, 53]. Although dysarthria can occasionally have a positive impact, such as by bringing families closer together, it generally negatively affects relationships. For example, difficulties in communication can prevent individuals from bonding with their children [49]. One of the most compelling findings is that impaired communication leads to a sense of social isolation [48–51, 53]. Social isolation is another concern. individuals with dysarthria find it challenging to keep up with conversations, leading to an avoidance of social interactions and limited self-expression. Affected individuals often restrict their conversations to only essential exchanges, further contributing to feelings of loneliness [49]. This avoidance of social situations has various underlying causes, one of which is the frustration of having to repeatedly clarify themselves. Emotional wellbeing is also affected [54]. individuals with dysarthria often report feelings of insecurity, helplessness and low self-esteem. Emotional disturbances, such as feelings of frustration, anger and sadness, are also common outcomes of dysarthria.

In addition to affecting communication, dysarthria affects individuals' independence [51]. Often accompanied by physical disabilities, individuals with dysarthria may lose their ability to perform everyday tasks. Independence is not just a matter of convenience; it also pertains to safety and security. An interview by Beukelman et al. [55] highlighted that the ability to make phone calls, for example, is crucial for the safety and security of individuals with dysarthria.

2.4 Assistive technology

As dysarthria progresses, many individuals turn to Assistive technology (AT), which was defined by the Technology-Related Assistance for Individuals with Disabilities Act of 1988 (PL 100–407) as 'any item, piece of equipment, or product system, whether acquired commercially off the shelf, modified, or customised, that is used to increase, maintain, or improve functional capabilities of individuals with disabilities'. These assistive devices can be for supporting mobility, communication, vision or cognition.

One specific type of AT is Augmentative and alternative communication (AAC) [56]. These systems are designed to help those who have speech difficulties, facilitating more effective communication and social interactions [57]. These systems can be unaided, such as sign language and gestures, or aided by devices, such as tablets. Input methods can vary and may include options such as a mouse, keyboard or joystick. Moreover, these systems often offer multiple input modalities to minimise user fatigue and allow users to select the most effective method for their particular situation. However, it is worth noting that AAC communication is generally slower than typical speech [33,58], could be counterintuitive and requires time to learn.

Voice input is another modality that warrants special consideration, especially for individuals with dysarthria. Despite its potential, the use of vocalisation as an input to AT is relatively rare [59]. Given the slowness and less reliable performance of AAC systems, many users prefer their own speech because it may increase communication effectiveness [33] and reduce effort and time [60, 61].

In the section that follows, we discuss VUIs, including their general applications and specific relevance for individuals with dysarthria.

2.5 Voice user interfaces

VUIs are speech-based interfacing substituting keyboard to interact with electronic devices [62]. In other words, they are the interfaces through which a user interacts with a speech-based application, system or device using spoken language and receives a feedback or reply. Given that the interaction uses voice only, this eliminates the need to be able to use complex computing devices [63, 64]. The feedback received could be of an audio, other modality or an action performed. VUIs have been studied for decades; however, recent improvements in automatic speech recognition system (ASR) technology have made VUIs a common feature in current computing systems. VUIs are increasingly being integrated into various devices, including mobile phones (e.g. Siri on Apple devices), computers (e.g. Microsoft Cortana), vehicles (e.g. Apple CarPlay) or standalone devices such as SVAs (e.g. Amazon Echo and Google Home) [65]. In addition, VUIs are being integrated in household devices, such as televisions, ovens and washing machines.

VUI devices use ASRs , which receive, understand and process audio commands [66]. First, ASR systems detect the user's utterance then extract features from the input signal, which is converting the input to a measurable characteristics, and identify the words. Then, a meaning of the word sequence should be detected. Finally, the system produces a response or action. ASR systems can be classified according to the acoustic input or the type of speaker [2]. In terms of acoustic input categories within ASR systems, there are three primary types [67]. The first is a 'single word', in which the system receives one word at a time. The second is 'sentence', in which commands consist of multiple words separated by pauses. Finally, there is 'continuous speech', in which words are spoken in a connected manner without noticeable pauses.

The other classification is according to speaker type [68]. The first are speaker-independent systems, which are designed to be used by the general public, such as commercial devices. These devices are trained on typical speech, so people who have difficult accents or speech difficulties may not be able to use it as efficiently as others [22, 69, 70].

However, speaker-dependent systems can be tailored to a specific individual. This approach requires significant resources and effort to train and maintain a separate system for each individual user [71]. In addition, speaker-dependent systems may not be as widely adopted as speaker-independent systems [72]. The third classification is speaker-adaptive systems, which are first trained on typical speech and then adjust to the specific speech patterns of the user [71]. These systems do not need a lot of training, but their performance may be poor at first until they have adapted to the user's speech and can improve over time .

One of the main devices featuring VUIs and ASR is SVAs, which enable users to interact with other VUIs and devices. A major advantage of using VUIs in general and SVAs in particular is hands-free and eye-free interaction. SVAs can be instructed to execute a variety of tasks, such as controlling a household device, adding an entry to a shopping list or requesting information, such as daily news [73]. SVAs can serve as an accessibility tool and be of great assistance to individuals who have disabilities, allowing them to control their surroundings without physically moving [74, 75]. For instance, someone who has a motor disability can use SVAs to control their thermostat or switch off lights without moving. In the same way, a person who has a visual impairment can listen to the news without needing additional technology. However, these devices may not yet be fully accessible to those who have other specific types of disability [74]. Interacting with SVAs is performed through verbal commands, such as words and phrases, which can cause two problems. First, people who have speech difficulties may face issues using SVAs [16]. This intensifies with an increase in case severity [18]. Second, these devices are not designed for their type of speech because it is difficult to collect recordings from individuals with dysarthria [16, 21]. Another concern for SVAs is security and privacy issues that make users hesitant to use them [76].

This leads to the next section, which focuses on people who have speech impairments, specifically, dysarthria, because people who have dysarthria are the target of our study.

The next chapter discusses the literature studying the interaction between people who have dysarthria and SVAs.

2.6 Voice technologies for people who have dysarthria

2.6.1 Automatic speech recognition for individuals with dysarthria

This section is based on the literature review paper I published [2]. However, the structure of the literature review in this thesis differs from that in the published paper. This adjustment allows for the seamless incorporation of the review's findings into the overall structure of the chapter. Paper Link: (`https://www.tandfonline.com/doi/abs/10.1080/10400435.2022.2061085`)

Individuals with dysarthria interact with voice technologies differently than those without speech disorders. The primary issue in this interaction is the performance of the systems, which tends to deteriorate for people who have dysarthria and worsens as the severity of the condition increases. Performance in ASR systems is commonly measured by various metrics, and word error rate (WER) is the most frequently used [77]. The WER is calculated as WER=(S+D+I)/N, where S represents the number of words substituted, D is the number of words deleted, I is the number of words inserted and N is (S + D + I + number of correct words). Given that this is the focus of the study, we conducted an exhaustive search in this specific area. This systematic review used the following search terms: (dysarthri* OR dysarthri* speech) AND (automatic speech recognition system* OR ASR OR virtual home assistant* OR voice-controlled digital assistant* OR conversational agent* OR commercial voice assistant* OR voice interface* OR personal assistant*). The search used the following databases: ACM Digital Library, Google Scholar, IEEE Xplore digital library, Springer Link and Science Direct. A backward and forward reference search was also performed: the backward search identified the references from certain studies and the forward search identified studies that cited a certain piece of research.

To identify studies, we used the following inclusion criteria: (i) studies published from 2011 to 2020 (Siri launched in 2011, making ASR mainstream and enabling people who have dysarthria to use ASR ubiquitously), (ii) studies evaluating interactions between people who have dysarthria and ASR systems or devices and (iii) studies using WER as the measurement criteria, the most commonly used metric for accuracy. Selecting studies that used the same metrics allowed the authors to compare one thing that was common across the studies. Our focus was on human–computer interactions (HCI) rather than on the disorder and therapeutic interventions; therefore, we excluded clinical and therapeutic studies and those that examined dysarthria in individuals who had language and cognitive impairments (e.g. aphasia and dementia) to eliminate factors that affected the interaction process. The preferred reporting items for systematic review and meta analysis (PRISMA) was followed [78]. One author conducted the screening process and another undertook the final review of the screening results. Out of 83 studies that were selected and assessed for eligibility, 32 were chosen to understand the characteristics of users' speech that affect their interactions with ASR systems, the types of acoustic input required to effectively interact with ASR systems, the ASR systems that have been evaluated in dysarthria studies and the evaluation of current ASR systems (see flowchart 2.1). During the thesis writing, recent publications from the period 2022 to 2023 were included.

In accordance with the systematic review we conducted [2], we summarised the WER results from various studies in Table 2.1. We found notable differences in performance when ASR systems were used by people who had dysarthria, despite advancements in ASR technologies. This is consistent with findings from [66, 79], which indicate that regardless of the systems used, study subjects or research methodologies, the results have not improved substantially.

Several studies have focused on understanding the factors affecting the accuracy of ASR systems [66, 80, 81]. One primary factor is speech intelligibility; lower intelli-

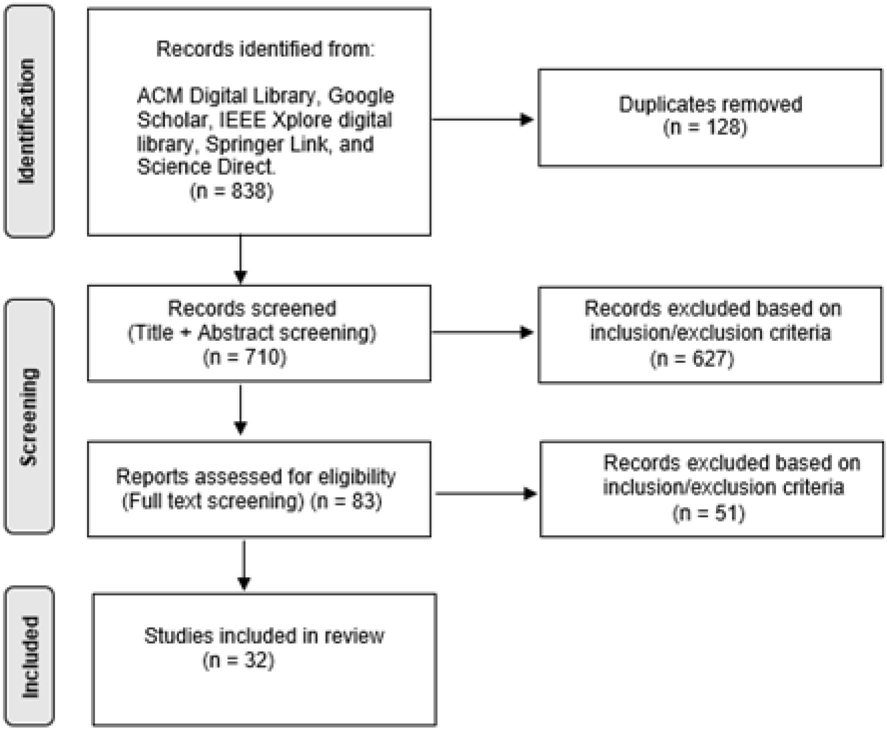

Identification
Records identified from:
ACM Digital Library, Google Scholar, IEEE Xplore digital library, Springer Link, and Science Direct.
(n = 838)
Duplicates removed
(n = 128)
Screening
Records screened
(Title + Abstract screening)
(n = 710)
Records excluded based on inclusion/exclusion criteria
(n = 627)
Reports assessed for eligibility
(Full text screening) (n = 83)
Records excluded based on inclusion/exclusion criteria
(n = 51)
Included
Studies included in review
(n = 32)

Figure 2.1: PRISMA-P 2015 flowchart

gibility correlates with lower accuracy. Table 2.2 presents the accuracy according to dysarthria intelligibility levels. Another factor closely related to intelligibility is severity; more severe cases typically have poorer intelligibility. The results according to severity are presented in Table 2.3. In addition, muscle fatigue influences interactions with ASR systems, reducing speech quality and increasing speech variability, thereby leading to lower recognition accuracy [80].

Other factors not directly related to the characteristics of dysarthric speech have also been found. One such factor is the speech mode, specifically, whether it involves isolated words or continuous speech. Many studies have found that isolated words result in better accuracy [66, 86] because it is easier for individuals with dysarthria to utter single words and it causes less fatigue. Another factor is the type of ASR

Study	Database	Input modality	Word error rate (%)
Geng et al. (2021)	UA-Speech	Isolated words	25.6
Jin et al. (2021)	UA-Speech	Isolated words	25.8
M. Kim et al. (2017)	Recordings	Isolated words	28.0
Liu et al. (2019)	CUDYS	Isolated words	28.2
	UA-Speech	Isolated words	31.0
M. Kim et al. (2016)	Korean	Isolated words	33.4
Marini et al. (2021)	IDEA	Isolated words	14.99
Hermann & Magimai-Doss (2020)	Recordings	Sentences	25.9
		Sentences	53.7
		Isolated words	42.9
Rudzicz (2012)	TORGO and MOCHA	Sentences	34.7
Turrisi et al. (2021)	Easycall	Sentences	61.9
oward a lightweight ASR	Recordings	Isolated word	15.6
Accurate synthesis of dysarthric			39.2
Benefits of pretrained	Copas	Sentences	40.73

Table 2.1: General accuracy of automatic speech recognition systems (Some papers were not included. They did not provide precise numbers. For example, they presented the word error rate in charts).

system used. Speaker-independent systems, designed for the general public, perform worse than speaker-dependent or speaker-adaptive systems [4]. However, obtaining training data for speaker-dependent systems is challenging, especially for people who have dysarthria, given the fatigue and frustration associated with recording a large number of samples [21]. Speaker-adaptive systems may initially offer lower accuracy but improve over time as they adapt to the user's speech.

Various methods were employed to address the previously mentioned interaction problems. To resolve the issue of variable dysarthric speech and enhance performance, prior studies have presented various models capable of modelling diverse phonetic variations [23–26]. Other approaches have aimed to resolve the problem of limited dysarthric speech data sets by either (1) using models that require less training data [27],

	Intelligibility			
Study	High	Medium	Low	Very low
[82]	7.91	16.80	27.16	59.83
[83]	7.75	16.50	27.37	61.42
[84]	7.55	16.47	26.84	62.37
[85]	5.81	16.38	4.47	50.36
[26]	16.17	31.41	45.79	83.36

Table 2.2: Accuracy (word error rate) of automatic speech recognition systems according to dysarthria intelligibility.

Study	Input modality	Dysarthria severity			
		Severe	Severe-moderate	Moderate	Mild
Xiong et al. (2019)	Isolated words	67.83	27.55	26.41	9.71
Joy & Umesh (2018)	All modalities	46.52	46.97	56.26	25.94
Yue et al. (2020)	Isolated words and sentences	57.6	–	33.0	14.3
Sriranjani et al. (2015)	Isolated words and sentences	43.61	–	32.63	21.14
Vachhani et al. (2018)	Isolated words	69.3	36.45	21.32	1.35
M. Kim et al. (2017)	Isolated words	–	–	28	28
Seong et al. (2016)	Sentences	–	–	46.00	29.11
Benefits	Sentences	51.54	–	31.26	39.4
Accurate synthesis of		50.1	–	36.8	12.6

Table 2.3: Accuracy (word error rate) of automatic speech recognition systems according to dysarthria severity (Some papers were not included. They did not provide precise numbers. For example, they presented the word error rate in charts).

(2) employing data augmentation as a source for new recordings, meaning artificially producing dysarthric speech [20, 28, 83–85, 87, 88] or (3) adapting data to particular speakers [82, 89]. Finally, 'data pooling' was used, that is, a group of normal speaker recordings were pooled from detests and combined with a dysarthric speech data set [90].

Another attempt to improve ASR performance is trying to create new data sets for dysarthric speech, which will enable researchers to train and test their models [91–96]. Examples include, TORGO database [42], UA-Speech [97], Nemours [98] and others. The latest effort is Google's Euphonia corpus [96], which aims to collect recordings from various speech disorders, including but not limited to dysarthria. However, this corpus is not open to the public. Another attempt for improving ASR performance is adding features to systems, for example, word prediction [99] or adding interfaces that help to formulate commands and reduce fatigue and frustration, which the authors discovered was a different way to interact because the system's structure was unable to comprehend incomplete commands [100]. However, adding interfaces may result in other issues for people who have dysarthria because it is often accompanied by physical disability [21]. For example, using a mouse or keyboard would not be easy for individuals with dysarthria. To reduce fatigue, some researchers have suggested using smaller vocabularies, which has been found to increase ASR accuracy [27].

2.6.2 Smart voice assistants for people who have dysarthria

SVAs, such as Google Home and Alexa, are primarily speaker-independent systems that are widely available for purchase. They can also be integrated into devices such as smartphones through Siri or computers via Cortana. However, prior studies have indicated that these commercial devices require further improvements to better accommodate users who have dysarthria [18, 22, 79, 101]. One primary challenge lies in the devices' accuracy, specifically, in understanding commands issued using dysarthric speech. Variability in volume and pitch adds another layer of complexity [22, 100]. For example, fluctuations in volume and pitch within a single word or sentence can confuse these systems, making it difficult to capture the intended command accurately [102]. Moreover, these devices may time out before the user finishes speaking [22, 100]. Although Alexa offers a follow-up mode that keeps the device listening for a few extra seconds, this feature has not been empirically tested on users who have dysarthria to the best of our knowledge. Another issue arises as a result of the unique characteristics of dysarthric speech, which often includes breaths between syllables. Current devices struggle to handle this as a form of input [79]. However, some studies, such as the one by Ballati [22], have suggested that these devices may be capable of understanding moderate dysarthria to some extent. Navigation presents another challenge. If the device does understand the initial command, it might issue a verbal prompt requesting additional information. This would require the user to speak another command, leading once again to potential errors and misunderstandings [22, 100].

In commercial devices, the WER for typical speech is 9%, according to kepuska [103], but it is significantly worse for those who have dysarthria. Multiple studies, including those by Ballati [22, 101] and De Russis and Corno [18], have assessed how these systems perform using recordings of speech from individuals suffering from dysarthria. The findings are presented in Table 2.4. Siri, which continuously attempts to transcribe every received command, scored worse than other platforms. However, some systems give feedback when they cannot comprehend a command, allowing the user to rephrase it [101]. These commercial devices do not rely solely on the transcribed text; they also consider the context for better understanding [22]. Consequently, WER is not the only metric used by researchers to measure system efficiency. They also qualitatively analyse the system's understanding of commands and its subsequent responses. Evaluating voice-controlled systems thus requires a multifaceted approach. Table 2.4 indicates that these systems generally underperform for users who have dysarthria, possibly because test data are not tailored for this specific type of ASR system.

Individuals with dysarthria can use assistive devices as an alternative. This ensures better access to SVAs; however, the interaction will be slower and require more steps to perform the desired action because the user needs to type the command instead of uttering it. This means that users who have dysarthria do not have the same experience as healthy users. Thus, the challenge is not solely about accuracy; rather, it represents a broader issue. These devices are not accessible or usable for this specific group of

Study	Dysarthria severity	Automatic speech recognition system				
		Google	Siri	IBM	Microsoft	Sphinx
De Russis & Corno (2019)	Severe	78.21	–	89.08	78.59	–
Ballati et al. (2018a)	Moderate	24.88	70.89	–	39.39	–
Ballati et al. (2018b)	Various	15.38	69.41	–	–	–
Moore et al. (2018)	Various	43	–	–	–	126

Table 2.4: Accuracy (word error rate) of commercial automatic speech recognition systems (Some papers were not included. They did not provide precise numbers. For example, they presented the word error rate in charts).

users [74]. Moreover, the experience for users who have dysarthria is not similar to that of nondisabled users. Existing guidelines, such as the Americans with Disabilities Act, the Web Content Accessibility Guidelines or the United Kingdom's Equality Act, may offer some framework for accessibility, but implementation in SVAs has been inconsistent. This results in a system that is not just technically inaccessible but also socially exclusive.

2.6.3 Examples

There are various projects aimed at assisting people who have dysarthria, each employing different approaches. One approach is training based. A key example is the STARDUST project [104], which is command only. Another significant initiative is the voice-input voice-output communication aid, which focuses on converting dysarthric speech into synthesised speech. CanSpeak [105] is similar to STARDUST in that it features a small vocabulary. In this project, users give computer commands using a predefined list of keywords tailored to their preferences. Likewise, Kim [99] designed a mobile voice interface that presents users with a list of predicted sentences or words in accordance with the initial letter of their utterance. Users can then select the desired word or sentence using a confirmation button, simplifying the interaction process. A later example is Mpass, software that helps users to train and generate a model according to their speech and preferred words.

There are also projects that focus on home automation. An example is the CloudCast [106] application in which researchers tried to create a low-cost voice-controlled system for home automation using Open Home Automation Bus open-source software. Another project is ALADIN by [100, 107]. This project is a VUI assistive device that involves users training the system by themselves so they can use any vocabulary.

Unlike ASR systems, there is a notable gap in the literature focused on SVAs [74]. This section incorporates findings on various voice assistants, including Siri. Ballati's study [101] emphasised the need for enhanced accessibility in voice assistants. The research found that these systems could identify 47% to 59% of dysarthric speech after testing three major voice assistants, Siri, Amazon Alexa and Google Assistant. The effectiveness of these platforms is not solely determined by their ability to transcribe speech accurately but also influenced by the context of the spoken words [22]. A single correctly identified word could substantially improve the system's response. Russis' findings [18] align with Ballati's but also revealed a level of transcription accuracy so low that it captured less than one correct sentence out of 51 for all tested platforms. Furthermore, the WER exceeded 78%, indicating a significant limitation in these systems. Although these studies did not account for the impact of contextual clues, they do reveal performance constraints. Another work by Ballati [22] investigated the performance of Siri, Google Assistant and Cortana for Italian speakers and found that user-specific variations affect systems' behaviour. One limitation of existing studies is the use of databases such as TORGO, which was not initially intended for virtual assistant systems, which potentially affects the reliability of the results. However, some research efforts have generated data sets specifically designed for these systems.

One work by Masina [74] involved users to understand their interactions with SVAs. In this study, the focus was on people who had cognitive or linguistic functions, which differs from our target group. However, their results align with prior studies on dysarthria. They showed that there is an issue concerning the timing of users uttering commands and the system timing out.

It is clear that the work in the literature focuses on improving the accuracy of the models for SVAs or understanding their performance. To the best of our knowledge, there is no study that has suggested a solution to directly interact with SVAs for people who have dysarthria. Moreover, the existing literature does not involve users in the evaluation and assessment process to understand their needs.

2.7 Nonverbal voice cue interactions

Nonverbal voice cue interaction refers to interacting with VUI systems or devices using nonverbal voice cues. In other words, this is using any sounds that a person can utter but not words or sentences. As the literature has found, and as discussed in earlier sections, the shorter the commands, the easier it is for an individual who has dysarthria to utter them. Furthermore, a more accurate ASR system could improve understanding. AAC users prefer to use their own speech because of poor performance and slow response times, as reported by Ferrier [33]. Furthermore, the approach of using vocalisation specially for people who have dysarthria is scarce [59]. To address these issues and accommodate a range of users who have diverse backgrounds, including those who have technical limitations and various types of dysarthria, it is recommended to design interfaces that are simple and user friendly [63, 64].

Therefore, we explored the literature on using nonverbal voice cues as a method of interaction. Nonverbal voice cues are any sounds that users can utter that are neither words nor sentences. Beyond the reasons mentioned earlier, nonverbal voice cues offer the advantage of being language independent. These cues can be classified into two categories: continuous input and discrete input. In the case of continuous input, users continue to utter the sound, receiving feedback while still vocalising. This method is particularly useful in gaming [108–110], interactive art drawing [111] and mouse simulation [112, 113]. However, discrete input occurs when the user finishes uttering the sound and the system responds according to that specific input. This method is beneficial when articulating longer utterances is challenging.

The concept of using nonverbal voice cues was first introduced by Igarashi [114]. Since then, various works have implemented this approach. Nonverbal voice interaction has been employed in various contexts, such as art, gaming [111], mapping [115] and accessibility [109]. Several types of nonverbal voice interaction have been used, including vowel sounds, humming, hissing, whistling and blowing.

Regrading continuous speech input, Igradshi [114] employed continuous nonverbal voice input to control various interactive applications. For example, Igradshi used the utterance 'volume up, aaaaah' to prompt the system to increase the volume continuously until the user ceased vocalising. Similarly, Mihara [116] employed this approach for cursor control. The cursor continues to move as long as the user maintains the sound. Other continuous sounds are vowel sounds, which have been used in voice-drawing applications in which they controlled the drawing brush and cursor [117]. Humming has also been used, and it was found that it was faster than speech commands in game control and served as an alternative keyboard input method [108]. Whistling and blowing have been used to control a mouse pointer through a microphone [112]. However, discrete input has not been as widely adopted as continuous input [118]. It has been used mainly for actions such as mouse clicks [117].

It is worth noting that the study described above did not focus mainly on accessibility research and did not specifically consider enabling people who have speech impairments. However, a recent study has proposed using nonverbal voice cues for interacting with mobile phones to assist people who have disabilities [119]. In this study, 15 nonverbal sounds were suggested and a model was trained using recordings from healthy individuals. During the evaluation phase, some sounds were removed because they were not suitable for various types of speech impairment. However, this study indicates that this area is worth investigating.

2.8 User involvement in system design

To improve our design quality and project outcomes, the involvement of humans, which includes the user and any individuals who contribute to the design of systems, is crucial [120]. Such involvement in the design, testing or evaluation of systems, products or projects provides rich information that can significantly benefit the project. From the literature, we found that few studies have involved users in the design process of systems for people who have dysarthria and ASR systems.

One example of a thorough user-centred approach is the work by Gemmeke [107]. His research focused on understanding user needs through various techniques, including interviews, storyboards, creating personas and others. Another study by [105] also involved users through using a participatory design approach in which the users were part of the design team. It involved users in two phases. The first phase included system customisation, in which users could choose the number of words they could utter to be used as commands. The second phase involved codesigning, in which the users made design decisions in addition to the research team. Users can also be involved in the testing phase to obtain feedback on proposed systems and designs [19, 121].

When focusing on SVAs specifically, researchers have relied on recordings from users to understand device performance or have used available data sets for evaluation [18, 22, 101]. Such approaches may have been used because of difficulties with recruiting people who have dysarthria. For example, M. Kim et al. [23] were unable to recruit people who had severe dysarthria and thus focused their study on mild and moderate cases only. A limitation of this approach is that these data sets were not designed to test SVAs and are not in the form of commands. Furthermore, we found no study that specifically tested SVAs for people who have dysarthria. One study evaluated the feasibility of using SVAs for people who have motor impairments and mild speech issues, without specifying their particular speech impairment. Therefore, we concluded that there was a need to evaluate user interactions with SVAs, understand their challenges and needs and identify feasible solutions.

2.9 Conclusion

From this literature review, we identified that SVAs are limited in accommodating the needs of individuals with dysarthria. The main reason for this is the variability in dysarthric speech, in addition to the difficulty in collecting enough recordings to train machine learning models. Moreover, the literature highlighted the significance of enabling these individuals to gain independence in everyday tasks and have better communication abilities and more safety. It is also clear that the alternative communication methods currently available are slower than direct voice interaction, which could lead to frustration. Therefore, people who have dysarthria prefer to use their voices to the maximum extent possible. Thus, individuals with dysarthria prefer voice modality.

Furthermore, we discovered that user involvement in SVA research is minimal. A possible explanation is the difficulty in recruiting individuals with dysarthria and the difficulty for them to give recording, as is discussed in Chapter 6, leading to a lack of solutions that accommodate their specific needs. It became evident that shorter voice commands are easier for users who have dysarthria. This realisation opened the door to exploring nonverbal voice cues as a means of establishing more directed interaction methods for individuals with dysarthria, an area that the existing literature has largely overlooked.

Our work builds upon these efforts, focusing specifically on dysarthria and actively involving users in the development process. We aimed to create an accessible solution that allows individuals with dysarthria to use the system independently. Moreover, the interface was designed to align with users' capabilities, ensuring that they do not experience fatigue or discomfort while using it. The objective was to provide users who have dysarthria with an experience similar to that of individuals without speech difficulties, thus eliminating the need for intermediary devices to interact with SVAs. Ultimately, users should be able to use their voice as their primary modality for interaction, just like individuals without speech impairments.

The proposed interaction approach uses nonverbal voice cues, satisfying all the above requirements. This method is well within users' capabilities: the use of short, nonverbal voice cues is less likely to cause fatigue than uttering full words or sentences. The interaction is direct: users can communicate with the SVAs using these nonverbal cues without the need for additional steps or intermediary devices. Furthermore, using voice as the sole modality for interaction is parallel to how individuals without disabilities typically interact with these devices, ensuring the approach is accessible.

The intention of this chapter was to highlight the issues and limitations found in the literature, which led to the following contribution:

C1 A systematic literature review that presents the limitations and gaps in the area of people who have dysarthria and automatic speech recognition systems and smart voice assistants.

Chapter 3

Methodology

The preceding chapter highlighted the low accuracy of SVAs when used by individuals with dysarthria. It also explained the difficulties faced by individuals with dysarthria when interacting with SVAs and how the capabilities of people who have dysarthria do not match the requirements of the SVAs. Moreover, it covered the prevalent lack of user involvement in research related to SVAs. These issues lead to inadequate results and solutions that do not meet the needs and preferences of individuals with dysarthria. Further, they limit user experience and hinder the full utilisation of these technologies.

In light of the issues highlighted, the primary aim was to develop solutions more attuned to user needs and contexts. To achieve this, users' input, feedback, preferences and recommendations were incorporated right from the planning phase to implementation and evaluation. This approach entailed the integration of various research methods.

To accomplish this, the present work employed a user-centred design approach, specifically, a method in HCI that actively includes users and focuses on understanding and addressing their needs. Furthermore, other contributors were involved to allow more involvement of in the design process, thereby fostering the creation of solutions that are well suited and responsive to user needs.

3.1 User-centred design and mixed methods approach

3.1.1 User-centred design

In the 1980s, the term ‘user-centred design’ was introduced by Donald Norman [122], whose philosophy centred on making the user the focal point of the design process. Two original publications introduced this concept and described the key principles [122, 123]. According to Gould [123], the principles focus on early consideration of users and tasks. First, the designer should fully understand the users for whom they are designing and the task. Second, measurements should be used to evaluate the performance of the early prototype and analyse users’ responses and feedback. The third principle emphasises that an iterative approach should be followed, which means that there should be an iterative cycle of design, evaluate and redesign. Over the years, several researchers and designers have introduced guidelines on implementing user-centred design, all aimed at incorporating the end user into the design process [124]. An example is the work by Maguire [125], who described a complete process and methods for every stage of the design process. It was not until the International Standards Organization (ISO), in 2010, expanded the scope of user-centred design under the standard 9241–210 [126], and the latest version is ISO 9241–210:2019, emphasising the importance of addressing the impact of various stakeholders or individuals related to the project at hand in the design process and not only the end users. User-centred design involves a variety of methods, ranging from observations and interviews to prototyping, surveys and others [127].

In the research outlined in this thesis, our goals were closely aligned with the foundational objectives of user-centred design, which seeks ‘to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques. Usable systems can provide a number of benefits, including improved productivity, and increased accessibility.’ To achieve our objectives, we adhered to the user-centred design process as outlined by ISO

9241–210 [126]. This process encompasses the following steps: (a) planning the process, (b) understanding the context of use, (c) specifying user requirements, (d) design and (e) evaluation.

3.1.2 Integrating user-centred design with mixed methods

Given that this study aimed to understand user experience and system performance, we used a mixed-method approach: qualitative and quantitative methods.

In HCI, qualitative methods are important for understanding the user and the area of research [128, 129]. Within some of the stages in this study, we used a qualitative approach aimed at obtaining an in-depth understanding of users' experiences when interacting with SVAs and the underlying reasons around the issues they face. Specifically, we conducted interviews because of their flexibility, ability to understand individual perspectives and interpretations and sensitivity to various expressions [130].

However, in other stages, we employed a quantitative approach focused on numerical data, logic and outcomes, this will give us information about system performance and users' performance as well. Thus, by utilising mixed methods, incorporating qualitative and quantitative techniques, we obtained more comprehensive answers to our research questions by understanding the relationship between the qualitative and quantitative results.

The following are the five steps of the user-centred design process, according to the ISO [126]. Within this process, we integrated user-centred design and mixed qualitative and quantitative methods. The processes were iterative; after the end of the cycle, we revisited previous processes to edit and refine the system.

1. Planning phase: During this phase, we formulated the research questions to guide the study, using an extensive literature review undertaken to explore the current state of research and identify gaps, as detailed in Chapter 2.

2. Context understanding: To attain a more comprehensive understanding of the challenges faced by users and how they currently use SVAs, a qualitative approach was used, specifically, interviews, because this is a common method in HCI [131]. We interviewed individuals who had dysarthria to grasp their challenges and discover how they use SVAs. This aspect is elaborated on in Chapters 2 and 4.

3. User requirements elicitation: In terms of eliciting user requirements, we also chose a qualitative approach. Interviews were held with individuals who had dysarthria to inform design decisions, as outlined in Chapter 4. In addition, an expert in communication sciences and disorders was involved who became part of the research team to enhance the collection of these requirements.

4. Design stage: The design stage was initiated according to the insights gathered during the requirements elicitation phase and using the involvement of the expert. A pilot study was conducted to test the first prototype. In line with our iterative approach, following the findings and feedback from the pilot study, the system was subsequently redesigned. More details about the design process and the considerations are in Chapters 5 and 6.

5. Evaluation phase: Finally, the evaluation phase employed qualitative and quantitative methods. Detailed in Chapters 7, 8 and 9, this stage involved field testing the system on actual users, followed by various questionnaires and post-study interviews. The questionnaires measured various aspects of the interaction. One measured usability (system usability scale [SUS] questionnaire [132, 133]), another measured user experience (subjective assessment of speech system interfaces [SASSI] questionnaire [134]) and the final one measured the workload required to interact with the system (NASA task load index [NASA-TLX] questionnaire [135]).

3.2 Methodological considerations

This section highlights the considerations we implemented to have an accessible research design. This means that the study was designed to remove the barriers to participants who have disabilities participating in the study [136]. This consideration was through all stages of the research process. In our work, we implemented specialised methodological considerations to accommodate the unique needs and challenges faced by individuals with dysarthria. The methodologies were chosen to ensure ethical rigour, participant comfort and accessibility and reliability of the collected data. One of the key factors considered was the nature of the disability under study, specifically, speech impairment because of dysarthria.

Starting with the planning and context understanding phases, an option to understand user experience is focus groups or workshops. However, in these settings, participants who have a speech impairment might struggle to be heard or might not have the time or capability to fully express their needs [136–138]. Such limitations could result in social pressure, which might further inhibit participants from expressing themselves freely [138]. Finally, some individuals might dominate the conversation in a group setting [136], limiting the representation and data collection for participants who have more severe impairments. To overcome this limitation, we decided to use individual interviews. Focusing on a single participant during each interview allowed us to collect richer, more nuanced data without causing undue stress or frustration to the participant.

Another barrier was accessibility, which refers to any obstacles that might prevent someone from accessing information or participating fully in an activity because of their unique challenges or disabilities. This could be, for example, the set-up of the study or materials being in a format that does not suit participants. Recognising that some participants might have challenges communicating on their own, we followed suggestions from prior studies [130, 139–141] during the interviews. We offered multiple communication modalities to accommodate the participants' needs. The participants could use speech if possible, typing or text-to-speech assistive devices. Moreover, the researcher was familiarised with each participant's preferred communication method prior to the interview. Given that it has been reported that users who have speech difficulties can communicate more effectively and experience less fatigue when supported by a partner [141], we allowed partners and caregivers to assist participants during interviews to ensure comprehensive communication.

To further address varied preferences and capacities, when administering the questionnaires, we made them available in digital and paper formats, providing participants the freedom to select the most convenient option for them.

By carefully considering these factors, we aimed to cultivate a research environment that was effective and considerate of our participants' unique challenges, ensuring that the research process was inclusive and respectful and yielded valid and reliable findings that accurately reflected the experiences and needs of individuals with dysarthria. By aligning our methodological choices with the research objectives and ethical standards recommended by Cardiff University's ethical review board, we aimed to contribute meaningful and trustworthy insights to the existing body of knowledge. Further details about these considerations are elaborated on in the chapters dedicated to each study.

3.3 Conclusion

This chapter outlined the methodology used in this study, which aimed to create alternative solutions allowing individuals with dysarthria to interact more effectively with SVAs. A mixed-method approach was used to gain deep insights into various aspects of the study. Given a noticeable lack of user involvement in prior studies, user-centred design approache was adopted to ensure comprehensive user participation in our study. We detailed the methods used in each of the four steps of the user-centred design process. The explanation of each phase in this chapter lays the groundwork for subsequent chapters, which provide more detailed insights. This chapter also presented a series of methodological considerations to ensure the reliability and validity of the data collected, addressing each aspect of the study with thorough consideration of the ethical, practical and scientific principles.

Chapter 4

Understanding User Experiences and Design Requirements

The primary objective of this study was to explore the challenges encountered by individuals with dysarthria when using SVAs and to obtain feedback on the proposed interaction modality, which involves nonverbal voice cues. To do this, we followed the user-centred design process, specifically, the stages of (b) understanding the context of use and (c) specifying user requirements. To address this objective, the subsequent section details the first study, which was designed to answer the research question (RQ1):

RQ1: How do individuals with dysarthria currently use smart voice assistants, and what are their present experiences with these devices?

Our approach involved conducting interviews with 19 participants who had dysarthria and analysing the interview data, which enriched our understanding of the specific issues faced and the areas that demand attention. This investigation led to the identification of key design components essential for developing voice assistants that cater to those who have dysarthria. Feedback and the requirements for a proposed nonverbal voice cue interaction system were collected from these interviews. The insights from this initial phase steered the direction for the subsequent phase, which focused on the design of a system aligning with user needs, as elaborated in Chapter 4.

The work in this chapter has been previously published in the Universal Access in the Information Society Journal [3].

4.1 Method

Semi-structured interviews was the main research method. The interviews covered various aspects. We focused on understanding how the participants use technology for assistance in their daily tasks. In addition, we asked them about their experiences using voice assistants in general or SVAs in particular. We also asked them about their views on how systems should be designed to use nonverbal voice cues and account for their enunciation capabilities and preferences. We adapted the interview questions from [99] and [142]. These studies targeted participants who had speech difficulties, and their general goal intersects with ours. Each of the studies examined the problem from a different angle. One focused on user experience [99] and the other [142] on system design for people who have dysarthria.

Ethical approval was obtained from the ethics committee at Cardiff University School of Computer Science and Informatics. There were no incentives provided for participation to ensure unbiased responses and maintain the integrity of the data collected. The interviews were 45 minutes long to allow for in-depth discussion without causing fatigue to the participants.

4.1.1 Participants

To recruit participants, announcements were made through various channels. Given that social media has been identified as an effective tool to overcome the challenges researchers might face during recruitment [143], we posted calls for participation on various social media platforms, including X (formerly known as Twitter), Facebook and Instagram, specifically, on dysarthria support groups' and charities' pages. We also reached out to various charity organisations and posted on the university's announcement platform. Furthermore, numerous speech and language therapists were contacted and asked to share the announcement with their patients.

The inclusion criterion for participants was that they should be adults who have dysarthria to the extent that it affects their speech yet still possess the ability to produce sounds. We focused on individuals whose speech was affected because we aimed to study those who have begun experiencing challenges when interacting with speech technologies. This would enable us to inquire about their specific issues and experiences. However, we excluded participants who had extremely severe cases to the point that they could not produce any sounds. This was because our target area did not encompass such cases; our work focused primarily on using nonverbal voice cues for interaction.

We recruited 19 adults: 10 males and nine females. Their ages ranged from 42 to 65+ years (see Table 4.1). The severity of the cases (see Table 4.2) and participants' speech capabilities varied so that our system and requirements could facilitate the entire spectrum of people who have dysarthria. All participants are native English speakers. In the section that follows, we describe how we accounted for this disparity and range.

Age group (years)	Number of participants
25–44	2
45–64	7
65+	10

Table 4.1: Participants' age range

Dysarthria severity	Number of participants
Mild	9
Moderate	4
Severe	4
Unknown $*$	2

Table 4.2: Dysarthria severity among participants (* Patients did not officially know the severity of their cases.

4.1.2 Procedure

Prior to undertaking the interviews, informed consent was obtained from all participants. Then, the participants completed a demographic survey in which they were asked to indicate their age, location, severity of dysarthria and preference regarding the method of communication or assistive device to be used in the interview.

We conducted the interviews during the first half of 2021. The timing is important because during this year much of the world was weathering the coronavirus disease (COVID-19) pandemic. Given that several countries, including many in which we conducted interviews, were under lockdown for varying periods, we conducted all interviews online.

One of the considerations during the interviews was that we gave participants as much time as they needed to answer, ensuring that they did not feel rushed or pressured. At the beginning of the interviews, we informed participants that the researcher's time was flexible. We emphasised that if they needed a break or wished to continue the interview at another time, it was entirely acceptable. Given that all the interviews were conducted online and participant speech capabilities varied, each interviewee communicated their answers using different methods. Some relied on their own speech to communicate whereas others, when unable to speak clearly or were tired of speaking, typed in the video call program's chat function or shared their screen to reveal their answers typed into a Microsoft Word document. Another group used an assistive device (e.g. a text-to-speech device) to communicate.

Several participants had another person to assist them in communicating. Considering that the person helping them could be biased from their own experiences, we wanted to clearly distinguish between the subjective feedback of the assistant and that of the participant. Consequently, we focused our questions and framed them to relate exclusively to the participants' experiences. For instance, rather than posing generalised queries, we opted for more personalised prompts, such as 'How did you find using smart voice assistants?' or 'What would you wish for ... ?'. This approach reinforced the primacy of the participants' perspectives and not the partners'. In addition, we made efforts to periodically reaffirm to participants and their assistants the core objective of our study: to understand the unadulterated experiences of the participants. By consistently emphasising this aim, we sought to encourage assistants to remain as neutral as possible, allowing the true voice of the participants to come through.

Given that the interviews were semi-structured, all participants were asked to answer a set of core questions and then, depending on their responses, follow-up questions. The interview questions can be found in (Appendix A). The first section was about their experience with dysarthria, including the history of their case, their speech style and the associated effects on their lives at home. The second section focused on how they coped with dysarthria. They were asked about the technologies or assistive devices they use and their experience using these. They were also asked whether they were using SVAs or any voice interface device or service. The last section concerned the proposed system. In this interview, our questions were not limited to SVAs but covered any voice technology. This was done to ensure broader data coverage. This approach not only allowed us to gather more comprehensive data but also offered insights into a wider range of user experiences. During the interviews, the interviewer explained the proposed system concept and its functioning. Then, the users were asked about their feedback on the proposed system, voice cues that they would find convenient and the system design they would prefer.

4.2 Data analysis

Given that the interviews were exploratory, we conducted a thematic analysis. Thematic analysis is a useful approach when conducting user-centred design research [144]. It helped in identifying themes and patterns in the interviews that would inform the design of the system. The thematic analysis was conducted using an open coding approach, which was driven by the transcripts (bottom-up approach), utilising the guidelines in [144, 145]. First, all interviews were transcribed verbatim. NVivo 12 was used to assist with the coding process. The process began by transcribing a few recorded interview files to generate the code logbook, and then moved on to transcribing all the files. The coding process was repeated several times to identify similar codes and refine these until a final set of codes was reached. The next step was to ensure the credibility of the coding process. When collecting qualitative data, the coding process needs to be validated before being accepted. A common method for this validation is using intercoder reliability. Intercoder reliability assesses the reliability of the coding process by employing a numerical measure to gauge the agreement between different coders on the data coded. There are several statistical measures available to determine intercoder reliability. In our work, we used Krippendorff's alpha, which has recently gained popularity among researchers because of its flexibility, which allows for the inclusion of more than two coders [146]. The result scale for intercoder reliability ranges between 1 and 1. Values between 0.61 and 0.80 indicate substantial agreement and those between 0.81 and 1.00 are almost perfect [147].

The software used to perform the intercoder reliability was Atlas.ti. The steps followed to assess the reliability after the data were initially coded and a codebook generated were as follows:

- First, I randomly selected approximately 25% of the data to be multiply coded [148].
- From the selected interviews, specific chunks of data were chosen for coders to process.
- Subsequently, two independent coders who had no prior involvement in the interview process were chosen. They both had a background in qualitative research. Each coder worked on the data separately. Both used the Atlas.ti software.
- After the coders completed their coding, I merged the two sets of codes and

performed an intercoder reliability agreement analysis. The Atlas.ti software offers built-in statistical features. Krippendorff's alpha was chosen because of its flexibility, should we decide to add more coders, and its ability to produce the most accurate possible results [149]. The result value was 0.86, which indicated the reliability of the coding.

It is worth noting that while this study employed thematic analysis based on Braun Clarke's approach [144] it diverged in the use of intercoder reliability. Thematic analysis, as described by Braun Clarke, does not incorporate intercoder reliability; instead, it focuses on the researcher's subjective interpretation of the data. The use of intercoder reliability is a somewhat controversial topic in the qualitative research community [148]. In this research, we opted to employ intercoder reliability to enhance the credibility of our findings. Future work should aim to adhere more closely to Braun Clarke's method or explicitly adopt a different analytical framework.

4.3 Results

In this section, we discuss the themes and the major points identified from the interviews with the study participants. The subheadings in this section are the themes we derived from our coding analysis. For each of these themes, we discuss the participants' responses and provide direct quotations from the interviews.

4.3.1 Tasks

Participants reported that they used voice interfaces in general and SVAs for different, specific tasks. These devices gave the participants some independence. One participant commented that using voice assistants

'would give my wife a break from me asking her to find that information for me'.

They mentioned various tasks, either those for which they were currently using the devices or those for which they would like to use one if their health condition deteriorated. The majority reported that they most often used the SVA device to play music. Similarly, they noted that they used their SVAs to acquire information, such as news, weather updates and football scores. Furthermore, they used SVAs for entertaining themselves. One participant stated that they often asked their SVA to tell a joke and another that they asked it to play the 'Alexa, I love you' song. Interestingly, few participants indicated that they used the device for communication, such as to send messages and make calls. They also gave examples of tasks for which they did not currently use a device but would like to have the opportunity to do so. Examples were to control smart devices and home appliances, such as lights, heating, alarm systems and curtains, and the volume of the television or music system.

4.3.2 SVA/VI experience

Participants shared their experiences using SVAs or other voice interfaces, such as Siri or smart devices. Our first finding is that the ways in which each participant interacted with the devices varied. Some participants who had mild dysarthria interacted with these directly through speech whereas others who had severe dysarthria used their mobile phone to type the command that they wanted to send to the device and then relied on their mobile phone to speak it. This would occur with the help of either a text-to-speech application or an application that relied on the user's stored voice, which is a recording made by the users when they were still able to utter phrases. These recordings are then transferred to synthesised speech. Another group used their AAC devices to speak on their behalf. These devices are tools that aid communication for people who have communication disorders. The device receives input through various devices, such as a mouse, keyboard and joystick, and converts it to speech.

The participants' experiences of interacting with the SVAs or other voice interfaces similarly varied, but the majority had negative experiences, regardless of the method with which they interacted. When the participants interacted by speaking, the device often failed to understand their speech. One interviewee said,

'I would rather type because nine times out of 10 you get the wrong answer.'

Their experiences with the voice interface failing to recognise their speech drove them to use a different method of interaction that sent their input directly. Another participant indicated facing the same problem when using the alarm feature, commenting,

'I find myself yelling at them a lot. My alarm clock, I can't turn it off anymore.'

Given that dysarthria can co-occur with a physical disability, using a regular alarm clock or phone alarm requires the use of hands, which the user might be unable to do. However, one participant observed that when the method of interaction was through a mobile phone, the process was complicated, requiring more steps to send their command to the device. Moreover, another participant commented,

'I suppose the only difficulty with ... all these voice banking systems is that the pronunciation that you get from your synthetic voice is not necessarily the word that you want to come out and therefore Alexa does not always respond correctly. So, you have to learn how to type certain words in to get the correct pronunciation.'

This comment raises several issues concerning the effects of system design in SVAs and other voice interfaces on people who have dysarthria.

Nevertheless, not every participant's experience was uniformly negative. Some were happy with their interactions with their devices. These individuals most often interacted through their mobile phones or an AAC device.

4.3.3 Proposed system design

Participants expressed a variety of perspectives about the potential of interacting with SVAs using nonverbal voice interaction. Overall, they responded positively to the idea.

One group was interested in nonverbal voice interaction as a means of decreasing their dependence on others for performing a task. Moreover, the technique would empower them to use other device features, especially if they only had limited use of their hands. Other participants compared it with other interaction methods, such as typing, believing that nonverbal voice interaction would be faster and more direct than typing. One participant noted that typing to interact with their SVA 'feels backward'.

Another group thought otherwise and disagreed that nonverbal voice interaction would be convenient. Some could not make any noise using their voice at all and thus believed that the proposed system would not suit them. Moreover, the dysarthria of some participants was severe enough to make it difficult to distinguish between even nonverbal voice cues. Others noted that given that they were physically able to perform all their tasks, they did not need to use their voices. One interviewee argued that using one's voice if one has dysarthria is fighting a losing battle. Another found typing to be faster and, because of the rapid deterioration of their condition, said that they preferred typing or using their eyes over using their voice.

The last group did not have a clear opinion. These participants had dysarthria that was still mild and they believed that nonverbal voice interaction might be helpful if their condition was to worsen.

Sounds

When asked about the sounds that they could make if they were able to make sounds at all, to be used as commands, participants were unable to give specific answers and found it difficult to decide on an exact sound. However, regardless of the severity level of their dysarthria, most of them found vowels to be easier, albeit softer, sounds to make than consonants. A participant who had severe dysarthria tried to enunciate the sound of the long form of the letter 'E' (e.g. 'eeee'), but their voice was breathy. The same participant found it challenging to enunciate combinations of vowels because their voice was similarly breathy when uttering all of them. Another vowel that was difficult for a participant who had a mild case of dysarthria to enunciate was ‘U’. The participant had difficulty pulling their tongue back. In terms of consonants, some participants were able to enunciate these; however, these required more effort and made the participants’ speech sound lazy or that they had not taken the time to pronounce the consonant properly. Participants noted that the most difficult consonant sounds to make were ‘G’, ‘H’, ‘S’, ‘D’, ‘K’, ‘F’, ‘Ch’, ‘T’, ‘P’, ‘B’ and ‘L’. Two participants observed that words that contained multiple consonants, such as the word ‘consonant’, were more difficult to pronounce than words that had only one consonant. For instance, a participant said that it was easy to pronounce ‘T’ and ‘L’ separately but it was much more difficult to pronounce the word ‘little’. In summary, the ability of each individual to pronounce letters and words varies and is affected by various factors, such as the level, cause and type of dysarthria. However, most of the participants found it easier to pronounce vowels.

List preference

When the participants were asked whether they preferred the system to have a predefined list of voice commands or to generate their own list, their opinions differed. Most respondents preferred to program their own voice commands. They argued that because every case is different, voice commands should be personalised and customised. One participant indicated that creating commands helped to make these easier to remember. However, a few participants preferred a predefined list. One individual stated that the choice was related to age or generation because this participant was used to ‘Plug and Play’ devices. Another commented that developers would know the sounds that worked best and therefore preferred a predefined list. Another participant alluded to the same idea but said it was preferable to have a predefined list initially and then move on to creating personalised commands. Many of the participants agreed that both options should be available to users. For instance, one noted using the predefined list to become comfortable with the system and then generating one’s own voice commands, and another suggested that developers offer guidelines to start or provide a list of voices that the user can choose from.

4.4 Discussion

The objectives of the interviews were to understand users' requirements; obtain feedback on the proposed interaction method, which involves nonverbal voice cues; and to determine the nonverbal voice cues that may be incorporated into the system. The interview results indicated that people who have dysarthria face challenges in using VIs and SVAs and that these devices are not fully accessible. Difficulty is experienced when interacting directly with the device and when interacting through an intermediary device. However, Ballati [22] indicated that people who have mild dysarthria could use these devices to a certain degree. This finding is consistent with the results of other studies that address the issue of interactions between people who have dysarthria and SVAs [150].

A person's quality of life is a major issue, and we found that these devices improve different aspects of the quality of life. For instance, they give users the ability to communicate. As per Light's definition of communication [52], the act of communicating involves expressing needs, exchanging information and engaging in social interaction. This was shown when the participants used the SVAs to tell a joke, call a partner or send text messages. This result supports that of prior studies, which have indicated that a person who has dysarthria could use music as a form of interaction with their children. Others who have dysarthria use music to express themselves and add a humorous effect when communicating with others [151]. Another aspect of improving quality of life is these devices giving individuals independence in performing various daily tasks [152].

The participants in our study reported that they performed various tasks using SVAs. However, a notable limitation in the tasks for which SVAs were used is the participant's ability to clearly enunciate the correct commands. In other words, instead of searching for a task that would perform the specific job they needed, the users started to examine which tasks they were able to utter. For instance, if the user was able to enunciate ‘weather’ and not ‘music’, then they would only use the SVA for ‘weather’. In terms of tasks, we also inferred from some users that there was a discoverability issue with SVAs. The participants sometimes did not know the capabilities of the device and how they could use it fully. This is an issue for all users and not specifically for people who have dysarthria [153]. However, this would be more difficult for people who have dysarthria because they would find it challenging to discover the tasks verbally, owing to their speech impairment.

In addition, we discovered from the interviews that people without physical disabilities could perform daily tasks themselves despite the severity level of their dysarthria. They were able to use their phones or other input devices if they wanted to use text-to-speech applications, perform a certain task or find an answer to something. Moreover, if their case was mild and their speech was still intelligible, they were able to use SVAs. However, their experiences would not necessarily be similar to those of people who have physical disabilities. Similarly, if the level of dysarthria becomes so severe that the participants cannot use their voices, they will obviously be unable to use SVAs. Consequently, the target audience of our system would be people who have moderate dysarthria and a physical disability. We hypothesise that people who do not have a physical disability but have dysarthria will prefer to use voice rather than the alternatives that they currently use. We intend to test this hypothesis in another study. However, the audience is not limited to this group and could also include people who have severe dysarthria as long as they are able to utter sounds using their voice.

4.5 Conclusion

Throughout the study, various insights, experiences and needs of participants who had dysarthria were discussed, in addition to the challenges they faced. The conclusion from this chapter is that there is a need for accessible solutions to enhance the quality of life for this group of users. Users who have dysarthria use SVAs to gain independence in daily life tasks; however, this is hindered by the capabilities of these devices. However, rather than the devices expanding their capabilities, users are adapting their needs to what they can utter and what these devices can understand.

The highlight of this study is that users positively responded to the proposed idea of using nonverbal voice cues for interaction, reflecting a willingness and interest in leveraging any potential mode of interaction that would streamline their experiences. They identified their preferences for the list of tasks, sounds and task-sound mapping.

These data help to enhance the understanding of users and thus the ability to build more effective and more usable systems. Moreover, they lay a foundation for other researchers to build on, driving the development of more accessible solutions. The chapter that follows describes the design process of the system developed according to users' needs and preferences.

The intention of this chapter was to explore the challenges encountered by individuals with dysarthria when using SVAs, which led to the following contribution:

C2 Analysis of interviews conducted with people who have dysarthria, exploring the use of smart voice assistants and revealing the challenges they face.

Chapter 5

Designing Daria, a Nonverbal Voice Cue System

This chapter represents an important phase in the user-centred design process; specifically, it focuses on understanding user requirements and initiating the design stage. Using the qualitative data gathered from Chapter 4, and lessons learned from prior studies, this chapter introduces a standardised approach for designing nonverbal voice cue interactive systems that enable people who have dysarthria to interact with smart voice assistance. This chapter also identifies key design components and shapes the design of a standardised framework. By providing a reproducible framework for developing nonverbal interactive systems for SVAs, we can increase the accessibility of said devices, their usability and their user experience.

In aligning with the user-centred design principles, this chapter delves into the design element considerations when designing our system. The focus was not only user centric but also on considerations for adaptability to the unique needs and preferences of users. The chapter also covers the list of sounds that resonate with users' preferences, ensuring ease of use while allowing for customisation to suit varied needs. Then, the chapter moves to describing the mapping between the action and sound list to ensure an intuitive and memorable mapping. The mapping approach employed principles of natural mapping and used metaphors from everyday life.

By outlining these design components and their alignment with user-centred design principles, this chapter contributes significantly to the development of a reproducible framework for nonverbal interactive systems. Such a framework is instrumental in enhancing the accessibility, usability and overall user experience of SVAs.

This chapter attempts to answer the following research question (RQ2):

RQ2: Can a standardised vocabulary that aligns with their unique speech capabilities and the range of sounds they can produce be developed for individuals with dysarthria?

The work in this chapter has been previously published in the Universal Access in the Information Society Journal [3].

5.1 Design elements

5.1.1 List

This system's design fundamentally revolves around creating an effective voice command list. Reflecting on the findings from Chapter 4, it became evident that participant preferences for command lists were split into two distinct groups. One group favoured a predefined list that allows immediate use of the system. However, the other group expressed a desire for customisation, seeking a more tailored interaction experience. Our approach aimed to accommodate both preferences.

To accommodate the need of the group that preferred customisation, provided structured flexibility in voice command customisation was provided. Users can select from a controlled list of voice commands, which were designed using input from our target user group, that they can then map to specific tasks as per their preferences. This approach ensures that although users can personalise their experiences, the system retains a level of standardisation for consistency and ease of use. This approach has been incorporated in the related literature, which has also emphasised user input in command selection and customisation. For instance, Kim et al. [99] collected preferred keywords—verbal commands in their case—from participants and programmed these into their system. When users start to use the system, they can choose various keywords that are easier for them, which are selected from the list provided by the participants in that study. In addition, Hamidi [105] relied on the same process, allowing users to customise their lists. Further, Hamidi [154] found that list customisation resulted in improvement in accuracy rates, and the improvement was significantly greater for groups whose caregivers and therapists participated in customising the list. Therefore, our approach draws on these insights, aiming to create an accessible, user-friendly interface that adapts to the unique communication styles of individuals with dysarthria.

To study the impact of list choices on users, this approach was compared to partially customising the list, which is discussed in detail in Chapter 8, with that of providing a predefined list in the system, which is elaborated on in Chapter 7. The lists in the two approaches were extracted from the interview results and users' preferences and capabilities identified from Chapter 4, as in Parker's study [104]. In this study, the participants provided a list of appliances they wanted to be able to control using their voice. The researcher then selected the words pragmatically according to the functionality of the appliance (e.g. 'on' and 'volume'). After participants recorded their voices, if the researcher found that certain voices were unclear, the researcher replaced the unclear word with another word.

5.1.2 Sounds

Sound and mapping: Mapping the sounds to their corresponding actions entailed connecting each sound to a command. To have a usable, learnable and memorable system, a framework was created for the sound–action mapping process (see Figure 5.1). First, the criteria for selecting sounds were established. Next, the sounds that met the criteria were listed, the mapping approach was decided upon, and finally, the mapping was conducted.

In the first step, setting the criteria for selecting sounds, the initial criterion was users' preferences. The second criterion was for the sound to be easy to utter. The third criterion was related to acoustic discriminability.

For the first criterion, which was users' preferences, almost all of the participants reported that vowels were convenient sounds for them to make. However, it was difficult for them to select specific preferred vowel sounds or any sound for the interactions. Therefore, the selection criteria were adapted according to the findings of [155] and [154].

The second selection criterion, namely, the sound should be easy to utter, aimed to lower the likelihood of vocal fatigue. In accordance with the users' preferences, vowel sounds were selected, in addition to nasal sounds, which are easy for people who have dysarthria to utter. Nasal sounds were added to increase the number of command combinations. The third criterion was related to acoustic discriminability, which is the ability to recognise different sounds [156]. The sounds of vowels may overlap in some cases of dysarthria [157]. Sounds that are in the corners of the International Phonetic Alphabet vowel chart were selected to minimise and avoid overlapping sounds.

The second step was sound selection. Given the disparate etiologies attributed to the various causes of dysarthria, various articulation capabilities emerged [158], which led to a limited number of vowels to choose from. For example, certain vowels (e.g. /i, a/) remain quite intelligible even in individuals who have severe dysarthria, unlike other vowels [45]. Given the capabilities of individuals with dysarthria, the vowel sound opted for comprised monophthongs, which are single-vowel voices. Notably, diphthongs (i.e. a combination of sounds) require changing the vocal tract configuration, which results in a steep second formant slope, a type of acoustic measure [159]. People who have dysarthria find it challenging to pronounce diphthongs. Given that a monophthong is composed of only one sound, it is less challenging to pronounce compared with diphthongs or other vowels. While working on these steps, we collaborated with an expert in communication disorders. Her specific work focuses on speech production for individuals with dysarthria. Her role was to help in curating and finalising the set of sounds that were distinct and easily recognisable by the participants.

The third and fourth steps, which overlap, entailed selecting the mapping approach and implementing it. As mentioned in the beginning of this section, the aim of the mapping process was to increase the usability and memorability of the voice commands by considering the users' preferences. Prior researchers have applied several mapping approaches to map sounds to actions or controllers. For example, [113] and [155] used the tongue's position to map vowel sounds with mouse movement and direction. Harada [117] incorporated a similar approach in a voice-driven drawing application. Norman [160] also discussed mapping as one of the design principles that aimed to increase system usability. The natural mapping design principle from Norman's design principles was followed. Natural mapping occurs when the knowledge in our heads is integrated with the knowledge from the world around us. In other words, through natural mapping the relationship between our knowledge and what we are trying to control is clear and obvious.

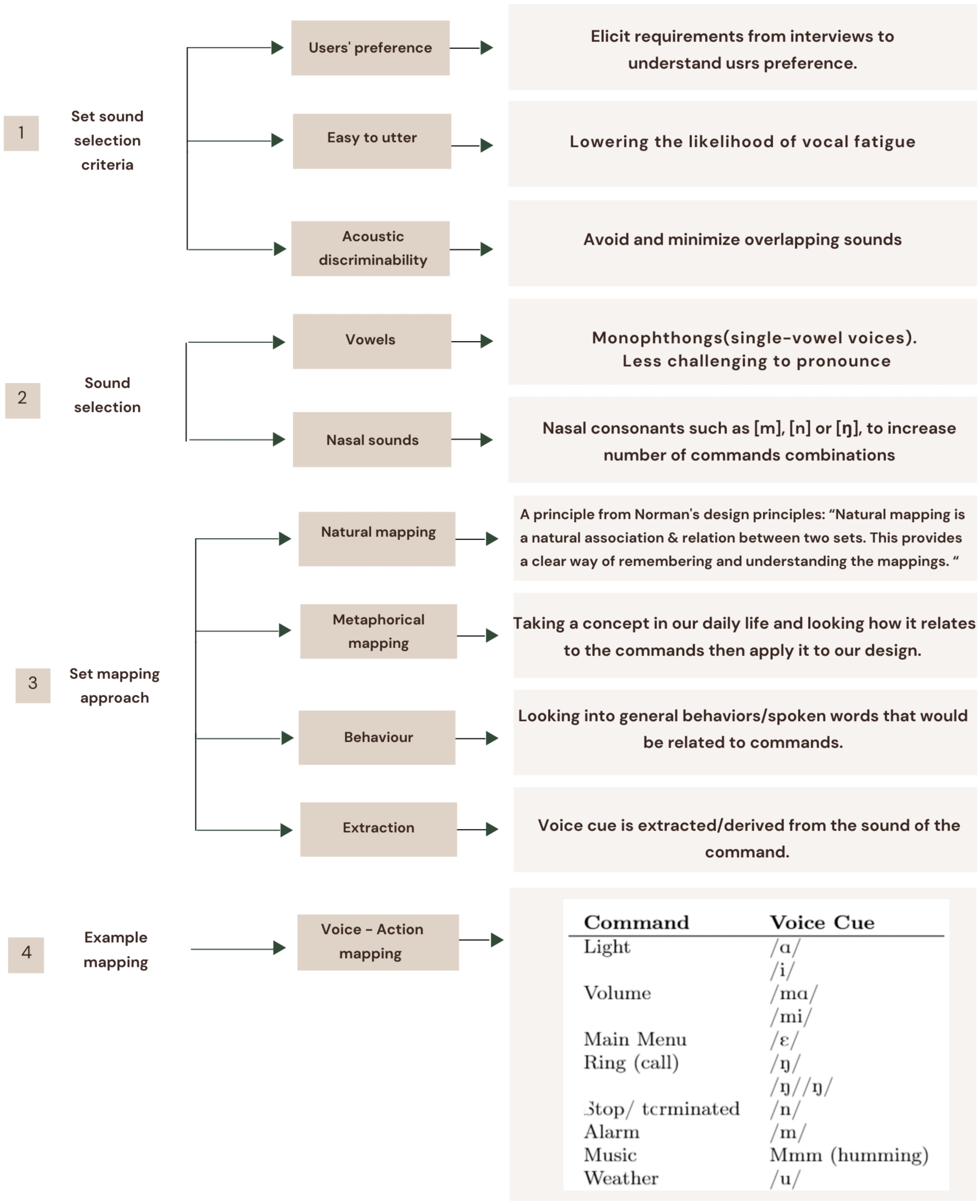
This principle was applied by taking a concept in our daily life and applying it to our design. This could also be described using life metaphors. An example from our daily life is the iPhone brightness controller, which increases the brightness by simply sliding the control up. Another example is the volume button in phones or remote controls, in which the upward direction represents an increase and the downward direction indicates a decrease. This metaphorical orientation (up means more, turn on or increase) is not arbitrary; it results from physical and cultural experiences [161]. From this concept, the /a/ vowel that most participants were able to utter was selected. This is an open vowel that requires opening the mouth and positioning the tongue far from the roof of the mouth. In our mapping, this sound represented open, up, raise and increase. This vowel was mapped with commands that had an increasing and turning on feature. Thus, the /a/ vowel was mapped with 'Turning on Light' and 'Increasing Volume'. However, a nasal letter was added before the vowels (ma) to differentiate between the commands for increasing volume.

Command	Voice cue
Light	/ɑ/
	/i/
Volume	/mɑ/
	/mi/
Main menu	/ɛ/
Ring (call)	/ŋ/
	/ŋ//ŋ/
Stop/terminate	/n/
Alarm	/m/
Music	Mmm (humming)
Weather	/u/

Table 5.1: Sound–action mapping

An effective design takes into consideration user behaviour [160]. Accordingly, users' behaviour was used as a basis for mapping one of our voices. The commands were chosen according to the users' behaviour or what they would say in certain situations. For the 'Weather' command, behaviors and spoken words related to the weather were examined. Then, 'oh' was chosen because it is used to communicate the sense that something has "just now" been noticed or realised' [162]. For example, a person could comment about the weather by saying, 'Oh, the weather is nice' or 'Oh, the weather is cold'. In her book [163], Diane described 'oh' in the sentence 'Oh, this weather is awful' as an attitudinal adverb that expresses emotion or attitude. From this, the /u/ vowel was used, which has the same sound, for the 'Weather' command. Next, given that people hum when they are trying to recall or repeat the lyrics of a song, for the 'Music' command, 'Hmm' was chosen. Table 6.1 summarises the sound–action mapping list.

The last technique was extracting the voice from the words, following Haradah [117], who used the ‘ck’ sound for the command ‘Click’. In our mapping, the sound ‘ing’ was used for the command ‘Ring’, which was intended for calling someone. Similarly, the sound ‘am’ was used for the ‘Alarm’ command.


1
Set sound selection criteria
Users' preference
Elicit requirements from interviews to understand usrs preference.
Easy to utter
Lowering the likelihood of vocal fatigue
Acoustic discriminability
Avoid and minimize overlapping sounds
2
Sound selection
Vowels
Monophthongs(single-vowel voices). Less challenging to pronounce
Nasal sounds
Nasal consonants such as [m], [n] or [ŋ], to increase number of commands combinations
3
Set mapping approach
Natural mapping
A principle from Norman's design principles: "Natural mapping is a natural association & relation between two sets. This provides a clear way of remembering and understanding the mappings. "
Metaphorical mapping
Taking a concept in our daily life and looking how it relates to the commands then apply it to our design.
Behaviour
Looking into general behaviors/spoken words that would be related to commands.
Extraction
Voice cue is extracted/derived from the sound of the command.
4
Example mapping
Voice - Action mapping

Command	Voice Cue
Light	/ɑ/
	/i/
Volume	/mɑ/
	/mi/
Main Menu	/ɛ/
Ring (call)	/ŋ/
	/ŋ//ŋ/
Stop/ tcrminated	/n/
Alarm	/m/
Music	Mmm (humming)
Weather	/u/

Figure 5.1: Sound–action mapping framework

5.2 Conclusion

This chapter presented a standardised approach for designing interactive systems using nonverbal voice cues for individuals with dysarthria.

Through this chapter, the main elements were identified as the basis for designing such systems. The key takeaways from this chapter emphasise the importance of customisation options in the design, leading to a substantial enrichment in the interaction between individuals with dysarthria and SVAs. Offering a ready-to-use system and an option to customise the mapping to the user's preferences showcases the importance of maintaining balance in the system and accommodating the needs of users.

Moreover, the process of mapping sounds to actions highlights the significance of intuitive design and demonstrates how metaphors from daily life can be used to foster seamless interactions. Furthermore, extending our research to include a wider array of sounds and mappings, while also considering the unique challenges faced by individuals with dysarthria, will be pivotal.

In conclusion, this chapter has described a standardised, user-centred approach to designing systems. The proposed framework lays the groundwork for future research into nonverbal interaction using SVAs. Thus contributing to existing knowledge on designing interaction systems by providing the steps followed to design the sound–action approach.

The intention of this chapter was to create a new language for interacting with SVAs for people who have dysarthria. In addition, the aim was to design a framework that future researchers could build on. This leads to the following contributions:

C3 An interaction framework for designing systems for smart voice assistants and people who have dysarthria.

C4 Creating a vocabulary that allows users to keep the verbal modality, known as nonverbal voice cues.

Chapter 6

Development and Evaluation of Daria, a Nonverbal Voice Cue System

This chapter introduces the design of the nonverbal voice cue interaction system, Daria, and represents Steps 3 and 4 of the user-centred design process: the design and evaluation phases. The name 'Daria' is derived from the word 'DysARthrIA' and has all the letters in the same order. The findings from the previous study indicate that there is a need for alternative interaction methods. Moreover, there is a need for interacting with SVAs directly, effortlessly and using the voice modality. The system design in this chapter responds to the results in previous chapters. In Chapter 4, the results from Study 1 identified participants' preferences concerning a list of tasks, sounds and task–sound mapping. Chapter 5 explained the framework for designing this system. Building on these foundations, this chapter presents the system implementation according to the identified requirements. It links the previous chapters, which described the theoretical understanding of interaction systems, to the practical application of these concepts.

The primary objective of this chapter was to explore the technical details of the system's design, presenting the steps we took to develop our nonverbal voice interaction system. In addition, we aimed to validate the efficacy of our system through a preliminary study. This study tested two design options, as outlined in the previous chapter (see Chapter 5): a pre-mapped list option and a customisation option. Another key objective was to understand user interactions with the system under these two distinct design scenarios.

In this chapter, we discuss the key technical decisions that influenced the structure of the system. We then describe the system components and architecture. Following this, we begin the process of collecting recordings from participants who have dysarthria because these recordings will be used to train the system. The chapter then describes the classification model and the iterative process of building it. Finally, we tested a prototype with seven participants to test its usability and, in turn, the soundness of our framework and thereby validate our proposed approach of system design. Of these participants, three were individuals with whom we had previously conducted interviews and the remaining four were new participants. Our work empirically demonstrates how an informed, structured design of a fast, direct (verbal rather than forcing users to change modalities or use an intermediate device) method of communication improves the usability of SVAs for people who have dysarthria while allowing for a more authentic experience. The data also highlight that using nonverbal voice cues is a convenient option.

This chapter contributes to answering the following research question (RQ2):

RQ2: Can a standardised vocabulary that aligns with their unique speech capabilities and the range of sounds they can produce be developed for individuals with dysarthria?

The study in this chapter has been previously published in the Universal Access in the Information Society Journal [3].

6.1 System design

In this section, we delve into the integral components and design principles that power our system.

6.1.1 Requirements

The design of Daria was informed by user needs, performance objectives and technical considerations. To effectively serve its intended purpose, the system should:

- be user centric and focus on ease of use
- be simple in its design
- take into account user preferences
- have a list of nonverbal voice cue commands that are memorable
- ensure that the list of nonverbal voice cue commands is within users' capabilities
- ensure that the mapping of commands to actions is in accordance with metaphorical concepts and not arbitrary, facilitating intuitive user interaction.

The rest of the chapter explains the system components and technical considerations.

6.1.2 System components

The main components of the system are the nonverbal voice commands and the system's hardware and software. The process of selecting the sounds is described in Chapters 4 and 5.

List of nonverbal voice cue commands: Having nonverbal voice cues as an input will make interactions with the system faster than speech because speech takes longer to produce and the system will be waiting for the utterance to end to perform actions. In addition, processing shorter voice cues will lead to faster processing than words and sentences [109]. Eleven nonverbal voice cues were chosen as commands (see Table 6.1). Each of these commands were mapped to one action. The commands were chosen according to the following criteria: (a) user preferences, which were elicited

Command	Voice cue
Light	/ɑ/
	/i/
Volume	/mɑ/
	/mi/
Main menu	/ɛ/
Ring (call)	/ŋ/
	/ŋ//ŋ/
Stop/terminate	/n/
Alarm	/m/
Music	Mmm (humming)
Weather	/u/

Table 6.1: Sound–action mapping

from interviews with target users; (b) the ease of uttering the command, to avoid fatigue; and (c) the need to use various sounds to avoid overlapping sounds (acoustic discriminability). The mapping between the voice cue and the commands were in accordance with human reactions or gestures. The detailed approach for choosing each of the voice cues and the mapping process is described in Chapter 4.

Raspberry Pi: We used the latest version of the Raspberry Pi, the Raspberry Pi 4 Model B, which is equipped with 4 GB of RAM and a 32 GB SD card. The device was connected to a USB microphone and USB speaker. The choice of using a Raspberry Pi was due to its affordable cost, its ability to handle multiple tasks efficiently and its compactness. Moreover, we followed the United Nations' 2015 Sustainable Development Goals (SDGs) [164], more specifically, SDG 12.5 (substantially reduce waste generation through prevention, reduction, recycling and reuse). No device was required except the Raspberry Pi, which is cheap, reusable and has a long life.

Classification model: This component was responsible for the nonverbal voice cue interpretation. The classification model was a machine learning model that received the users' inputs (nonverbal voice cue commands) and mapped them to their corresponding meanings. This means that the model output was to the class or sound to which the input belonged. For example, is the input an '/ɑ/' or an '/i/'.. etc. The classification result was then mapped to the intended action, which was sent as a text request to Google Assistant to execute the desired task. This means that if '/ɑ/' is mapped to 'turn on the light', 'turn on the light' as a text is sent to Google Assistant. To ensure the privacy of the users and their voices, the classification process was executed locally on the Raspberry Pi, ensuring no data transmission to external servers, which is called edge computing [165]. Applying edge computing concepts means that the data were processed at the edge of the network rather than the cloud [165].

Google Assistant service: The Raspberry Pi was equipped with the Google Assistant service. To use Google Assistant on the Raspberry Pi, first the Google Assistant Software Development Kit released by Google was installed. The installation process involved several steps, including setting up a project on Google's developer platform (actions.google.com console), registering the Raspberry Pi device and downloading the OAuth credentials. Once Google Assistant was set up on the Raspberry Pi, it could receive commands from users in audio or text formats and act on them. The responses were through audio or as an action performed in accordance with the command.

It is important to note that at the time of implementation, Google Assistant did not offer the same comprehensive functionality as devices such as Google Home or Google Nest. Its command set was restricted, and it could not execute three of our specified commands. To address this, we recorded responses for these specific commands as if they were played by Google Assistant. The final list of commands are described in Section 6.2.3.

RabbitMQ: In software development, the challenge often arises in which there is a need to integrate diverse software systems into one cohesive unit. To connect systems, we used message-oriented middleware technology (MOM). MOM is a technology that simplifies the transfer of data and enables systems that have different hardware or software to interact with each other [166]. In our system, among the available solutions, we opted for RabbitMQ [167], an open-source message broker. We chose RabbitMQ because it was easy to implement and used the advanced message queuing protocol that enables connection between different platforms [168]. RabbitMQ received messages from the classification model described above and added these to the message queue. The messages in the queue, which indicated the requested actions, were then received and executed by Google Assistant. By streamlining this message processing, we achieved enhanced responsiveness and efficiency in our overall set-up.

6.1.3 System architecture

To summarise this section, we integrated all the components of the system. The system was implemented on a Raspberry Pi that was connected to a microphone and speaker. First, users spoke a nonverbal voice cue into the microphone, and this command was interpreted in the Raspberry Pi to the action it was mapped to. The user's commands were then converted into text-based commands that were sent to Google Assistant, which processed the request and produced a response. The response could have been a verbal response or an action performed in accordance with the user's request. Figure 6.1) describes this process.

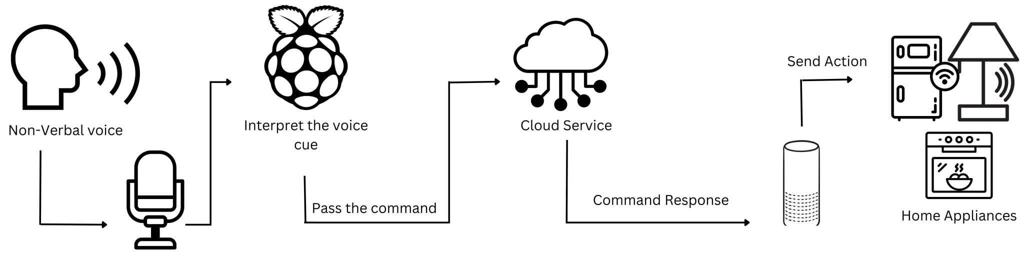


Figure 6.1: System architecture

6.2 Classification model

In this section, we cover the steps for building and training our classification model.

The first step involved collecting recordings to train the model. Then, we decided on the classification technique to use. This was followed by discussing the model architecture. After that, we trained the model, enabling it to learn and understand the features of the nonverbal voice cues. Finally, we initiated the model prediction process, in which a sound was fed into the model and the model was expected to predict which sound it was.

6.2.1 Data set

To train the machine learning classification model, the model needed to be supplied with numerous recordings of nonverbal voice cues. Given that these recordings were highly specific, to the best of our knowledge, there was no available open-source corpus that contained the 11 nonverbal voice cues required for our system, especially when uttered by individuals with dysarthria. Therefore, we undertook the process of collecting recordings from individuals who had dysarthria.

Participants

The recordings were conducted with a total of 10 participants who had dysarthria. Of these, six were participants we had previously interviewed (see Chapter 4) and four were newly recruited. The recordings were conducted in two rounds: the first round involved six participants, and the second round involved six participants as well, which included four new individuals in addition to two who had participated in the first round (see Table 6.3). All participants were above the age of 18. Their dysarthria severity levels varied: seven had mild severity, two had moderate severity and one had severe dysarthria.

Among the participants we initially interviewed, several challenges arose: some did not respond to our follow-up requests and others expressed reluctance to have their voices recorded. This reluctance was sometimes because of concerns about fatigue and a desire to conserve their voices for essential daily communication. In addition, some participants were concerned about the exertion involved and the potential strain on their voices. Most regrettably, we also faced the unfortunate circumstance of one of the participants we interviewed passing away.

Round	Total participants	Participants from previous interviews	Newly recruited participants
First round	6	6	0
Second round	6	2 (from first round)	4

Table 6.2: Distribution of participants across two recordings rounds

Procedure for collecting recordings

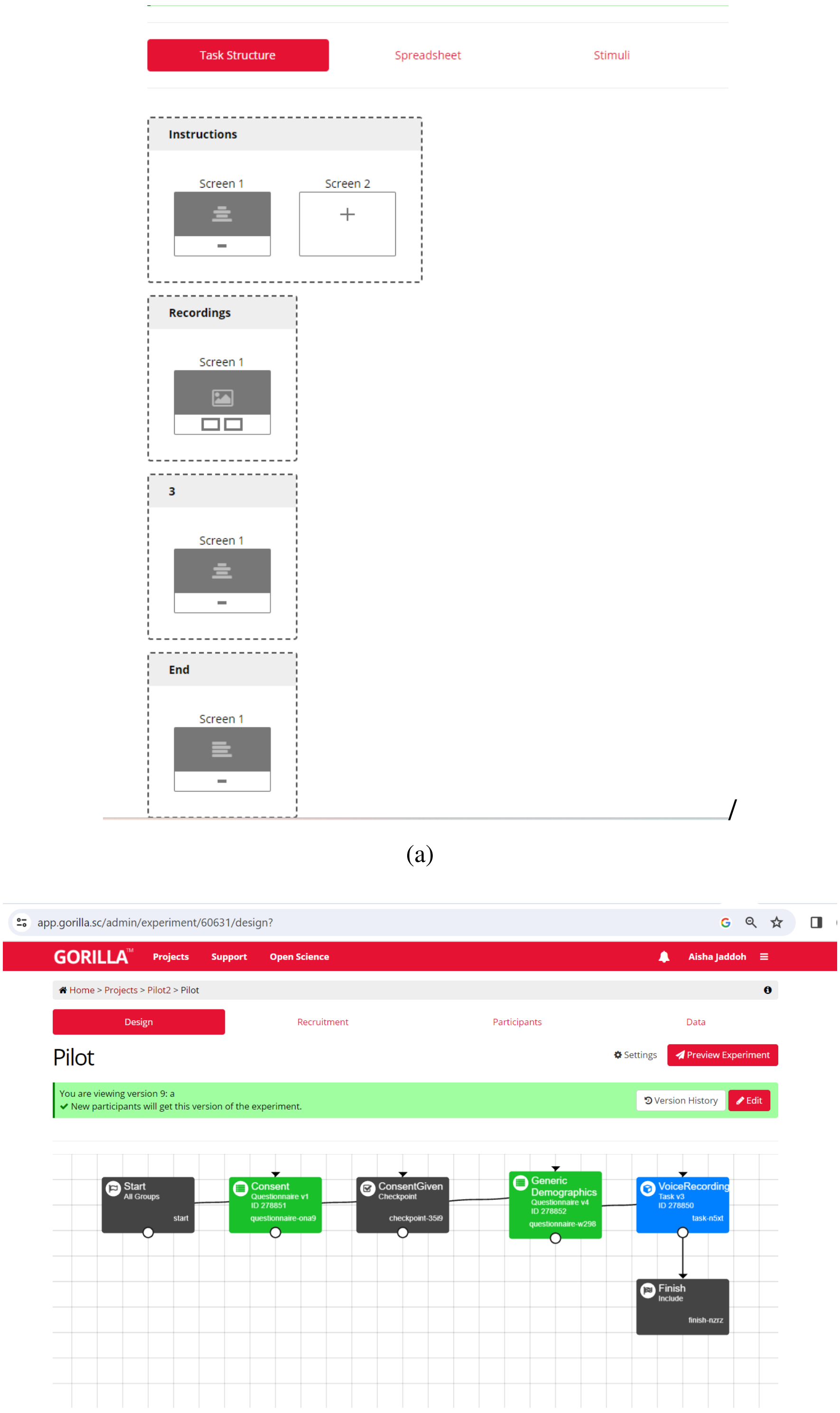
Designing the online recording experiment: The recordings were collected in 2021. Given the COVID-19 pandemic, all recordings were gathered online. We used the online platform Gorilla.sc [169] https://gorilla.sc/. This platform serves as an experiment builder, enabling researchers to design and conduct experiments online. We chose this platform because it met our requirements and made it straightforward to construct the necessary components. First, we designed the experiment, which involved designing the building blocks of the task, as presented in Figure 6.2a, in which each block on the page represents a screen. Next, we created the stimuli needed to be recorded and the survey. The total number of nonverbal voice cues to be recorded by the participants was 11 sounds. Figure 6.2b shows the final design of the experiment structure.

We sent a link to the participants, which they could click on to do the recordings at any time of their convenience. Although we could not control the recording process or the quality of the recordings, we provided users with instructions to mitigate these issues. We advised them to use a microphone and record in a quiet place 6.3a. The recording did not start unless the user was using a microphone. Figure 6.3b shows the message that prevented users from moving forward unless a microphone was plugged in. However, collecting recordings in a home environment might have been beneficial because it represented the actual setting in which the Daria system will be deployed. Furthermore, there are available data sets that have gathered recordings from participants in home environments, emphasising the realistic nature of such settings, such as the homeService corpus [93].

Recording process:

Recording Round 1: In the first recording round, data were collected from six participants. Five of the participants had mild dysarthria and one case was moderate. This recording round was divided into two sessions. We asked the participants to record each session at different times of the day or on other day because, as a result of dysarthria, their voices could differ throughout the day, depending on their level of fatigue. Thus, having recordings at various times would give us more extensive varieties of recordings, which were needed to train our system. In addition, this approach helped to prevent participant fatigue.

In Round 1, the total number of repetitions for each sound was five times. These five repetitions were divided between the two sessions: three repetitions in the first session, and two in the second session. We divided the recording sessions to not cause fatigue

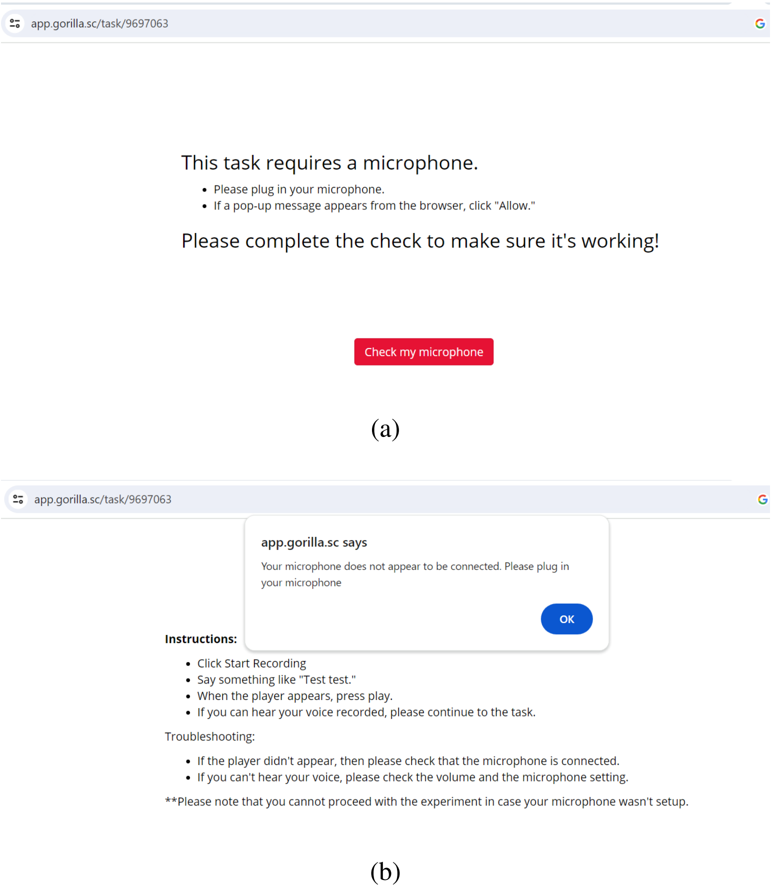


You are viewing version 9: a
✓ New participants will get this version of the experiment.

(a)

(b)

Figure 6.2: Recording experiment design on Gorilla



This task requires a microphone.

- Please plug in your microphone.
- If a pop-up message appears from the browser, click "Allow."

Please complete the check to make sure it's working!

(a)

app.gorilla.sc says

Your microphone does not appear to be connected. Please plug in your microphone

Instructions:

- Click Start Recording
- Say something like "Test test."
- When the player appears, press play.
- If you can hear your voice recorded, please continue to the task.

Troubleshooting:

- If the player didn't appear, then please check that the microphone is connected.
- If you can't hear your voice, please check the volume and the microphone setting.

**Please note that you cannot proceed with the experiment in case your microphone wasn't setup.

(b)

Figure 6.3: Recording experiment instructions

for the participants.

At the beginning of the recording session, users gave their consent to participate in this study and then completed a demographic survey in which they were asked to provide their names and indicate the severity of their dysarthria. The names were anonymised immediately after ensuring that the recordings were completed accurately. The severity question was also extended to participants who were a part of the interviews described in Chapter 4, in case there had been any changes in their condition since the time of the interview.

After the survey, the first recording session commenced. A microphone check was displayed to the participants to ensure that their microphones were connected and functioning correctly. Then, instructions were displayed to guide the participants through the recording process. After the instructions, a page appeared displaying a nonverbal voice command as text and play button. Participants could press this button to listen to the sound uttered by the same female speaker (the researcher). After listening to the sound, participants clicked ‘next’ and the recording started immediately. Once they finished uttering the sound, they clicked ‘next’ again and a new page appeared with a new word for recording and so on.

The order in which the stimuli appeared to the participants was randomised, ensuring no specific sequence was followed. Each stimulus appeared thrice throughout the recording process in the first session and twice in the second session. However, the stimuli were presented one at a time to avoid any priming effect. The term ‘priming effect’ refers to a phenomenon in which exposure to a stimulus affects the response to later stimuli, which occurs without the users’ awareness [170]. For instance, if participants heard and then recorded a particular nonverbal voice command, their responses to this stimulus could unconsciously influence how they perceived and recorded the next command. This approach was essential for ensuring the accuracy and reliability of the recordings. By randomising the order of the stimuli and presenting them one at a time, we minimised the chances of one sound influencing the participants’ responses to the next one.

Data pre-processing

Having collated our data, the next step was its pre-processing to make it suitable for feeding into our model. After the completion of the recording session, I listened to all the files to ensure that the recording process was conducted correctly because some of the files had no sounds and others contained noise, which made the sound unclear, so these files were discarded. I standardised the files into a single format then converted all the files to WAV format because this was the required input for the machine learning model.

TORGO database

To increase the overall number of recordings and enhance the data set for a more robust one, we supplemented our data with 36 files from the TORGO database [94]. TORGO, a joint effort between the University of Toronto and the Holland Bloorview Kids Rehabilitation Hospital in Toronto, contains recordings from eight speakers (three female and five male) who have dysarthria and seven speakers (four male and three female) from a non-dysarthric control group. The database is open and free for academics. We selected this data set not only because it is, to the best of our knowledge, the only one that includes nonverbal voice recordings by individuals with dysarthria but also to enhance the diversity and volume of our data. The nonverbal voice recordings in TORGO included ‘aa’, ‘ee’ and ‘oo’.

For each of the three classes ‘aa’, ‘ee’ and ‘oo’, we used 12 nonverbal voice cues from the TORGO database. Given that our data set comprised five classes in total, we needed to even out the distribution across all classes. To achieve this balance, we applied data augmentation techniques to the remaining two classes. This process involved adding noise and altering the speed of the audio files, thereby creating additional variations in the data set. These augmentation strategies enabled us to ensure an even number of samples for all five classes, including ‘aa’, ‘ee’ and ‘oo’, thus contributing to a uniform and comprehensive data set suitable for robust analysis.

6.2.2 Classification technique

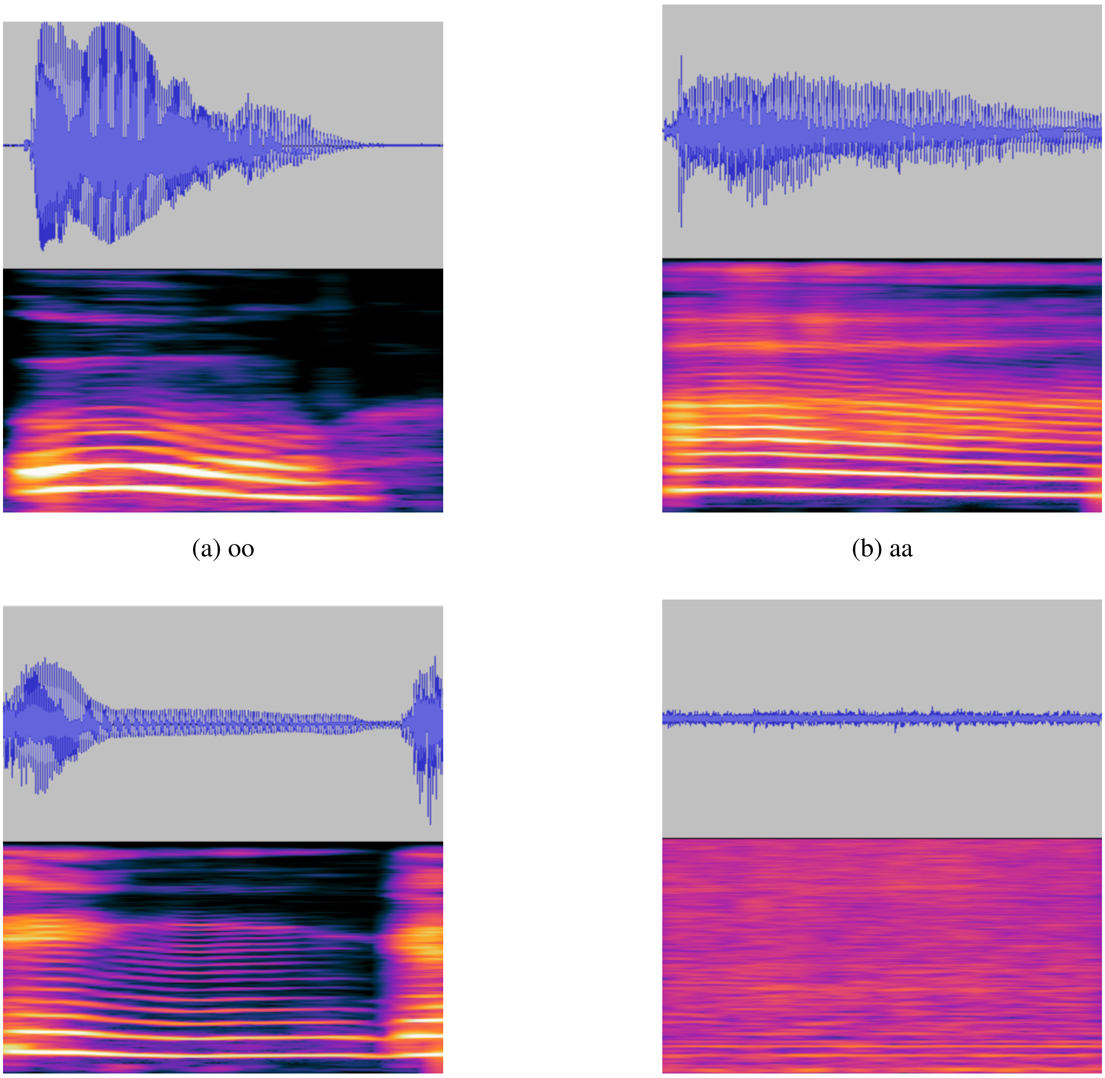
Numerous techniques exist for audio classification. Among these, deep neural networks are emerging as one of the most promising solutions. Notably, the performance of convolutional neural networks (CNNs) in terms of computational efficiency and accuracy has been shown to surpass other neural network architectures [171,172]. CNNs offer several advantages, such as robustness against background noise, proficiency in handling distorted sounds, suitability for small devices and a design that maintains effectiveness even when scaled down [173]. Furthermore, CNNs have demonstrated an ability to reduce the error rate by up to 10% compared with other deep neural networks [174]. Their applications span various tasks, including speech recognition [175], vowel recognition [174, 176, 177] and environmental sound recognition [178, 179].

Given CNNs' widespread use in image processing, to harness the efficiency of CNNs in image processing, we can transform audio into image form, such as spectrograms, to leverage this efficiency. A spectrogram converts a one-dimensional temporal sequence, typically an audio clip, into a two-dimensional image, preserving the original data while accentuating the relationship between time and frequency [180]. This is achieved by extracting time and frequency features from a signal. Extracting features from a spectrogram can be more efficient than using those from an audio file because the data from two-dimensional images provide a richer context than those from one-dimensional audio files [174]. To illustrate the distinction between the WAV format and the spectrogram format, Figure 6.4 presents a comparative representation of an audio file in the two formats; the upper is the WAV format and the lower is the spectrogram format. Each of the subfigure represents different sounds from our data set.

6.2.3 Model training

For building and training the model, we used an edge impulse framework [181], which suggested a one-dimensional CNN that contained four convolutional layers. The data were divided into 80% training data sets and 20% testing data sets. The training process involved two steps: first training the model using the training set and then testing the model using the testing data set to ensure the system's accuracy.

The model training began with the uploading of the recordings in a WAV file format. In this context, each group of similar sounds was labelled and uploaded as a separate 'class'. In machine learning, the term 'class' refers to a distinct category used in classification tasks. Each class in our model corresponded to a specific group of sounds.



(a) oo

(b) aa

(c) ing

(d) silence

Figure 6.4: Audio files in various formats

Further, these classes served as the target labels for the machine learning algorithm to recognise and differentiate. This classification step was crucial because it lay the foundation for the subsequent training process. After that, the training commenced within the framework. The system automatically converted these files into spectrograms.

The spectrograms were then input into a neural network architecture that was specifically designed to discern and learn the patterns present in them. The initial steps of the training involved the neural network learning these patterns and extracting salient features. The types of features extracted depended on the nature of the sounds and their spectrograms, which could include frequency patterns, amplitude variations and temporal characteristics. This feature extraction was pivotal because these features directly influenced the model's ability to accurately classify and interpret sounds.

Owing to the limited number of data sets available and the challenges in collecting recordings from people who had dysarthria, there was a need to find alternative methods to enhance the training of the model. One effective solution to this challenge was data augmentation. Data augmentation techniques helped to increase the number of samples available for the model during training.

These techniques were particularly beneficial because they reduced the risk of model overfitting, contributing to the development of a more robust model and saving the time and effort required to collect additional data. Although the use of more real data would have ideally produced a better model, data augmentation presented a viable alternative. In this project, data augmentation was implemented using the edge impulse framework. Two specific augmentation techniques were employed: adding noise to the spectrogram and mask time band.

Data augmentation can be applied before or after converting audio into spectrograms, depending on the augmentation type. For spectrogram augmentation, it has been proven that it enhances model accuracy [182]. Time masking, in which portions of the spectrogram are masked, helped the model to perform reliably even if the given sound was imperfect. This technique enabled the model to learn from a variety of scenarios, mimicking real-world imperfections in sound recordings.

Throughout the training, there was a dynamic flow of features and information from one layer of the neural network to another. This involved not just the transmission of data but also their refinement and transformation, enhancing the model's learning. Each feature was mapped to its respective class, allowing the model to categorise sounds according to their distinct characteristics.

Moreover, during the training, the framework incorporated optimisation and validation steps to ensure the model's efficacy and accuracy. Optimisation involved tweaking the neural network parameters to improve performance whereas the validation steps involved testing the model using unseen data (testing data set) to evaluate its generalisation capabilities. These steps were vital for ensuring that the model not only learned effectively but also performed reliably in real-world scenarios. Through these processes, the model progressively advanced in learning and understanding the features, ultimately contributing to its overall task of sound classification (see Figure 6.5).

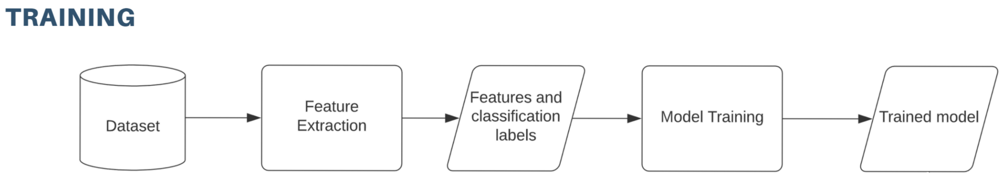


Figure 6.5: Model training

The model was trained on 11 sounds of nonverbal sound cues. Given the limited data collected and the similarities among some voice cues (e.g. 'aaa' and 'aaam'), the model underperformed: its accuracy did not surpass 50%. Figure 6.6 represents a confusion matrix that played a role in evaluating the model. A confusion matrix is a table used to describe the performance of a classification model on a set of test data for which the true values are known. In our case, it helped us to understand how well the model was identifying each of the 11 nonverbal sound cues. The diagonal elements of the matrix represent the percentage of correct predictions made for each sound class, which is where we would ideally see higher numbers. However, as indicated by the highlighted numbers in Figure 6.6, our model showed a low count of correct predictions for several classes. This suggests that the model often confused certain sounds with others, leading to misclassifications. For example, sounds such as 'aaa' and 'aaam' might have been incorrectly identified as one another, indicating that their acoustic similarities were challenging for the model to distinguish. This analysis was pivotal in our decision to reduce the number of sound classes in the model, because we aimed to eliminate those sounds that were most prone to misinterpretation by the model. After several iterative rounds of training and class removal, we reduced the number of classes of nonverbal voice cues to five (see Table 6.6). Although this constrained the number of commands available for testing, the primary aim of this work, as stated, was to validate the concept of using nonverbal voice cues as an interaction method. The decision to reduce the sound classes to five simplified the model's classification task. This reduction was crucial for tailoring the model to be more effective and efficient. These results highlight the challenges of implementing such a project for individuals with dysarthria. When the quantity of recordings is limited, this inherently constrains the variety of commands that can be included. This limitation could be mitigated if more recordings were available.

At this stage it became evident that our approach needed to evolve to gather more recordings. The initial recordings provided valuable insights, but we recognised the necessity of expanding our data collection to improve our system's performance. This realisation led us to conduct a second round of recordings using a revised approach. Our goal was to capture a broader spectrum of speech patterns in dysarthria, which was crucial for enhancing the system's robustness and accuracy. This strategic shift in our data collection methodology was aimed at obtaining a richer and more diverse data set, which was essential for the nuanced refinement of our system.

Recording Round 2: In the second round, we gathered data from four new participants and rerecorded two participants from the first session. Although some of the participants had already recorded in the previous session, because we did not collect all the sessions simultaneously, the speech of the users could have differed because a participant's speech can deteriorate or improve over time. Consequently, even though the same participants were recording, the variations in their speech could be so significant that it was akin to collecting data from new participants.

The process of conducting the recording was similar to in Round 1. However, the number of repetitions in the second session was increased to 12 because we needed a larger volume of recordings in each class. In addition, we increased the number of repetitions because of the difficulty in finding participants, thereby ensuring a more extensive collection of data from each participant. Unlike the previous round, in which recording was divided into two sessions to avoid fatigue, this time we adopted a different approach. We conducted the recordings in a single session but informed participants that they were free to stop whenever they felt it was necessary and could resume recording at their convenience. This adjustment was made to maintain participant comfort while accommodating the increased demand for data. By allowing participants this flexibility, we aimed to mitigate fatigue while still capturing the broader range of speech samples needed for our study.

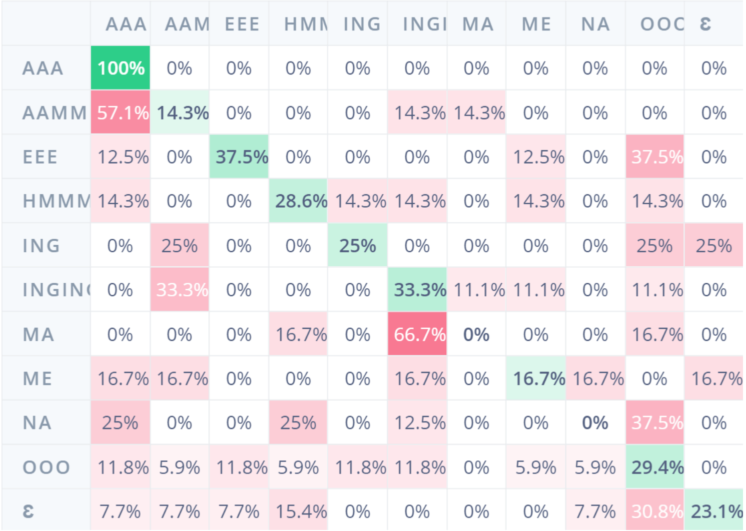
Round	Total participants	Participants from previous interviews	Newly recruited participants
Second round	6	2 (from first round)	4

Table 6.3: Distribution of participants across two recordings rounds

Finally, after retraining the model using the five classes, we added two more classes to the data set: silence and noise. These classes were important because the model needed to be able to classify whether the input was just noise or silence, rather than one of the five nonverbal voice cues. Figure 6.7 presents the confusion matrix and the seven classes.

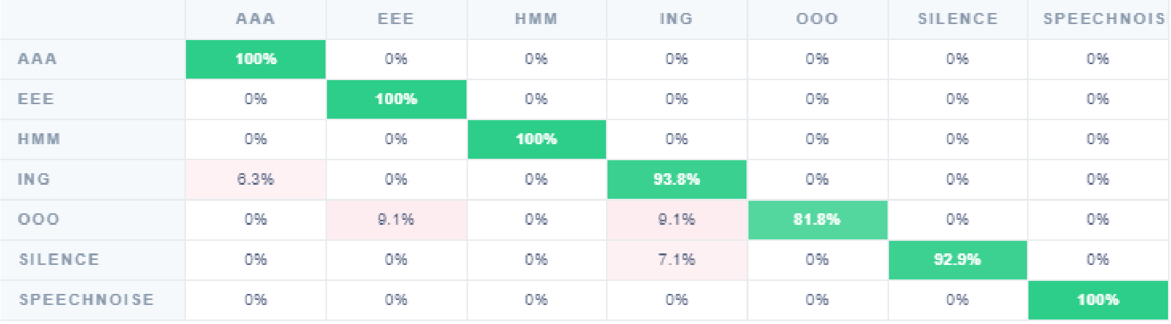
6.2.4 Model prediction

After the model was trained, we proceeded to optimise it, tailoring it to meet the specific requirements of the Raspberry Pi. The platform offered options for additional layers of optimisation, which was critical given the limited computational resources of the Raspberry Pi. This optimisation process was designed to make the model more



	AAA	AAM	EEE	HMN	ING	INGI	MA	ME	NA	OOC	ε
AAA	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
AAMM	57.1%	14.3%	0%	0%	0%	14.3%	14.3%	0%	0%	0%	0%
EEE	12.5%	0%	37.5%	0%	0%	0%	0%	12.5%	0%	37.5%	0%
HMMM	14.3%	0%	0%	28.6%	14.3%	14.3%	0%	14.3%	0%	14.3%	0%
ING	0%	25%	0%	0%	25%	0%	0%	0%	0%	25%	25%
INGIN	0%	33.3%	0%	0%	0%	33.3%	11.1%	11.1%	0%	11.1%	0%
MA	0%	0%	0%	16.7%	0%	66.7%	0%	0%	0%	16.7%	0%
ME	16.7%	16.7%	0%	0%	0%	16.7%	0%	16.7%	16.7%	0%	16.7%
NA	25%	0%	0%	25%	0%	12.5%	0%	0%	0%	37.5%	0%
OOO	11.8%	5.9%	11.8%	5.9%	11.8%	11.8%	0%	5.9%	5.9%	29.4%	0%
ε	7.7%	7.7%	7.7%	15.4%	0%	0%	0%	0%	7.7%	30.8%	23.1%

Figure 6.6: Confusion matrix: 11 classes



	AAA	EEE	HMM	ING	OOO	SILENCE	SPEECHNOIS
AAA	100%	0%	0%	0%	0%	0%	0%
EEE	0%	100%	0%	0%	0%	0%	0%
HMM	0%	0%	100%	0%	0%	0%	0%
ING	6.3%	0%	0%	93.8%	0%	0%	0%
OOO	0%	9.1%	0%	9.1%	81.8%	0%	0%
SILENCE	0%	0%	0%	7.1%	0%	92.9%	0%
SPEECHNOISE	0%	0%	0%	0%	0%	0%	100%

Figure 6.7: Confusion matrix: five classes

efficient, enabling it to run using 25 to 55% less RAM, a crucial factor for its operation on the Raspberry Pi. Once optimised, the model was ready for deployment. Once the model was deployed to the Raspberry Pi, it was ready to receive inputs and perform predictions. This involved receiving an input, extracting its features, comparing these features with previously learned sounds and then determining which sound it matched. Figure 6.8 presents the prediction process.

Command	Voice cue
Light	/ɑ/
News	/i/
Ring (call)	/ŋ/
Music	Mmm (humming)
Weather	/u/

Table 6.4: Sound–action mapping

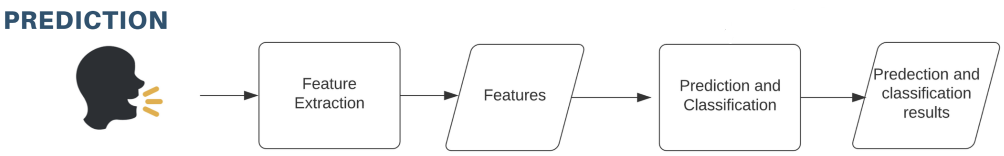

PREDICTION
Feature Extraction
Features
Prediction and Classification
Predection and classification results

Figure 6.8: Model prediction

6.3 Initial prototype and preliminary evaluation

This pilot study, as a part of the user-centred design process, primarily served as an initial test to gather early feedback and make necessary adjustments. In this section, we describe the prototype system in accordance with our framework presented in Chapter 5 and the system design and implementation presented in Chapter 6.

The initial system comprised a list of five actions and five nonverbal voice cues. Each voice cue was mapped to one action, as shown in Table 6.6. The list was selected according to two factors: first, the interview results, which indicated the activities for which the participants used SVAs and, second, voice cues that had only one sound, rather than, for example, commands that contained a vowel and nasal sounds. This was after eliminating some sounds, as described in the previous chapter. A new command was added to the list, which was not included in Table 6.1 because of the limitations of the Google Assistant functions. Specifically, given that Google Assistant used a single command for controlling lights—turning them off if they are on, and vice versa—we faced a limitation in testing distinct commands for turning lights on and off. Given this limitation, and to test more commands, we used the nonverbal voice command /i/ for another action. In addition, given that we now had five commands, instead of having two commands for lighting, we changed the command /i/ to be used for accessing news instead of for turning off the light. The mapping approach was similar to that described in Chapters 4 and 5, because the nonverbal voice command was extracted from the word itself, in this case, the letter ‘e’.

At the end of the design and implementation phase, we conducted preliminary testing using seven participants. This user acceptance test primarily aimed to explore two specific aspects: the memorability of nonverbal voice cues and user preferences concerning predefined versus customisable mapping designs, in addition to assessing the general efficacy of the system, focusing on its ability to accurately recognise and respond to nonverbal voice cues.

First, we tested the system using one participant to ensure that it performed well and had no bugs. In this test, we encountered some issues concerning the system’s sound clarity, leading us to implement external speakers to enhance the audio quality. This adjustment was made before testing the system using the remaining six participants. Despite these issues, the participants gave positive feedback about the mapping and found it clear and easy.

6.3.1 Method

Participants: The initial testing included six participants who tested the system remotely. Of these, three were male and three were female. Three cases had mild dysarthria and three had moderate dysarthria.

Set-up: The study was conducted using Zoom on a Windows laptop, which was connected to an Anker speaker to ensure superior sound quality. The system was installed on a Raspberry Pi, which was connected to a microphone. The set-up (see Figure 6.9) remained consistent for all participants and was situated in the same location within a room in the Abacws building at Cardiff University. One of the actions requested by the users was turning on the light, so a smart light bulb was also included in the set-up.

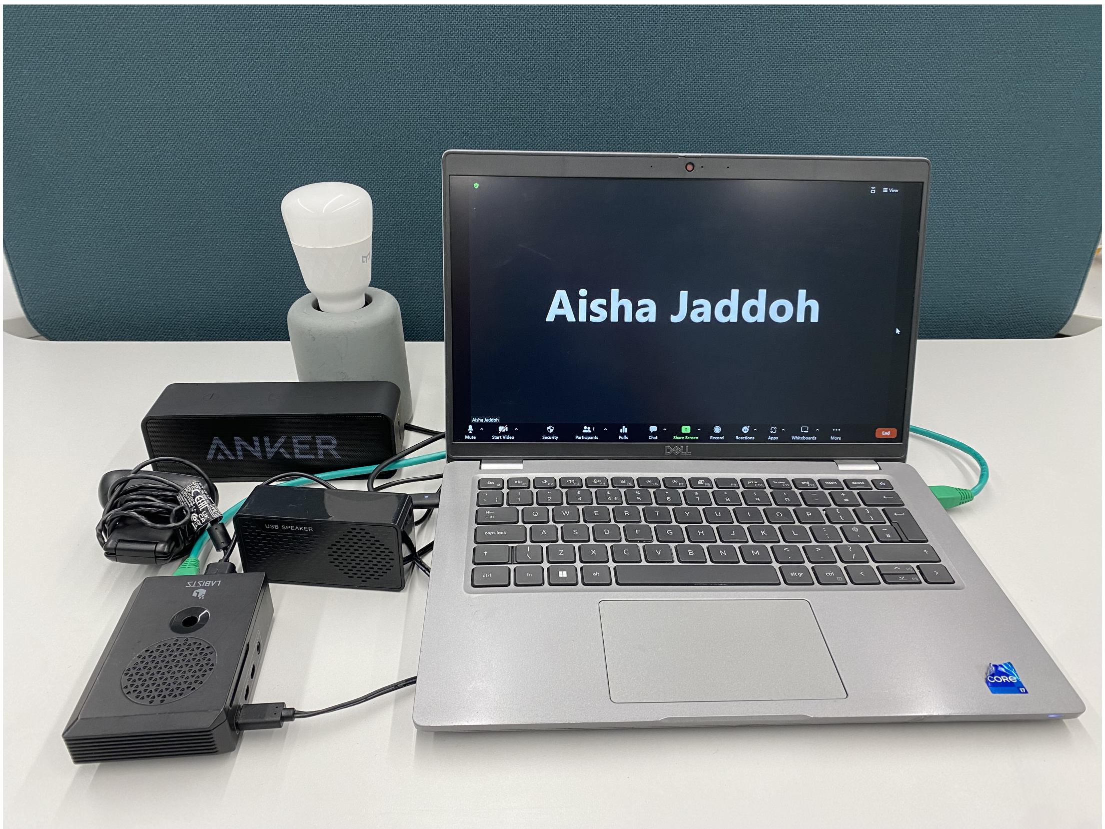


Figure 6.9: Study set-up

Procedure: This testing was conducted online. We conducted a between-subjects test, dividing participants into two distinct groups. Each group tested the system using a different list option: the first group used a predefined list and the second group used a customisable list. Our aim in having two groups was to assess the memorability of commands in each scenario. Conducting a between-subjects study helped to eliminate the carryover effect [183], especially when different scenarios were present in the study. This was crucial because users' performances might vary according to their experiences in a previous study if they were to test the two scenarios. Moreover, being introduced to the nonverbal voice commands twice could also introduce a carryover effect.

The first group (Group 1) tested the system using a set of pre-mapped commands, that is, we had already mapped the commands (voice cues) to the desired action. The second group (Group 2) tested the system using commands that they mapped to actions; thus, they chose the command for each action.

The sessions started with training, which lasted for approximately five minutes. During this, we explained the system to the two groups and introduced the list of commands. For Group 1, we explained the process we followed to map the commands to their corresponding actions, in addition to the metaphorical concepts these represented. This was conveyed through presentation slides. However, participants in Group 2 were provided with a set of nonverbal voice commands and a list of actions and were asked to create their own command-to-action mappings. This was also conveyed through presentation slides.

Next, to ensure familiarity, the participants from the two groups were asked to practise each command that was on the screen aloud at least once. During the practice, we provided immediate feedback on their pronunciation and command execution to ensure accurate recognition by the system. After starting the system for the actual test, we prompted the users to select and utter one of the nonverbal voice commands shown on the screen (the sound and the expected action were on the screen). After completing an action, participants could choose another command. If a command was not detected or was detected incorrectly, they could repeat it. There was no limit on how many times they could repeat it; we gave them the choice. The commands were listed in a random order, and this order was randomly changed for each participant to avoid the serial position effect [184], which explains why our memory often recalls the first and last items in a list more vividly than the middle items. The participants were instructed to follow any order that they preferred.

Finally, we conducted post-test semi-structured interviews to collect feedback from the participants about the system, specifically, the sounds uttered and the mapping. We asked the participants how easy it was for them to utter the sounds and sought their feedback about the system, specifically, whether they would use it again and how simple they found it to use. Their feedback and insights helped us to understand system usability and user satisfaction. We also measured the effectiveness of the system by evaluating its success in performing the task and understanding the sounds, that is, the accuracy of the system.

In terms of the mapping, we asked the two groups for their feedback. Group 1 was asked for its perspective on the mapping we provided, specifically, its intuitiveness, to understand how natural these mappings felt. For Group 2, given that these participants had created their own mappings, we asked them to share the underlying reasons for their choices. This was to understand how individuals might approach the task if given the opportunity, providing insights for future design.

To understand which design option users would prefer, a standardised option for direct use (compromising customisation) or a customised option that requires initial effort, we asked the participants about their preferences. Specifically, we inquired whether they would like the system to be pre-mapped for immediate use or they would prefer to undertake the mapping themselves.

In terms of memorability, to compare it between the two groups and analyse the effectiveness of our mapping approach, we emailed them 24 hours later. We asked the participants to recall the nonverbal sounds associated with each action to gauge the memorability of the commands. The results concerning memorability will aid in enhancing system usability and inform future design considerations. Finally, task success was evaluated as a sound and system effectiveness measure for the two types of mapping.

6.3.2 Results

Through this preliminary testing, we found that the selected nonverbal voice cues were appropriately simple, being just one syllable in length, which prevented user fatigue during articulation. Moreover, they were utterable, indicating that users could produce the sound consistently. All participants reported that they had no difficulty in uttering the sounds, found the system easy to use and would like to use it again. For example, one participant stated that the system was 'easy to use because [it does not] use difficult letters like R, Z or S'.

Although the aim of this preliminary study was not to test system performance, we did observe that vocal volume is important. Levels sometimes differ between voice cues, which is accentuated when users do not increase their vocal volume for nonverbal voice cues. This challenge was attributed to dysarthria, which limits the ability to modulate volume. For example, two of the participants spoke more quietly when uttering 'hmmm', and the system did not detect their almost voiceless commands. Similarly, one participant's voice was quieter when uttering '/ŋ/'. Nevertheless, although this participant had to repeat it for the system to detect the commands, the sounds were not reported as difficult to utter.

The results also showed that Group 1 found the mapping provided by us to be learnable and memorable. This opinion was parallel with the memorability measurements we used, in that all the participants remembered all the commands 24 hours after using the system. When asked whether they would prefer to have the voice and actions pre-mapped, all the participants in Group 1 expressed a preference for the voice and

Table 6.5: Mapping memorability

	P1	P2	P3
/ɑ/	1	0	0
/i/	0	1	0
/u/	0	1	0
hmm	1	0	0
/ŋ/	1	1	0

actions to be pre-mapped. When Group 2 was asked about the mapping 24 hours later, only 40% of the commands were remembered; however, all of the participants said that they would prefer to do their own mapping. We also asked Group 2 about the reasons behind the mapping decisions. Two participants indicated that they had done the mapping randomly; one of these participants did not recall any sound and the other recalled three out of five. Another participant stated that he 'tried to choose sounds that sounded a bit like the commands' and recalled three out of five. The memorability results are presented in Table 6.5.

6.3.3 Discussion

The findings show that the system that was built using the framework proposed in Chapter 5 has the potential to be used as a method of interaction as an alternative to using different ATs that have varying modalities for people who have dysarthria. The observations about system performance and volume require further investigation. Although the devices did not detect some of the commands that were uttered quietly, the fact that the study was conducted remotely and the users were not directly by the device could have had an effect.

The preliminary findings related to the question of mapping show that participants' opinions differed regarding mapping customisation. Each group preferred to use the system in the way they tested it given that they reported it was easier. Group 1 found that the mapping made sense whereas Group 2 preferred customisation. A possible explanation for this variation is what Ellsberg [185] defined as ambiguity aversion. This concept suggests that individuals prefer known risks to unknown risks. Consequently, participants might have chosen the option they knew and thus avoided the risk of unknown factors. Therefore, further testing is required to determine whether user preferences for the pre-mapped and self-mapped systems differ.

In terms of memorability, the results suggest that meaningful mappings enhance recall capabilities. However, when mappings are done randomly, it becomes more difficult to remember the information. However, it is worth noting that other factors might enhance memorability when users create their own mappings rather than doing so in an arbitrary manner. Therefore, additional research is required to delve deeper into this aspect.

Table 6.6: Sound–action mapping

Command	Voice cue
Light	/ɑ/
News	/i/
Ring (call)	/ŋ/
Music	Mmm (humming)
Weather	/u/

6.4 Conclusion

This chapter provided details on the design of the Daria system, which was crafted according to the recommendations from previous chapters. It delved into the integration of hardware and software, in addition to the associated machine learning aspects.

The brain of the system, represented by the Raspberry Pi, demonstrated the capabilities of a tool that is affordable and readily available. In this system, we employed a classification model, specifically, a CNN. The CNN exhibited robustness and capability even when the recordings were made in suboptimal environments. The initial model struggled to differentiate between similar voice cues, especially when working with a limited data set. Thus, we adopted an iterative approach of removing and refining similar sounds, underscoring the pragmatism required for deploying systems in real-world scenarios. This exposed the challenges in designing a system for this group of users and the potential difficulties that could be encountered. The primary difficulty was in recruiting participants who had dysarthria and collecting representative voice recordings from them.

Furthermore, the process of data collection, specifically, the fact that it was conducted online in home environments, presents advantages and disadvantages. The advantage lies in the real-world applicability of the data, because this is the environment in which the system will be used. However, the lack of controlled recordings conducted in labs did affect the model's performance, and better results could have been achieved, especially using similar voice cues. In addition, although the range of severity among the participants who contributed to the recordings was valuable, expanding this range and including a greater number of participants could lead to a more robust and better system performance across a wider spectrum of cases.

In the preliminary study presented in this section, we discovered that individuals who had dysarthria found using nonverbal voice cues to be a convenient option. Nevertheless, this study does have a limitation: the study was conducted online, so this could have an effect on the system's performance. Second, only six participants were included in the preliminary testing process. This test was conducted primarily to understand users' experiences and system performance. Therefore, a larger sample should be included in future tests to gather more quantitative and qualitative data and conduct a more insightful analysis on the specific utterances that should be mapped. Despite this limitation, the study undeniably contributes to our current understanding of the research topic by providing insightful perspectives on the participants' experiences and feedback.

Although the results of this study offer valuable insights into this field, further research involving more participants is necessary to validate the findings. In addition, a broader range of evaluation metrics should be employed. This section sets the stage for further discussions and evaluations in the section to follow.

This chapter leads to the following contribution:

C5 Creating a (currently non-existing) specialised utterances data set.

Chapter 7

Standardisation in Nonverbal Voice Cue Interaction

Chapter 6 provided a comprehensive overview of the system requirements and design for Daria, our nonverbal voice cue interaction system. It also presented the insights gained from the pilot study. This foundational chapter set the stage for a deeper exploration of Daria's functionality and potential user experiences. Building on this foundation work, this chapter, which aligns with Step 4 of the user-centred design process, delves into a detailed evaluation phase.

To evaluate and test Daria, a between-subject study was conducted to assess the two system options. The study tested the system options using two distinct groups. The first group interacted with Daria using a pre-mapped list of commands, which is the primary focus of this chapter. The second group used a customised list tailored to their preferences or needs, which is the focus of Chapter 8. We also compared the nonverbal interactions using a different modality, specifically eye gaze, for both groups, which was explored in Chapter 9. The study was divided into different chapters to allow for an in-depth and comprehensive analysis of each aspect of the study. Unlike the preliminary study in Chapter 6, which provided initial insights, this study was conducted in person and involved a larger group of participants. This methodological shift allowed for more robust and nuanced data collection and analysis, offering a deeper understanding of user interaction with nonverbal voice cue systems. This approach also aimed to address any limitations identified in the preliminary study, providing a more comprehensive evaluation of the system's performance.

The primary objective of this study was to investigate the usability and user experience of the Daria system, in addition to providing comparative insights into different user experiences with the Daria system. By comparing the insights gained from the two groups and the preliminary findings, we aimed to paint a holistic picture of the Daria system's effectiveness and user friendliness. Given that this chapter focuses primarily on the study using the pre-mapped list group, this chapter aimed to determine how intuitive and user friendly the pre-mapped list was, analysing its effectiveness in aiding smooth and efficient user interaction.

Furthermore, in our study, the objective of the comparison between off-the-shelf SVAs and our proposed nonverbal interaction system was to evaluate and understand the extent to which each system, a conventional SVA and our nonverbal system, can accommodate the needs of users who have dysarthria. This comparison was crucial to ascertain which system offers greater accessibility and user friendliness for this demographic. By analysing the performance of Alexa, which is optimised for verbal interactions, we aimed to establish a baseline of current SVA capabilities. Concurrently, we assessed the effectiveness of our nonverbal system in providing an alternative means of interaction. The contrast between these two systems allowed us to identify gaps in existing technologies and explore how innovations in nonverbal interactions could potentially enhance the accessibility of SVAs for individuals with dysarthria. This comparative analysis was not just about evaluating our system in isolation but also situating it within the broader context of existing SVA technologies and their usability for people who have diverse communication abilities.

The primary contribution of this chapter is that through the study it provides evidence that using nonverbal voice cues to interact with SVAs is a viable option. It also suggests that users would employ nonverbal voice systems if they were made available. This not only opens new avenues in AT but also underscores the potential for wider adoption of nonverbal voice systems.

In this study we assessed:

- the usability of the proposed interaction technique
- the user experience
- the task workload.

Our study aimed to answer the following research questions (RQ):

RQ3: How does the use of nonverbal voice cue interaction techniques affect the user experience and usability of smart voice assistants?

RQ3.1: How memorable are the nonverbal voice cues for users?

The work in this chapter is under review in the ACM Transactions on Accessible Computing journal. (Paper number 5 in page xvi)

7.1 Methodology

7.1.1 Participants

To ensure the applicability of our system across a broad spectrum of dysarthria severities, we recruited 20 new individuals who had varying levels of dysarthria to participate in this study. These participants were distinct from those involved in previous chapters, ensuring a diverse perspective on the system's usability and effectiveness. All participants had been diagnosed with dysarthria and did not have any cognitive issues. Fourteen participants were male and six were female. Six participants had mild dysarthria, 10 moderate, and four severe. The severity level was determined by their speech and language therapists. To maintain consistency and reliability across assessments, all therapists employed the same standardised assessment known as the 'Motor

Speech Assessment'. The participants also had various etiologies. All participants are Arabic speakers, so the study was conducted in Arabic. All participants were patients at Sultan Bin Abdulaziz Humanitarian City (SBAHC), which is a rehabilitation hospital [186]. The diversity of our participants in terms of gender, severity of dysarthria and etiology was essential to our study's objective of developing a system that is effective and adaptable for a broad spectrum of individuals with dysarthria.

7.1.2 Measures

In this chapter, we expand on the methods used to measure usability, user experience and workload, building on the initial approaches explored in Chapter 6.

We assessed various attributes of usability, focusing primarily on effectiveness. According to the ISO definition [187], effectiveness is the 'extent to which planned activities are realised and planned results are achieved'. To evaluate the effectiveness of the system's interaction, we documented the number of times the command was successfully executed after the user directed a request to any of the systems. Each command, including nonverbal voice cues and verbal commands, was repeated five times. For each command, the user would utter it clearly then wait for the system to execute the task before repeating the command. The repetition was intended to reduce the effect of randomness [188]. This was to ensure consistency of the response pattern. This number is in line with other studies in many domains, in which a range of three to five repetitions has been commonly used to ensure data reliability without compromising participant comfort and study efficiency [188–191]. This measure directly reflected the practical usability of the system in real-world scenarios.

Another important aspect of usability is memorability. We assessed the memorability of the command mappings 24 hours after the study, choosing this time frame because SVAs are typically used on a daily basis. This measurement involved calculating the percentage of correctly remembered mappings. Using the measure, we aimed to understand how intuitive and user friendly the command mappings were, which is essential for ensuring long-term user engagement.

For usability, we used the SUS [132,133], which is the most widely used instrument for testing usability [192] and has been applied in various domains. Notably, it has been used in studies to measure VUIs [193]. The SUS contained 10 questions and used a 5-point Likert scale (rating from strongly disagree to strongly agree). This questionnaire helped us to quantify user satisfaction and identify areas for improvement.

To measure user experience, and gain a comprehensive understanding of it, the SASSI [134] was used. We selected this questionnaire recommended by [194], in addition to taking into consideration the suitability of the system. The SASSI consisted of 34 items divided into six categories of user experience design, according to the definition proposed by [194]. These categories included:

- system response accuracy (nine items): the users' perceptions of the system as accurate and therefore doing what they expected
- likability (nine items): the users' rating of the system as useful, pleasant and friendly
- cognitive demand (five items): the perceived amount of effort needed to interact with the system and the feelings resulting from this effort
- annoyance (five items): the extent to which the users rated the system as repetitive, boring, irritating and frustrating
- habitability (four items): the extent to which the users know what to do and what the system is doing
- speed (two items): how quickly the system responded to user input.

Each of these dimensions was scored using the mean ranges outlined in Table 7.1. These ranges provided a framework for interpreting the users' level of agreement with each dimension, from strongly disagree to strongly agree. This method of interpretation helped us to quantify and understand the participants' experiences with the system in a structured and consistent manner.

Table 7.1: Mean range and its interpretation

Verbal interpretation	Mean range
Strongly disagree	1.00 – 1.85
Disagree	1.86 – 2.71
Slightly disagree	2.72 – 3.57
Neutral	3.58 – 4.42
Slightly agree	4.43 – 5.28
Agree	5.29 – 6.14
Strongly agree	6.15 – 7.00

The third questionnaire we used was the NASA-TLX [135], which is used specifically to assess the workload required to perform a task. In this questionnaire, the lower the value, the lower the workload. Given that the NASA-TLX is typically used as a relative measure to compare results between two tasks, we used this questionnaire to compare the two systems. The questionnaire contained six questions and each question covered a dimension of workload: mental demand (How mentally demanding was the task?), physical demand (How physically demanding was the task?), temporal demand (How hurried or rushed was the pace of the task?), performance (How successful was the user in accomplishing what they were asked to do?), effort (How hard did the user have to work to accomplish their level of performance?) and frustration level (How insecure, discouraged, irritated, stressed or annoyed was the user?). The dimensions ranged from low to high, and the performance dimension ranged from good to poor. This tool was vital for understanding how the system's usage affected user workload, an important factor in overall system design and user satisfaction.

Following the completion of the interaction tasks, we conducted a post-study interview to qualitatively assess the participants' experiences and gather detailed feedback. The interview questions were designed to explore various dimensions of the user experience. Participants were asked about their perceptions of the mapping process, specifically, its ease of use and any aspects they might not have liked. We also inquired whether participants had a preference for customising the mapping themselves. In addition, the interviews sought to understand the participants' experience with the nonverbal voice cues, specifically, the choice of sounds and the inclusion or exclusion of certain letters. This qualitative approach aimed to complement our quantitative data, providing a richer, more nuanced understanding of the participants' interactions with the system and their subjective assessments of its usability and effectiveness.

7.1.3 Procedure

Before the study, ethical approval was obtained from both the Cardiff University School of Computer Science and Informatics Ethics Committee and the SBAHC Review Board.

The study was conducted at a clinic in SBAHC, on a one participant at a time basis, with each session lasting an hour. Before the study commenced, consent forms were filled out by the participants. We began by explaining the purpose of the study and ensuring that participants understood that they could request breaks or stop the study whenever needed. The study was conducted in three parts: the first to test the usability and performance of Alexa, the second to test the Daria system, and the third to test the eyegaze interaction.

In the first part, the participants were instructed to ask Alexa, verbally, to perform five tasks. It is important to note that Alexa was accessed through a mobile phone because it supports Arabic, given that all our participants were Arabic speakers. Each command was repeated five times. After the participants had completed the five tasks, they were asked to complete three questionnaires: the SUS, the SASSI and the NASA-TLX. The questionnaires were available in hard copy and soft copy formats, allowing users to choose the format most suitable for them. We used Alexa on a mobile phone, primarily because it supports the Arabic language and all participants were Arabic speakers. This decision was crucial in ensuring that the participants could interact with the system in their native language, thus facilitating a more natural and accurate representation of their user experience with voice-activated technology.

The second part of the study was conducted immediately following the first part. The participants were instructed to ask Daria, using nonverbal voice cues, to perform five tasks. It commenced with a 10-minute briefing, during which we explained the system to the participants and introduced the list of nonverbal voice commands. We explained the process we followed to map the commands to the actions, in addition to the metaphorical concepts they represented. Next, the participants were asked to utter the commands to ensure that they were capable of doing so. After this task was completed, the system testing process commenced. To accommodate participants, especially those of older ages who might have faced challenges with reading, we adapted our approach. Instead of asking participants to read the commands, we verbally communicated each command to them, in addition to its expected output. We then requested the users to send commands to the system. Each command was repeated five times. Each time, we told the participants the command and the expected action. The order of the commands was random. After this exercise, the participants were asked to complete the same three questionnaires from Part 1 to evaluate the performance of Daria. At the end, a post-study interview was conducted to gather feedback about the systems. We asked about their preferences between the two interaction systems and their feedback about Daria, the nonverbal sounds and the mapping.

Finally, to assess the memorability of the commands, we contacted the participants 24 hours after the session. We employed various approaches in contacting them. For those who were still admitted in the rehabilitation centre, we approached them directly. Meanwhile, for those who had been discharged, we reached out via text messages. During these follow-ups, we asked each participant to recall the nonverbal commands associated with each action to assess how well they remembered the mappings. It is important to acknowledge a practical limitation in this approach: the response times varied because not all participants replied immediately. This variability in response times potentially affected the accuracy of the 24-hour recall measure. Future studies should consider this factor and explore alternative solutions to ensure a more consistent time frame for assessing memorability, thereby addressing this practical limitation.

7.1.4 Analysis approach

The analysis was divided into five steps. In each step, we analysed the results according to the suitability of the measurement, using quantitative and qualitative data. The five steps included:

- user interaction and command success, to determine the success of the interaction with the system in accordance with the number of times the command was successfully executed
- memorability, to summarise the memorability result
- usability, according to on the SUS results
- user experience, according to the SASSI results
- workload, according to the NASA-TLX results.

To describe the results statistically, we used the Wilcoxon signed-rank test. This is a scale-free statistical test that is appropriate for smaller and non-normally distributed data sets. Furthermore, this test is ideal for comparing differences between two related groups, which aligned with our research design. We used this testing when analysing the significance difference between two systems when used by the same participants.

We also used the Mann–Whitney U test, a scale-free statistical method that is suitable for smaller and non-normally distributed data sets. This test is particularly appropriate for comparing independent groups. We used this test when comparing two groups of participants who tested different system options. However, these tests also come with limitations, such as less statistical power compared to parametric tests, ordinal interpretation of results, and limitations in handling complex experimental designs. [195]

7.2 Pre-mapped list group

The results of the preliminary study presented in Chapter 6 revealed a dichotomy in user preferences for nonverbal voice cue interaction: some users preferred a predefined and mapped list of nonverbal voice cues and actions whereas another group preferred customisation. Similarly, the results of the interviews in Chapter 4 showed differences in users' preferences. Thus, it is important to focus on each option separately. This section focuses primarily on testing the Daria system using a predefined list of mappings. Understanding the dynamics of this interaction is critical before we delve into customisation in more detail in a subsequent chapter, which then shifts focus to the second group. This staged approach allowed us to comprehensively explore and compare the effectiveness and user experiences associated with standardised and customisable interactions within the Daria system.

7.2.1 Participants

This group had 10 individuals who had varying levels of dysarthria. Six participants were male and four were female. Three participant had mild dysarthria, five moderate, and two severe. The severity level was determined by their speech and language therapists. The participants also had different etiologies, as shown in Table 7.2. All participants were patients at SBAHC, which is a rehabilitation hospital [186]. The diversity of our participants in terms of gender, severity of dysarthria and etiology was essential to our study's objective of developing a system that is effective and adaptable for a broad spectrum of individuals with dysarthria.

Table 7.2: Participants details

Participant	Gender	Severity	Age range	Diagnosis
P1	Male	Severe	25–44	Stroke
P2	Male	Mild	18–24	Traumatic brain injury
P3	Female	Mild	25–44	Cerebral palsy
P4	Male	Mild	25–44	Traumatic brain injury
P5	Male	Moderate	65+	Traumatic brain injury
P6	Male	Moderate	45–65	Stroke
P7	Male	Severe	18–24	Traumatic brain injury
P8	Female	Moderate	45–65	Stroke
P9	Female	Moderate	18–24	Stroke
P10	Female	Moderate	65+	Amyotrophic lateral sclerosis

7.3 Results

User interaction and command success: To assess the effectiveness of the system's interaction, we calculated the percentage of successful actions, as shown in Table 7.3. The upper section presents the results for the percentage of successful nonverbal sound utterances for each dysarthria severity level and the lower section indicates the percentage for the verbal utterances for Alexa.

On average, the rate of success when interacting with Daria was 72% for mild cases, 64% for moderate cases and 66% for severe cases. Notably, the command '/ɑ/' in nonverbal cues showed a high rate of success across all severities, which indicates its effectiveness as a command sound. However, the command '/i/' in nonverbal cues had a notably lower rate of success in severe cases, indicating potential challenges in articulation for users who have more pronounced dysarthria.

When comparing these results with Alexa, we found that Alexa performed better for mild cases. We also observed that Daria's performance remained relatively consistent across the severity levels whereas Alexa's performance varied significantly. Figure 7.1 illustrates the clear decline in Alexa's rate of success as the level of severity increased. Specifically, the average rate of success for mild cases using Alexa was 82.67%; for moderate cases, it was 33.60%; and for severe cases, it was 24%. A statistical analysis further supported this observation, as presented in Table 7.4, indicating a significant difference in the performance of four out of the five commands between Daria and Alexa. These results demonstrate Daria's overall better performance and potential suitability for users who have dysarthria, especially in moderate to severe cases.

Another analysis was conducted (see Table 7.5) to gain insights into the effectiveness of various nonverbal voice cues. The command 'aa' was identified as the most successfully recognised voice command by the system, demonstrating significantly higher rates of recognition than other tested commands. The system appeared to be more adept at recognising certain nonverbal voice cues ('ɑ', 'hmm', '/u/') than others ('/i/', '/ŋ/'). The reason may be that 'ɑ' requires less precise tongue and mouth movements.

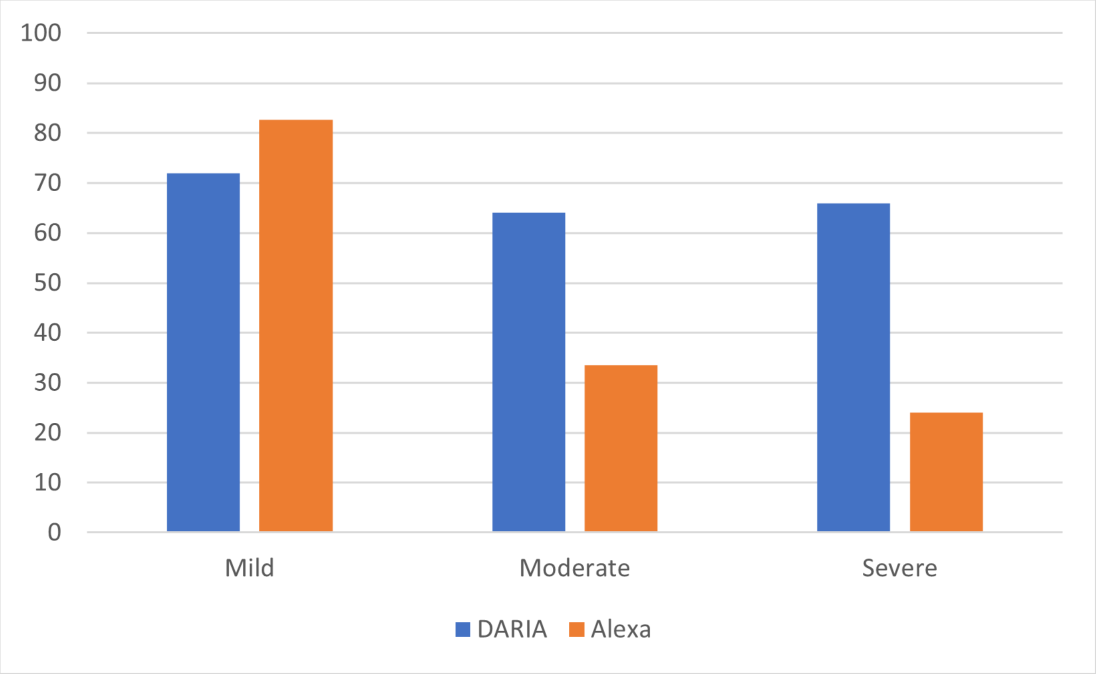


Figure 7.1: Percentage of successful commands for each severity group

Memorability: To measure the users' ability to recall the mapping of the sound and

Table 7.3: Rates of success for individual commands according to severity group

Nonverbal voice cues

	/ɑ/	/i/	/u/	hmm	/ŋ/	Overall
Mild	80	86.67	60	60	73.33	72
Moderate	80	44	64	68	64	64
Severe	100	10	60	80	80	66
All	84	50	62	68	70	

Alexa

	Light	News	Weather	Music	Call	Overall
Mild	86.67	66.67	86.67	80	93.33	82.67
Moderate	32	48	32	24	32	33.6
Severe	40	30	0	40	10	24
All	50	50	42	44	46	

command, we calculated the percentage of correctly remembered mapping, which was 80%. This high score indicates that the majority of users was able to remember the mapping. This observation aligns with the findings from the preliminary study, in which all participants who tested the system using the predefined mapping list were able to remember all commands 24 hours after use. Table 7.6 shows the memorability of each command. Here, 1 indicates correctly recalled mapping and 0 indicates not recalled mapping. From the table, the most remembered mapping was 'hmmm' for music and '/ɑ/' for turning on the lights whereas the least recalled sound was '/i/' for news.

Usability: To calculate the SUS score, we followed the procedure described in [132] taking into consideration negative and positive questions, which should be treated dif-

Table 7.4: Command success. Significance between Daria (predefined) and Alexa. The significance level was 0.05..

Command	**System**	**Mean**	***p*-value**	***p*-value assessment**
Light	Daria	0.840	**0.0004**	**Significant**
	Alexa	0.500		
News	Daria	0.500	1.0000	Not significant
	Alexa	0.500		
Weather	Daria	0.620	**0.0412**	**Significant**
	Alexa	0.420		
Music	Daria	0.680	**0.0233**	**Significant**
	Alexa	0.440		
Call	Daria	0.700	**0.0186**	**Significant**
	Alexa	0.460		

ferently in analysis.

The system was evaluated positively by the participants. They found it user friendly, given that the SUS score was 85.75 and thus rated the system as excellent [196]. For Alexa, the score was 71.5, indicating that users found Daria more usable. The analysis presented in Table 7.7 yielded significant results (p = 0.027). The bold p-values in this table and throughout the study indicate significant results.

User experience:

The mean score analysis for Daria, as shown in Table 7.8, revealed that participants strongly agreed about the speed dimension (mean = 6.45). The participants also agreed about the accuracy dimension (mean = 5.33) and the likability dimension (mean = 6.09), suggesting that they perceived the system to be fast, accurate and likable. Fur-

Table 7.5: Significance assessment between voice cues (predefined). The significance level was 0.05..

Voice cue	Mean	p-value	p-value assessment
/ɑ/	0.840	<0.001	**Significant**
/i/	0.500		
/ɑ/	0.840	0.012	**Significant**
/u/	0.620		
/ɑ/	0.840	0.059	Not significant
hmm	0.680		
/ɑ/	0.840	0.071	Not significant
/ŋ/	0.700		
/i/	0.500	0.257	Not significant
/u/	0.620		
/i/	0.500	0.095	Not significant
hmm	0.680		
/i/	0.500	0.050	**Significant**
/ŋ/	0.700		
/u/	0.620	0.549	Not significant
hmm	0.680		
/u/	0.620	0.371	**Significant**
/ŋ/	0.700		
hmm	0.689	0.808	**Significant**
/ŋ/	0.700		

ther, the participants expressed slight agreement with the habitability dimension (mean = 5.10). For the annoyance dimension, the participants provided a rating of neither disagree nor agree (mean = 3.80). Finally, the participants disagreed that the system required a cognitive demand (mean = 2.14). This is a positive result because a higher score would have reflected a negative perception. In accordance with this analysis, we

Table 7.6: Mapping memorability

Voice cue	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	# of errors	# of recalled mappings
/ɑ/	1	1	1	0	1	1	1	1	1	1	1	9
/i/	0	1	0	1	1	0	1	1	1	0	4	6
/u/	1	0	1	1	1	1	1	1	1	0	2	8
hmm	1	1	1	1	1	1	1	1	1	1	0	10
/ŋ/	1	1	1	0	1	0	1	1	1	0	3	7

Table 7.7: SUS: significance between Daria (predefined) and Alexa

System	Mean	Z	*p*-value	*p*-value assessment
Daria (predefined)	3.01	–2.11	**0.035**	**Significant**
Alexa	2.78			

can conclude that the participants had an overall positive perception of the system.

We also analysed the mean values of the responses when using Alexa, as shown in Table 7.9. When comparing the results, our system had a more positive rating for the accuracy and likability dimensions whereas habitability and cognitive demand had an equal rating. However, Alexa obtained a lower rating for the annoyance dimension.

Further statistical analysis of the data was conducted using the Wilcoxon signed-rank test, and the results are presented in Table **??**, which shows that there were no signific-

Table 7.8: Level of agreement: Daria

Dimension	Mean	Std. deviation	Verbal interpretation
System response accuracy	5.33	0.903	Agree
Likability	6.09	1.234	Agree
Cognitive demand	2.14	1.323	Disagree
Annoyance	3.80	1.360	Neutral
Habitability	5.10	1.088	Slightly agree
Speed	6.45	1.165	Strongly agree

Table 7.9: Level of agreement: Alexa

Dimension	Mean	Std. deviation	Level of agreement
System response accuracy	4.56	0.934	Slightly agree
Likability	4.46	0.839	Slightly agree
Cognitive demand	2.24	0.970	Disagree
Annoyance	2.34	0.971	Disagree
Habitability	4.70	0.985	Slightly agree
Speed	6.00	1.491	Agree

ant differences between the two systems, given that both were rated as having positive user experiences. However, there was a significant difference in the likability dimension, and our system scored higher.

When examining the results concerning the likability dimension of our system, we observed a similarity between this result and the responses to the post-interview questions. When asked about their preference for system interaction, seven of the 10 participants expressed a preference for Daria. Among these seven participants, two were categorised as having mild dysarthria, three with moderate dysarthria and two with severe dysarthria. This diversity in levels of dysarthria severity among the participants who preferred Daria provided further insight into the system's appeal across user profiles. The other noteworthy finding pertains to the annoyance dimension. Participants reported a higher level of perceived annoyance associated with Daria. It is important to note that this analysis was conducted post-study, making it challenging to return to participants for further clarification on this specific aspect of frustration. Furthermore, this aspect of frustration was not highlighted in the post-study feedback. However, a possible explanation for the annoyance could be attributed to two main factors: the limited customisation options and technical challenges during interaction. In terms of the customisation, four out of the 10 participants expressed a preference for using the system that had a customised list. The lack of such customisation options in Daria could have led to frustration among users who desired a more personalised interaction. In addition, instances in which the system failed to detect commands accurately or misinterpreted them could have further contributed to user annoyance. This highlights the importance of customisation and accurate command recognition in enhancing user satisfaction.

Table 7.10: SASSI: significance assessment between Daria (predefined) and Alexa. The significance level was 0.05..

Dimension	System	Mean	*p*-value	*p*-value assessment
System response accuracy	predefined	5.33	0.609	Not significant
	Alexa	5.20		
Likability	predefined	6.09	**0.007**	**Significant**
	Alexa	4.46		
Cognitive demand	predefined	2.14	0.498	Not significant
	Alexa	2.24		
Annoyance	predefined	3.80	**0.007**	**Significant**
	Alexa	2.34		
Habitability	predefined	5.10	0.284	Not significant
	Alexa	4.70		
Speed	predefined	6.45	0.128	Not significant
	Alexa	6.00		

Workload: We hypothesised that nonverbal voice interactions would demand a smaller workload compared with words and sentences. Consequently, the NASA-TLX [135] was used to test the amount of workload required for the two types of interactions. The results are presented in Figure 7.2 and Table 7.11. The figure shows that Daria required a lower level of workload across most dimensions. However, it is noteworthy that the workload dimensions of performance and effort were higher for Daria. In the result for the performance dimension, for which participants rated their degree of success in accomplishing the task, Alexa scored better than the Daria system. This result contradicts the result presented in Section 7.3(User interaction and command success), which showed that the rate of success when interacting with our system was greater. This was due to some participants' commands not being detected because they spoke at a lower volume and so had to repeat their command. A similar observation was made in the pilot study discussed in Chapter 6. The repeated instances across studies underline the need for ongoing improvements, especially in terms of sensitivity to varying speech volumes. In addition, two participants encountered a similar issue with Alexa, which timed out before they could complete the task. This could also have contributed to the result for the effort dimension, in which there was a small difference between the two systems, specifically, Alexa required less effort (Daria 8.33, Alexa 6.67). Finally, despite these variations in performance and effort, the overall workload between the two systems showed no significant difference.

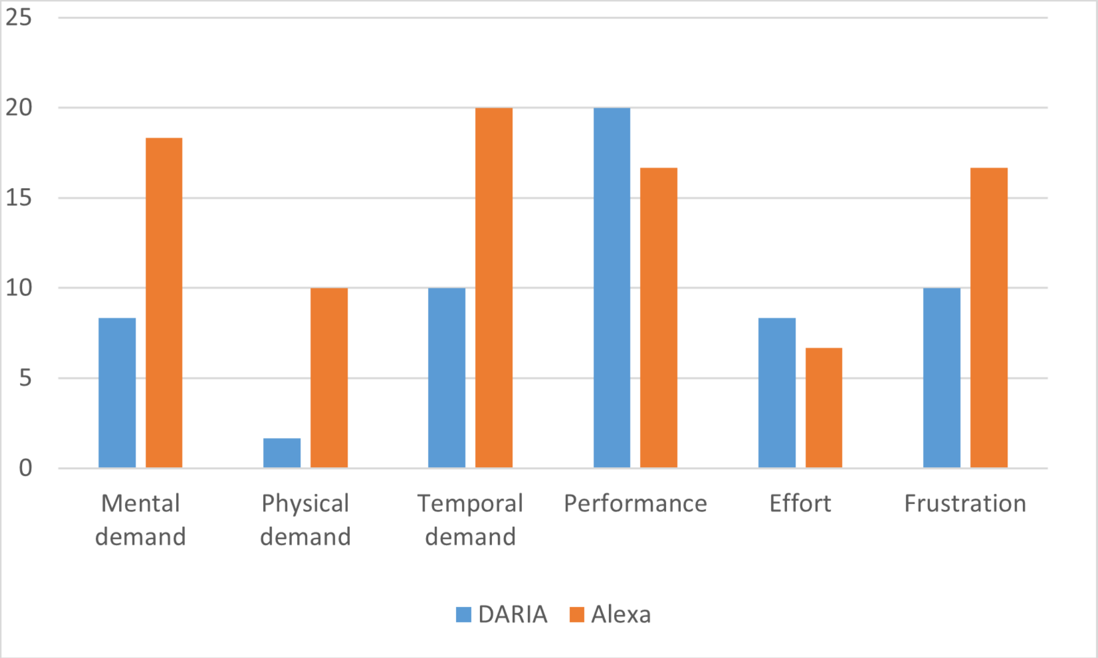


Figure 7.2: NASA-TLX results

Table 7.11: NASA-TLX: significance and effect size between Daria (predefined) and Alexa.

Dimension	System	Mean rank	Mann–Whitney U	p-value	p-value assessment	Effect size (r)	Effect size assessment
Mental demand	Daria Alexa	9.95 11.05	44.50	0.585	Not significant	0.122	Small
Physical demand	Daria Alexa	9.40 11.60	39.00	0.234	Not significant	0.266	Small
Temporal demand	Daria Alexa	9.50 11.50	40.00	0.234	Not significant	0.210	Small
Performance	Daria Alexa	10.00 11.00	45.00	0.669	Not significant	0.096	Small
Effort	Daria Alexa	10.20 10.80	47.00	0.786	Not significant	0.061	Small
Frustration	Daria Alexa	9.60 11.40	41.00	0.400	Not significant	0.188	Small

7.4 Discussion

The study revealed that there was no substantial difference in the average performance of Daria across degrees of dysarthria severity. This was unlike Alexa, for which performance noticeably declined as severity increased. These findings indicate that Daria is suitable for users who have varying degrees of dysarthria, as long as they are capable of producing vocal sounds. Using nonverbal interaction is more effective in moderate and severe cases. The performance of Daria could be further enhanced through technological advancements and additional training using a larger data set. In terms of Alexa, the limitations of this system become more pronounced as the severity of dysarthria increased, making it difficult for individuals to generate words and sentences in interactions.

Concerning the question of memorability, we found that nonverbal voice commands remained memorable even 24 hours after the study, demonstrating a high rate of recall. This indicates that meaningful mapping contributes significantly to the memorability of voice cues. Further, our detailed explanation of the meaning and reasoning behind these mappings enhanced the users' ability to remember them. This finding validates the effectiveness of the mapping process, which was explained in detail in Chapter

5. These results are consistent with those of [197, 198], who tested the memorability of earcons and musicons. Their findings highlight the importance of establishing a meaningful connection between sounds and their associated tasks.

A positive result was also obtained from the usability evaluation. The SUS score was rated excellent according to the SUS rating. This result was higher than that achieved for Alexa, one of the leaders in the smart speaker market. A significant difference was also noted in the results of Alexa and Daria. This finding demonstrates that participants found the system usable and user friendly and interactions using nonverbal voice cues were preferred over Alexa. Prior studies [2, 101] have indicated that SVAs still require improvement to cater for the needs of people who have dysarthria. Although some prior studies have focused on the accuracy of voice assistants [18,22], to the best of our knowledge, no study has measured the usability of smart voice assistance for people who have dysarthria.

The results of the SASSI questionnaire and the interviews align with the results of the SUS questionnaire. The participants had overall positive perceptions of the dimensions of the system. The statistical analysis indicated a significant difference in the 'likability' dimension, and the effect of using nonverbal voice cues in relation to 'likability' was high. This result suggests that users would employ nonverbal voice cues for interaction if this was available as an alternative.

Another finding is that, in general, Daria scored better for most of the workload dimensions (i.e. mental demand, physical demand, temporal demand and frustration), which is what we hypothesised. The reported inflexibility and lack of customisation, which contributed to 'annoyance' and 'frustration,' will be addressed in future studies.

Finally, this study demonstrated that using nonverbal voice cues to interact with SVAs is a usable and efficient alternative method. It also revealed that users can recall mapping at a high level of accuracy. In accordance with the results of this study, and to gain a better understanding of the effectiveness of different system options(customisation), an additional study was conducted, which is described in the following chapter.

7.5 Conclusion

This chapter expanded our knowledge on the use of SVAs for individuals with dysarthria. It covered commercial off-the-shelf SVAs and an alternative SVA using nonverbal voice cues, namely, Daria. The study examined usability and user experience when interacting with both systems.

The study demonstrated how the performance of off-the-shelf SVAs declines with increasing levels of dysarthria severity. However, using nonverbal voice cues remains consistently effective across various levels of dysarthria severity, presenting potential suitability for a larger and more inclusive group of users who have a speech impairment. This advancement is particularly notable in nonverbal voice interaction technology, especially for moderate and severe cases.

In addition, the study addressed the aspect of memorability. It confirmed that the mapping approach we followed aided in command mapping retention, making the system intuitive and more user friendly. Moreover, this can significantly enhance user experience and memorability.

The current data highlights users' preferences and their consideration of Daria as being more usable than Alexa. This finding is significant in the context of designing ATs that are accessible to a broader range of users. In alignment with this finding, the user experience questionnaire showed that participants had a positive perception of various dimensions of the interaction process. Moreover, using nonverbal voice cues, in general, did not show a difference in workload, indicating that users did not find a difference in the effort between the interaction methods.

There are aspects of the study that require further investigation. Future researchers could benefit from asking for more specific feedback, such as the factors contributing to users' annoyance and frustration. Moreover, to generalise the findings of this study, further studies need to be conducted using a larger number of participants and covering different etiologies and severities.

Finally, this chapter sets the stage for the next chapter, which explores the effect of customisation on the interaction experience, thereby broadening our understanding of effective SVA design for individuals with dysarthria.

This chapter leads to the following contributions:

C6 Analysis of smart voice assistants, revealing the effectiveness of nonverbal voice cue systems in accommodating diverse dysarthria severities, a significant advancement in voice interaction technology.

C7 Validating the effectiveness of our bespoke system, Daria.

7.6 Acknowledgement

We would like to express our acknowledgement and appreciation for the support provided by SBAHC. Moreover, special thanks go to the research centre and the SLP team for their support. We would also like to thank all the participants who took part in this study.

Chapter 8

Customisation in Nonverbal Voice Cue Interaction

Following the in-depth exploration of nonverbal voice cue interaction using a premapped list in Chapter 7, this chapter shifts focus to the second group of our between-subjects study: interaction with the Daria system using a customised list. This chapter continues the narrative of our comprehensive research into enhancing the usability and accessibility of SVAs for individuals with dysarthria.

In the previous chapter, various measurements were employed to understand users' interactions with SVAs when using a pre-mapped list of commands. The results of that study presented positive usability results and proved the efficacy of the system. However, according to the findings in Chapter 4, a group of the participants expressed a preference for system customisation options. This customisation allows participants to create their own mappings between a given set of sounds and specific actions. Customisation could potentially enhance the system's intuitiveness and memorability for users. Moreover, this customisation offers participants the opportunity to map the sounds that are easier for them to the most frequently used tasks. This can lead to greater satisfaction with the system.

In response, a test using a second group was initiated. This aimed to meet users' needs and understand the impact of this customisation on interaction. Specifically, it aimed to investigate how this flexibility affects the usability, memorability and overall user experience of the system for individuals with dysarthria.

Therefore, this chapter represents a continuation of Step 4 of the user-centred design process, which is evaluation. By understanding the specific needs and preferences of users who have dysarthria, we can develop more inclusive and effective SVAs. Consequently, this study addressed the following research questions:

RQ3: How does the use of nonverbal voice cue interaction techniques affect the user experience and usability of smart voice assistants?

RQ3.1: How memorable are the nonverbal voice cues for users?

RQ3.2: How does the usability, user experience and workload differ between the proposed interaction technique and using verbal interaction

RQ4: What is the impact of allowing customisation rather than standardisation on the interaction?

The work in this chapter is under review in the ACM Transactions on Accessible Computing journal.

8.1 Method

8.1.1 Participants

In this study, a new group of participants who had dysarthria was recruited, distinct from those involved in previous chapters, to eliminate potential biases from prior experiences. This approach was particularly important for accurately assessing memorability, a key measure of interest. One of the main measures that would be affected is memorability. The reasons are as follows. First, when asked about the mapping, participants could confuse the mapping from their first experience with that of the second one. Second, participants should be unaware that they would be asked to recall the mapping after the study. A total of 10 individuals who had dysarthria participated in the study. Eight of the participants were male and two were female. The participants had different levels of dysarthria severity. Three participants had mild dysarthria, five moderate, and two severe. The severity level was provided by their speech and language therapists. The participants also had different etiologies, as shown in Table 8.1. All participants were patients at the rehabilitation hospital SBAHC [186].

Table 8.1: Participant details

Participant	Gender	Severity	Age range	Diagnosis
P1	Male	Moderate	45–65	Spinal cord injury
P2	Male	Severe	25–44	Traumatic brain injury
P3	Male	Severe	18–24	Traumatic brain injury
P4	Male	Moderate	18–24	Traumatic brain injury
P5	Male	Mild	45–65	Stroke
P6	Male	Moderate	25–44	Traumatic brain injury
P7	Male	Moderate	45–65	Stroke
P8	Female	Mild	25–44	Cerebral palsy
P9	Female	Mild	25–44	Cerebral palsy
P10	Male	Moderate	65+	Stroke

8.1.2 Procedure

The procedure was exactly the same as for Study 1 (see Section 7.1.3), in which participants asked to send commands to Daria and Alexa, with the exception that we asked the participants to map each command to an action. To do so, they were given a set of nonverbal voice commands and a list of actions and were allocated time to conduct the mapping.

8.1.3 Results

User interaction and command success: We calculated the average of successful interactions according to the number of correctly completed tasks. Table 8.2 presents the results for Daria and Alexa. Starting with Daria, the average rates of success for participants who had mild and moderate dysarthria (68% and 62.4%, respectively) were relatively similar; however, the rate of success for Daria was significantly higher for severe cases (80%). For Alexa, the rate of success was lower for severe cases (38%) whereas mild and moderate cases had a higher rate of success (57.33% and 41.6%, respectively). In general, the mild cases tended to score higher; however, the rate of success decreased as the level of severity increased (see Figure 8.1). Table 8.3 presents the significant differences between various voice cues. When comparing the results with the highest voice cue, specifically, '/ɑ/', we observed that if the rate of success was lower than 52%, a significant difference started to appear.

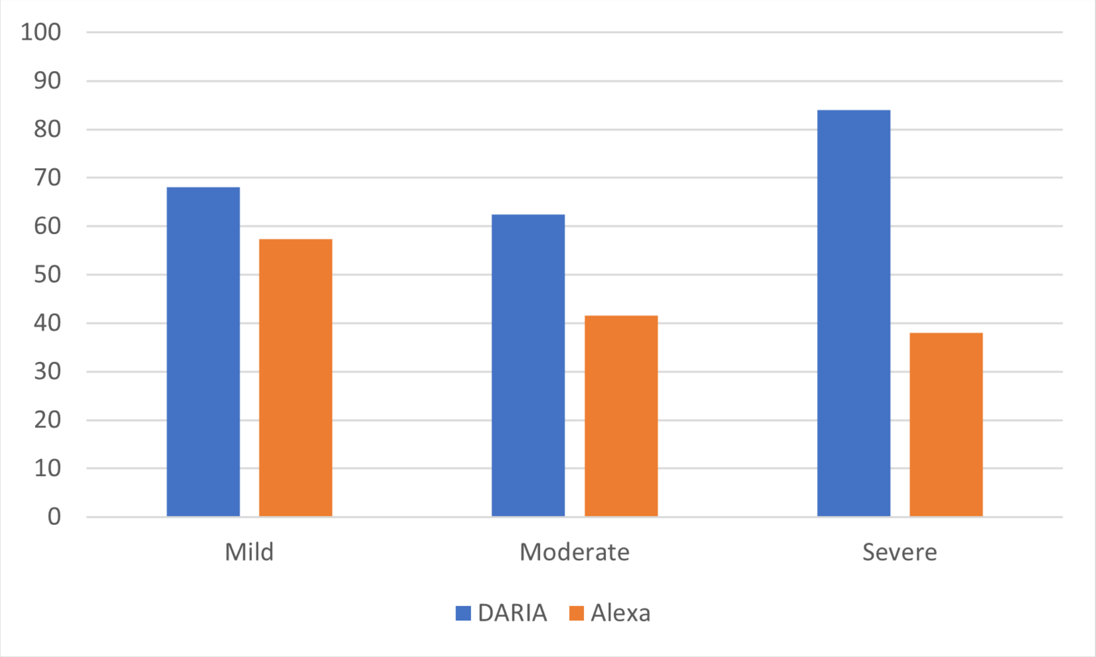


Figure 8.1: Percentage of successful commands for each severity group

Memorability: The percentage of correctly recalled mapping was 28%. This indicates a significant difference between the group who had a predefined mapping from the

Table 8.2: Rates of success for individual commands according to severity groups.

Nonverbal voice cue						
	/ɑ/	/i/	/u/	hmm	/ŋ/	Overall
Mild	93	20	93	66.7	66.7	68
Moderate	84	28	72	92	36	62.4
Severe	100	80	90	80	70	84
All	90	36	82	82	52	
Alexa						
	Light	News	Weather	Music	Call	Overall
Mild	53	73	53	53	53	57
Moderate	60	16	32	56	44	41.60
Severe	50	40	40	20	40	38
All	56	38	40	48	46	

previous chapter, and its percentage of correctly recalled mapping of 80%, and the group in this study. This significant variance highlights the impact of mapping type on memorability. Table 8.4 shows the mapping memorability of each command. Here, 1 indicates correctly recalled mapping and 0 means incorrectly recalled mapping.

Usability: The SUS score for this group was 80.6, which, according to the SUS rating scale, is considered ‘Good’ [196]. For Alexa, the score was in the same range: also ‘Good’, scoring 81.2. In the statistical analysis (see Table 8.5), we found no significant difference between the two systems, which suggests that the two systems are comparably usable.

User experience: Table 8.6 presents the SASSI questionnaire results for each of the questionnaire dimensions. Beginning with the dimensions that received the highest

Table 8.3: Significance assessment between voice cues (custom). The significance level was 0.05..

Voice cue	Mean	p-value	p-value assessment
/ɑ/	0.900	**<0.001**	**Significant**
/i/	0.360		
/ɑ/	0.900	0.248	Not significant
/u/	0.820		
/ɑ/	0.900	0.285	Not significant
hmm	0.820		
/ɑ/	0.900	**<0.001**	**Significant**
ing	0.520		
/i/	0.360	**<0.001**	**Significant**
/u/	0.820		
/i/	0.360	**<0.001**	**Significant**
hmm	0.820		
/i/	0.360	0.074	Not significant
/ŋ/	0.520		
/u/	0.820	1.0000	Not significant
hmm	0.820		
/u/	0.820	**0.003**	**Significant**
/ŋ/	0.520		
hmm	0.820	**0.003**	**Significant**
/ŋ/	0.520		

mean rate, this result indicates that participants strongly agreed with the system's 'likability' (mean = 6.48), 'accuracy' (mean = 5.34), 'habitability' (mean = 5.88) and 'speed' (mean = 5.95) dimensions. However, the mean of the 'cognitive demand' dimension was very low, which means that the participants strongly disagreed that the

Table 8.4: Mapping memorability

Voice cue	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	# of errors	# of recalled mappings
/ɑ/	0	1	0	1	0	0	0	1	0	0	7	3
/i/	0	0	0	1	0	1	0	0	0	1	7	3
/u/	0	0	0	1	0	0	0	1	1	0	7	3
hmm	0	1	0	0	0	0	0	0	0	1	8	2
ing	0	1	0	0	0	1	0	0	1	0	7	3

Table 8.5: SUS: significance between Daria and Alexa

System	Mean	*p*-value	*p*-value assessment
Daria (custom)	4.35	0.719	Not significant
Alexa	4.25		

system required a cognitive demand. Finally, participants slightly disagreed with the 'annoying' dimension. Having a low mean for these two latest mentioned dimensions is a positive result because the lower the annoyance and cognitive demand, the better the system. In summary, participants had a positive perspective about the system.

Table 8.7 presents the comparison between the Daria system and Alexa. According to the results, the mean for the 'response accuracy', 'likability' and 'habitability' dimensions was between 2.72 and 3.75, which suggests that participants slightly agreed with these dimensions. However, the results were more positive for Alexa for the 'annoyance' dimension. In terms of 'cognitive demand', the results for the two systems were equal, which shows that participants did not find the system cognitively demanding. When the results were compared with the feedback received from post-interview questions on the preferred method of interaction between the two systems, six of 10 participants preferred Daria.

A Wilcoxon signed-rank test was conducted, and Table 8.8 presents the p-values dimension. There was no significant difference between any of the dimensions, with the

Table 8.6: Level of agreement for Daria

Dimension	Mean	Std. deviation	Level of agreement
System response accuracy	5.34	0.821	Agree
Likability	6.48	0.810	Strongly agree
Cognitive demand	1.68	0.994	Strongly disagree
Annoyance	3.10	1.851	Slightly disagree
Habitability	5.88	1.042	Agree
Speed	5.95	1.383	Agree

Table 8.7: Level of agreement for Alexa

Dimension	Mean	Std. deviation	Level of agreement
System response accuracy	4.74	1.138	Slightly agree
Likability	5.17	1.485	Slightly agree
Cognitive demand	1.64	1.006	Strongly disagree
Annoyance	1.86	0.755	Disagree
Habitability	5.03	0.901	Slightly agree
Speed	6.80	0.422	Strongly agree

exception of likability, for which Daria scored higher. Overall, the results demonstrate that the Daria system was more likable than the Alexa system.

Workload: The results of the NASA-TLX questionnaire are presented in Figure 8.2. As can be seen, Daria required less ‘mental’, ‘physical’ and ‘temporal’ demand and had a lower ‘frustration’ dimension. For Alexa, the ‘effort’ dimension was slightly lower. Further, for the ‘performance’ dimension, the participants felt that they completed the task successfully using Alexa than Daria. The statistical analysis demonstrated that there was no significant difference between the two systems (see Table 8.9).

Table 8.8: SASSI: significance assessment between Daria and Alexa. The significance level was 0.05..

Dimension	System	Mean	*p*-value	*p*-value assessment
System response accuracy	Custom	5.34	0.201	Not significant
	Alexa	4.74		
Likability	Custom	6.48	**0.028**	**Significant**
	Alexa	5.17		
Cognitive demand	Custom	1.68	1	Not significant
	Alexa	1.64		
Annoyance	Custom	3.10	0.114	Not significant
	Alexa	1.86		
Habitability	Custom	5.88	0.137	Not significant
	Alexa	5.03		
Speed	Custom	5.95	0.092	Not significant
	Alexa	6.80		

To further understand these results, we examined the qualitative data collected. The analysis revealed that six participants encountered issues with Daria, including commands not being detected. This could be attributed to factors such as participants' low voice volume. However, one participant experienced problems with Alexa; specifically, the system timed out twice, requiring the user to repeat the command. These findings may provide a possible explanation for the differences observed in the workload results.

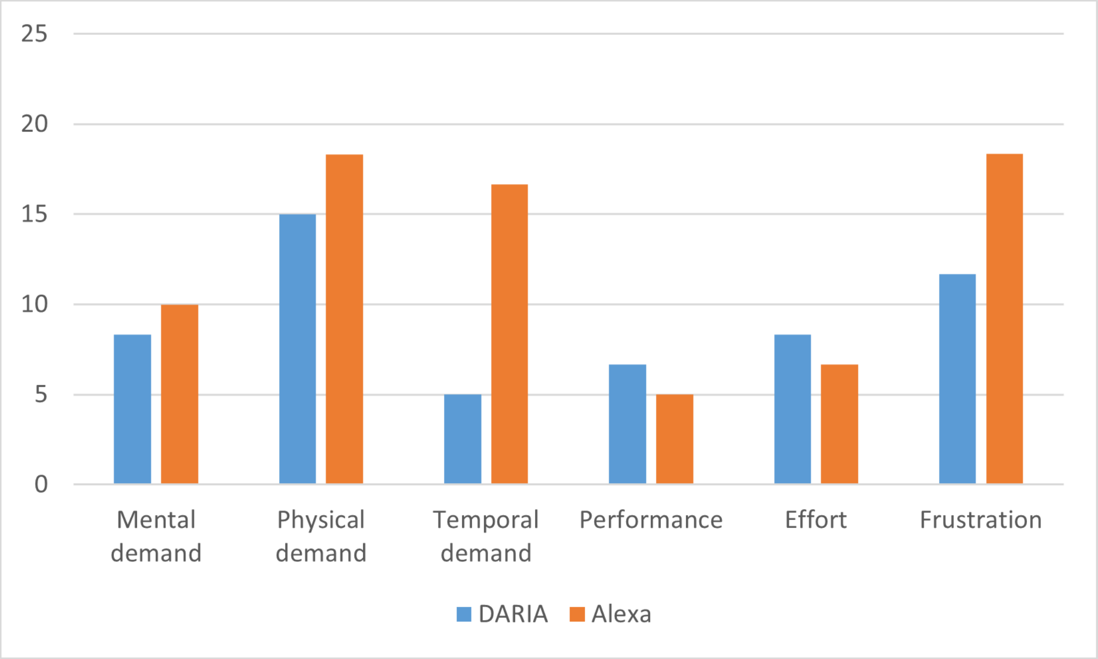


Figure 8.2: NASA-TLX score

8.1.4 Discussion

The study revealed that regardless of dysarthria severity Daria generally achieved a higher rate of success and better recognition than Alexa. As hypothesised, shorter phrases, and in our cases nonverbal voice utterances, were more effective to be used to interact with SVAs. This result aligns with prior studies, which have indicated that shorter commands are associated with enhanced recognition accuracy [66, 86]. In addition, studies [2, 158] have shown that the vowel sounds used in Daria are easy to articulate and within users' capabilities, allowing for effective detection by the system. However, it is worth noting that Alexa performed better in mild cases in study in Chapter 7.

Another notable observation was the inverse relationship between the results for Alexa and the severity of dysarthria. As the level of severity increased, Alexa's performance tended to decrease. This is explained by the decrease in speech intelligibility associated with greater dysarthria severity, leading to reduced ASR accuracy [80, 199–201]. It is

Table 8.9: NASA-TLX: significance assessment between Daria (custom) and Alexa. The significance level was 0.05..

Dimension	System	Mean	p-value	p-value assessment
Mental demand	Daria	91.667	0.317	Not significant
	Alexa	90.000		
Physical demand	Daria	85.001	0.655	Not significant
	Alexa	81.668		
Temporal demand	Daria	94.999	0.285	Not significant
	Alexa	83.333		
Performance	Daria	93.333	0.317	Not significant
	Alexa	95.000		
Effort	Daria	91.666	0.317	Not significant
	Alexa	93.333		
Frustration	Daria	88.334	0.655	Not significant
	Alexa	81.667		

important to note that Alexa was not specifically trained on dysarthric speech. However, Daria exhibited performance variations across severity levels. As can be seen in Table 8.2, there was no large difference between the results for mild and moderate cases in Daria. However, the severe cases reflected notably improved performances.

A possible explanation may be that even individuals who have the same level of dysarthria may exhibit variations in voice characteristics. For instance, two users who have moderate dysarthria may differ in terms of voice volume, that is, one speaks at a lower volume and the other at a regular volume. Such variations in voice characteristics can affect the system's performance. In addition, intra-speaker variability, which refers to variations in speech within the same speaker, can also influence results, and this variability tends to increase as severity levels increase [80]. When examining specific commands, we observed that the '/i/' command had a lower rate of success for mild and moderate cases but a higher rate of success for severe cases. This could be caused by the fact that it requires tongue raising [202].

In terms of memorability, the score was lower than in the pre-mapped list study in Chapter 7. When the participants were asked about their mapping process, all 10 reported that mapping was performed randomly, despite having as much time as they wanted for mapping. This non-systematic mapping approach made it difficult to associate the sound with the action, making participants more likely to struggle with recalling the mapping. The authors of [197, 198] supported this point, highlighting the importance of establishing meaningful connections between mapping elements to enhance memorability. Interestingly, the study conducted by [197], which employed earcones as navigation cues, demonstrated that some participants attempted to construct meaning from the sounds to help them to recall the sounds, even though the sounds were abstract and had no direct association with the task. The conclusion drawn from this study, combined with the results of Chapter 7, emphasises the significance of mapping between sound and action to promote higher recall scores. Such mapping would not necessarily need to be predetermined; users could be provided with guidance or training on how to create a mapping that would make sense to them.

When analysing the usability data, the results indicate that the SUS scores for the Daria system and Alexa were almost identical and there was no significant difference between the two systems, suggesting that the usability of the two is nearly identical. Therefore, the results suggest that the presence of a customisation option did not have a significant positive or negative impact on usability when compared with Alexa, which does not offer customisation.

In contrast, from the post-interviews, six out of the 10 participants reported a preference for a pre-mapped system for various reasons. First, it allowed them to directly use the system without the need for additional set-up. Second, a pre-mapped system was perceived as being easier to use. Third, it offered a faster user experience by eliminating the mapping step. Finally, it provided independence because users did not need assistance from others to perform the mapping.

However, the participants who preferred the customisation option highlighted that it would help them to remember the mappings and tailor the system to their preferences. In accordance with the results of this study and the results of Chapter 4, it can be concluded that users generally prefer a fast, direct and easy-to-use system that offers independence rather than relying on external assistance. However, there is a subset of users who value customisation because it aids the memorability of answers and aligns with personal preferences.

A positive outcome was obtained from the user experience questionnaire, indicating higher scores for Daria than Alexa in terms of likability. This finding is consistent with the post-study interviews, in which seven out of 10 participants expressed a preference for Daria over Alexa. The participants reported that Daria was less tiring, did not require them to speak, could understand their requests and was easy to use. However, the three participants who preferred Alexa cited different reasons, such as the availability of the system on mobile phones and more advanced technology.

The SASSI questionnaire revealed different effect levels. For the 'habitability' dimension, there was a higher favourability towards Daria (r = 0.41). However, for the speed dimension, there was a favourability towards Alexa (r = 0.13), which was expected given that Alexa is a commercial product that has been on the market for several years.

In terms of the workload questionnaire, similar to the results of Group 1, there was no significant difference. Future research will be conducted to further understand this result.

The section that follows explores the difference between Daria used through a pre-mapped list, as described in Chapter 7, and users customising the system themselves, as described in this chapter.

8.2 Overall discussion of pre-mapped study and customisation study combined

In accordance with the results of the study using the two groups, it is evident that the group using a predefined list had a higher percentage of participants who remembered the commands after 24 hours (80%) than the group that performed its own mapping (28%). Thus, users were able to recall the mapping more effectively when the mapping was predefined than when asked to map them themselves. Consequently, various recommendations emerged depending on the customisation option. If the mapping is predefined, system designers should ensure that the mapping is meaningful and understood by the user. This could be achieved by providing a guide with the system, which could be in written format or video format, for example. The same could apply to systems that have a user customisation option. Users should be educated about the importance of meaningful mapping and how it affects usability, possibly through a user guide. However, it is still necessary to examine the effects of using the system over multiple days and how this might influence memorability despite having meaningful mapping.

Another finding is that having predefined mapping is substantially more user friendly. The results indicate that the SUS score of the first group (Study 1) was higher where the interpretation of the score was 'excellent' (SUS = 85.7) whereas the second score (Study 2) was 'good' (SUS = 80.6). Various factors contributed to the pre-mapped system being more usable. As mentioned in the preceding sections, users preferred a ready-to-use system that could be directly used without the need for additional set-up or dependence on someone else. This preference was statistically significant in Study 2, in which a distinct difference emerged in the physical demand dimension. This demonstrates that the group from Study 2 required greater physical effort to perform the tasks. However, this was not the case for other workload dimensions. The negative effects of customisation have been discussed in the literature. Prior studies have found that customisation may require more time and effort from users, thereby decreasing usability [203, 204]. Although these studies were conducted on different systems and VUIs, the concept of having an additional step before using the system remains the same. Moreover, it has been found that there could be a trade-off between immediately using a system and allowing time to customise it [205, 206]. Other studies that focus on speech assistance have indicated that there is an inverse correlation between user satisfaction, usability and the effort expended to complete a task [207].

Turning to user experience, there was no significant difference between the two systems. Thus, the user experience in general was positively similar. This result was expected because the two systems were identical after the mapping step. The only difference we found was in the habitability option: the 'habitability' dimension had a medium effect, indicating that users were more familiar with the system's behaviour when using the custom option. This could suggest that during the study although the users still remembered their mapping, they better understood what the system would do. Focusing on the 'annoyance' dimension yielded contradictory results in Study 1; specifically, annoyance with Daria was higher than with Alexa whereas the frustration dimension in the NASA-TLX exhibited the opposite. The results in Study 2 differed: the 'annoyance' level in the SASSI was not significant and Daria scored lower 'frustration' levels in the NASA-TLX.

Finally, the results of the NASA-TLX presented in Table 8.10 for the two groups show a significant difference only for the physical demand dimension, specifically, that the custom option required greater 'physical demand'.

8.3 Conclusion and future work

This chapter provided a comprehensive analysis of the efficacy of nonverbal voice cue interaction in the Daria system, focusing on the use of a customised list. The findings have significant implications for the design and application of ATs for individuals with

Table 8.10: NASA-TLX: significance assessment and effect size for Study 1 and Study 2.

Dimension	System	Mean rank	Mann–Whitney U	p-value	p-value assessment	Effect size (r)	Effect size assessment
Mental demand	Predefined	10.90	46.00	0.627	Not significant	0.109	Small
	Custom	10.10					
Physical demand	Predefined	8.30	28.00	**0.039**	**Significant**	0.461	Medium
	Custom	12.70					
Temporal demand	Predefined	10.30	48.00	0.842	Not significant	0.045	Small
	Custom	10.70					
Performance	Predefined	10.95	45.50	0.675	Not significant	0.094	Small
	Custom	10.05					
Effort	Predefined	10.10	46.00	0.720	Not significant	0.080	Small
	Custom	10.90					
Frustration	Predefined	10.50	50.00	1.000	Not significant	0.000	No effect
	Custom	10.50					

dysarthria.

The findings reinforce the idea that establishing meaningful connections between voice cues and actions is crucial for enhancing memorability, thereby improving the overall usability of the system. In addition, the results suggest that customisation, although beneficial in certain contexts, may not always be the preferred option, especially for people who have dysarthria. The need for additional steps or dependence on others for customisation can inadvertently increase the physical effort required and reduce the user's sense of independence. The primary takeaway from this study is the significance of designing SVAs that are not just accessible but also thoughtfully adaptable to meet the varied needs of users who have speech impairments. This approach will ensure that SVAs can be used effectively by a broader spectrum of users, maximising utility and user autonomy.

This study sets the groundwork for future research into these interaction techniques. Future work includes creating a wider range of command to provide users with a better experience when using SVAs. Deeper studies could focus on the memorability of commands to understand their retention over longer periods or the impact of various types of training on system usage. Future researchers should also continue to seek to optimise the balance between ease of use, customisation and overall user satisfaction.

Finally, to generalise these results, larger groups should be involved in system testing to gain richer feedback.

The contribution of this study lies in:

C6 Analysis of smart voice assistants, revealing the effectiveness of nonverbal voice cue systems in accommodating diverse dysarthria severities, a significant advancement in voice interaction technology.

C7 Validating the effectiveness of our bespoke system, Daria.

8.4 Acknowledgement

We would like to express our acknowledgement and appreciation for the support provided by SBAHC. Moreover, special thanks go to the research centre and the SLP team for their support. We would also like to thank all the participants who took part in this study.

Chapter 9

Evaluating Eye Gaze Interaction and Voice Command Modalities

This chapter shifts focus to the third part of the study, which compares non-verbal interaction using a different modality: eye gaze. Our theis up to this point has focused primarily on verbal and nonverbal voice interactions. However, given the complex nature of dysarthria and the wide range of abilities among this group of users, we recognised the need to compare our approach to a different modality beyond voice. Considering that dysarthria can deteriorate to a level at which the user might have full motor disability, the most logical alternative to explore was eye gaze interaction. Therefore, we decided to test interaction using eye gaze to understand how this modality compares with the other Daria and Alexa. By doing this, we acknowledged the multifaceted nature of dysarthria. Consequently, we incorporated eye gaze interaction into our research paradigm, expanding our exploration of accessible communication technologies for individuals who have varying degrees of dysarthria.

This chapter aimed to determine how the usability of each interaction method—namely, direct speech commands through Alexa, nonverbal voice cues through the Daria system and eye gaze control—vary for individuals with dysarthria when interacting with SVAs. These methods were chosen for their potential to accommodate the varying communication capabilities of individuals with dysarthria, ensuring a broad and inclusive approach to enhancing their interaction with SVAs. By evaluating the usability of various interaction methods, this study will help to inform a more comprehensive approach to future research efforts. Understanding the strengths and limitations of each method will enable us to enhance and optimise these technologies in parallel. This will also affect the daily lives of individuals by offering them greater independence and ease of communication.

Notably, the eye gaze interaction test was conducted in the same session as the between-subject tests described in the previous chapters, 7 and 8. This approach was taken to ensure consistency in participant experience and facilitate direct, immediate comparison among the interaction methods. However, it also posed unique challenges, such as managing potential fatigue effects, which were carefully considered in our study design and analysis.

The first interaction method involved using speech. Users directly sent commands to the SVA by uttering a sentence command, for example, 'Alexa, what is the weather today?'. The second method used nonverbal voice cues in the Daria system. The third method employed eye gaze control, in which users controlled a tablet connected to the SVA using only their eyes. These methods were chosen for their potential to accommodate the varying communication capabilities of individuals with dysarthria, ensuring a broad and inclusive approach to interaction with SVAs. Understanding the strengths and limitations of each method enabled us to enhance and optimise these technologies. Usability measures for each modality were assessed according to the ISO, which defines usability as the 'extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' [187]. Furthermore, the workload required for each interaction was also evaluated.

The work in this chapter has been previously published in Applied Sciences Journal [1].

9.1 Background

Although SVAs are mainly controlled through voice, they could also be controlled through other modalities to overcome the challenges faced by people who have dysarthria. A recommendation by Masina [17] has suggested that systems may use additional modalities to overcome the voice accessibility issue for people who have speech impairments. In this study, we evaluate interaction with SVAs using eye gaze because people who have dysarthria usually do not suffer from issues with eye movement or visual impairment thus using eye gaze technology is a practical option for them [208]. Eye gaze technology works by tracking eye movement to determine the eye gaze position, which enables users to perform various actions. For example, if the user looks at the light button, the light is turned on. This could be through detecting dwell time or blinking, depending on the eye-tracking device setting.

To date, there is a relatively small body of literature that is concerned with the usability of eye gaze systems when used by people who have speech impairments in general and dysarthria in particular [209–211]. Donegan [212] explored some aspects of usability by examining user satisfaction and how eye gaze can affect the quality of life of people who have disabilities. This study pointed out that users were satisfied with the use of eye gaze and it did have a positive impact on their daily lives. Similarly, a study by Najafi [213] also evaluated the use of eye gaze devices. In addition to receiving user feedback on the use of the device, it evaluated the issues that may occur from using these devices and the adjustments that are required to be able to use them efficiently. Recently, Hemmingeeon and Borgestig [209] used surveys to assess the usability of eye gaze technology for people who have physical and communication impairments. The results reported that most of the participants were satisfied with using eye gaze in controlling computers and it was efficient to use.

9.2 Study

9.2.1 Method

In this study, we measured usability, effectiveness, satisfaction and workload. As in previous chapters, the usability attribute measured the effectiveness of interacting with the system. This was measured by the SUS.

The effectiveness attribute measured the user's ability to complete a task (task success rate). The task was considered successful if the SVA replied or correctly performed the requested command. The success was recorded during the study and confirmed by video recordings.

The satisfaction attribute measured user satisfaction with the system. It could be measured through various methods; however, in this study we measured it qualitatively and quantitatively. We measured it using the SUS questionnaire, in addition to post-study interviews to receive feedback.

Finally, given that we aimed to compare three interaction options, we considered the workload required to perform the tasks. This was measured by the NASA-TLX questionnaire. This questionnaire contained six questions. Each question covered a dimension of workload: mental demand, physical demand, temporal demand, performance, effort and frustration level.

9.2.2 Participants

Eight of the participants who had dysarthria participated in this section of the study. These participants also participated in the studies discussed in previous chapters. However, we could not conduct the study with all participants for various reasons. Some participants had issues with their eyes or vision, or the eye-tracking device did not work effectively for them because of uncontrolled head movements caused by different

Participant	Gender	Severity	Age range	Diagnosis
P1	Male	Mild	25–44	Traumatic brain injury
P2	Male	Mild	45–65	Stroke
P3	Female	Mild	25–44	Cerebral palsy
P4	Male	Moderate	45–65	Spinal cord injury
P5	Male	Moderate	25–44	Traumatic brain injury
P6	Male	Severe	25–44	Stroke
P7	Male	Severe	25–44	Traumatic brain injury
P8	Male	Severe	18–24	Traumatic brain injury

Table 9.1: Participants' details

cases of dysarthria. None of the participants had cognitive issues, ensuring that their responses and interactions with the systems were solely influenced by their dysarthria condition. All the participants were patients at SBAHC. Participants were adults of various ages, ranging between 18 and 65, and had varying levels of dysarthria severity. The level of severity was provided to us by the language and speech therapists of the patients. Although all participants had experience using voice technologies, none had prior exposure to eye-tracking systems. Table 9.1 provides details about the participants.

9.2.3 Set-up and Equipment

The study occurred in a clinic in the medical city. For testing the verbal interactions, we used Alexa on a mobile phone, primarily because it supported the Arabic language and all participants were Arabic speakers. To test the nonverbal voice interactions, we employed the Daria system. Finally, to test the eye gaze interactions, we used the Tobii Eye Tracker 4C, an off-the-shelf eye tracker from one of the leading eye-tracking companies [214]. This eye tracker was compatible with Windows PCs and easy to use. The tracker was affixed magnetically to the bottom of a laptop screen. For the eye-tracking interactions, the key components were the eye-tracker device and a HTML web page that contained buttons. Each button represented a command. This page was connected to the Raspberry Pi through the RabbitMQ message broker. The user interface of this HTML page was designed in accordance with guidelines that recommend large buttons to ensure ease of selection and interaction [215].

During the study, participants were asked to instruct the devices to perform five tasks. These tasks were turning on the lights, playing music, playing the news, calling someone and asking about the weather. First, the participants verbally asked Alexa to perform these five tasks. The order of the tasks was randomised and varied for each participant. Next, the participants switched to Daria, and they gave commands using nonverbal voice cues. Finally, they tested the eye gaze system.

When the participants used the eye tracker, a bar containing buttons appeared at the top of the screen, as shown in Figure 9.1. A red circle, functioning as a cursor (see Figure 9.1), also appeared, positioned on the second item from the left on the bar, which participants controlled using their eyes. Participants were instructed on how to use the tracker, including identifying which buttons represented the left-click mouse function and the confirm button. The users were required to first select the mouse click button and then, once this bar disappeared, needed to choose (Figure 9.2) one of the boxes that appeared on the screen using the cursor. Each box represented a command. Once the user pointed to one of the boxes, it would be highlighted and the action would be performed.

After each part, the participants filled out the SUS and the NASA-TLX questionnaire. At the end of all three parts, post-study interviews were conducted to ask about their preferences among the three systems.

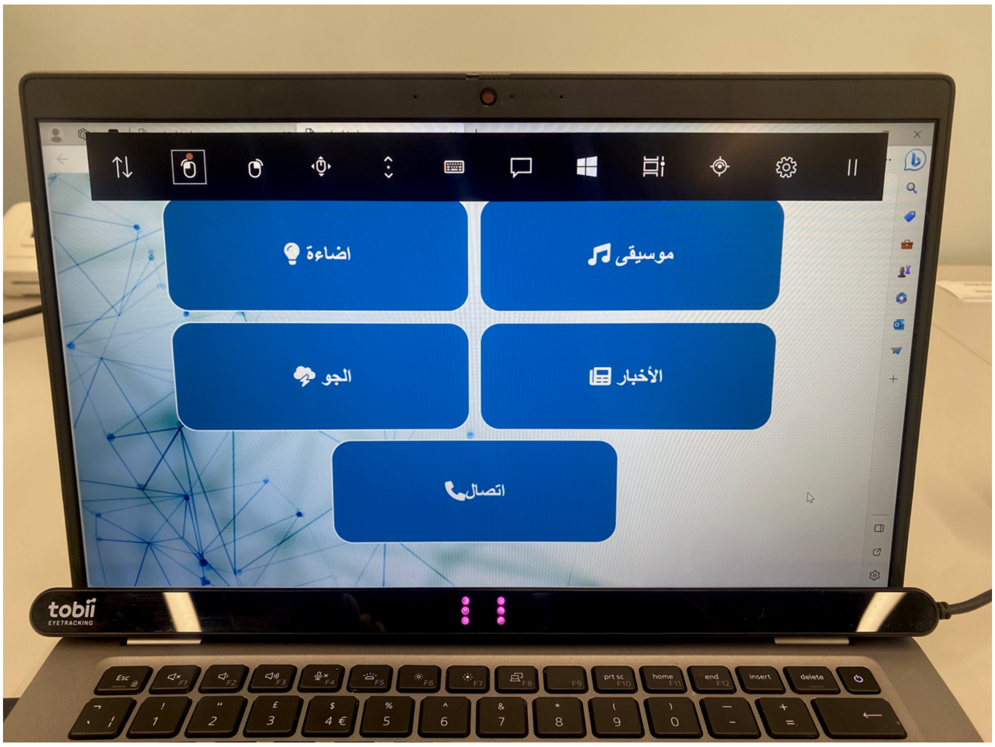


Figure 9.1: Eye-tracker control bar

9.3 Results

9.3.1 SUS

The SUS result for Alexa was 79.06. According to the SUS rating scale this is equivalent to ‘Good’. For the Daria system, the score was 84.68. This score is also equivalent to ‘Good’. Finally, for the eye gaze system, the score was 52.81, which indicated that the evaluation was ‘OK’.

To compare the three approaches, a statistical analysis was conducted using the Friedman test, which is suitable for small sample sizes and comparing the same subjects. The results indicated that there was a significant overall difference between the options. To better understand these differences, we conducted a pairwise statistical analysis using the Wilcoxon test. This test revealed significant differences when comparing the eye gaze approach with the Daria system (p = 0.011), favouring the Daria system.

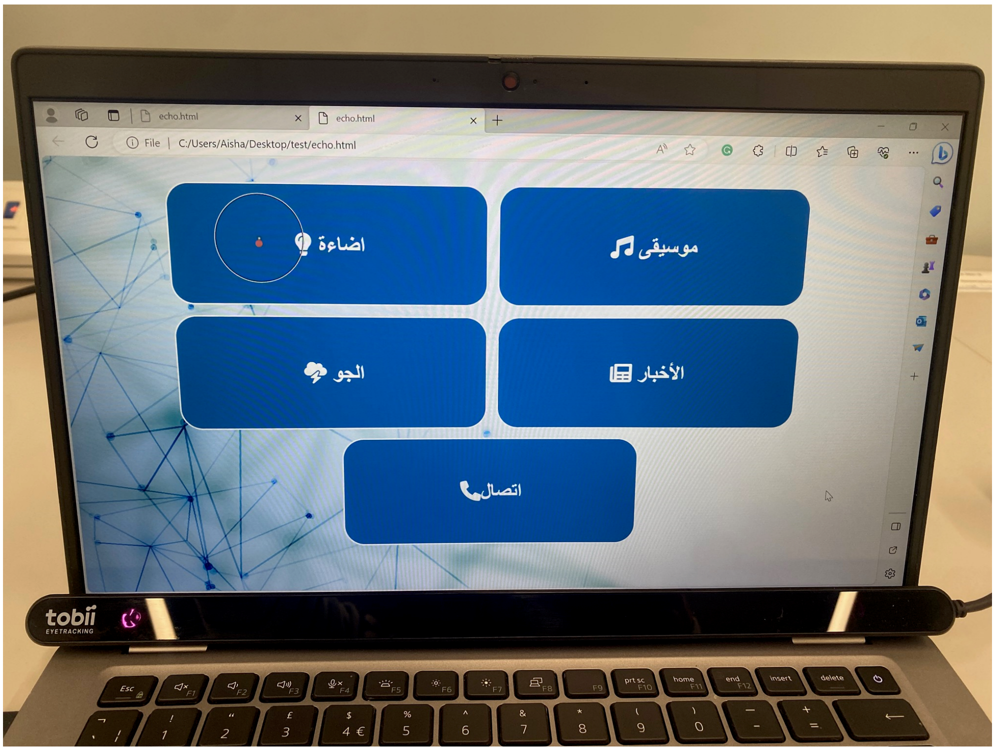


Figure 9.2: Selecting a command

Moreover, there was a significant difference between the eye gaze approach and Alexa (p = 0.011), in which Alexa was favoured. There was no significant difference between the Daria system and Alexa. To further understand how these usability scores related to the workload experienced by participants during the interactions, we turned to the NASA-TLX assessment.

To further understand how these usability scores related to the workload experienced by the participants during the interactions, we turned to the NASA-TLX assessment.

9.3.2 Workload

Given that the NASA-TLX is commonly used to compare results between tasks, we used this questionnaire to evaluate the differences between the three systems. In this questionnaire, a lower score indicated less workload and effort, which translated to better results. The averages are presented in Figure 9.3. Starting with mental demand, Daria had the least demand, followed closely by Alexa and then the eye gaze interaction, which showed a significant difference to the other two. In terms of physical demand, Alexa scored the lowest, followed by Daria, which had only a slight difference. However, the eye gaze interaction imposed a notably higher level of demand. For temporal demand, the pattern was similar to that of mental demand: Daria scored the lowest, followed by Alexa and then the eye gaze interaction. In terms of performance, which is how successful the user was in accomplishing what they were asked to do, Alexa had the lowest score, then Daria, showing a marginal difference, and the eye gaze interaction, which had a significantly higher level of demand. Finally, considering effort and frustration, Daria had the lowest score, followed by Alexa and then the eye gaze interaction.

Similarly to the statistical analysis for the SUS, the Friedman test was used to determine whether there were statistically significant differences in the workload scores across the three interaction methods (see Table 9.2). The test revealed statistically significant differences in physical demand (p = 0.003), performance (p = 0.040) and effort (p = 0.011); however, no significant differences were found for mental demand (p = 0.174), temporal demand (p = 0.054) and frustration (p = 0.244).

To understand the specific differences between each pair of methods, we conducted a Wilcoxon test for the categories that had significant Friedman results (see Table 9.3). For physical demand, the pairwise comparison showed significant differences in favour of Alexa over eye gaze interaction (p = 0.024) and Daria over eye gaze interaction (p = 0.025), indicating that Daria and Alexa required less physical effort than the eye gaze method. Similarly, for temporal demand, significant differences were noted between Daria and eye gaze interaction (p = 0.020), in which Daria required less time to perform tasks. For performance, there was a significant difference between Daria and eye gaze interaction (p = 0.041). Participants found themselves to be more successful in performing the tasks using Daria. For effort, a significant difference was found again in favour of Daria over eye gaze interaction (p = 0.018), suggesting that Daria interactions demand less effort from users. Similarly, Alexa required less effort than the eye gaze method (p = 0.042).

These findings suggest that for individuals who have varying degrees of speech impairment, nonverbal and verbal voice command methods (Daria and Alexa) may impose a lower workload and be more accessible than eye gaze interaction methods. However, it is noteworthy that no significant differences were found between Daria and Alexa, indicating that the two voice-based interaction methods performed similarly in terms of workload.

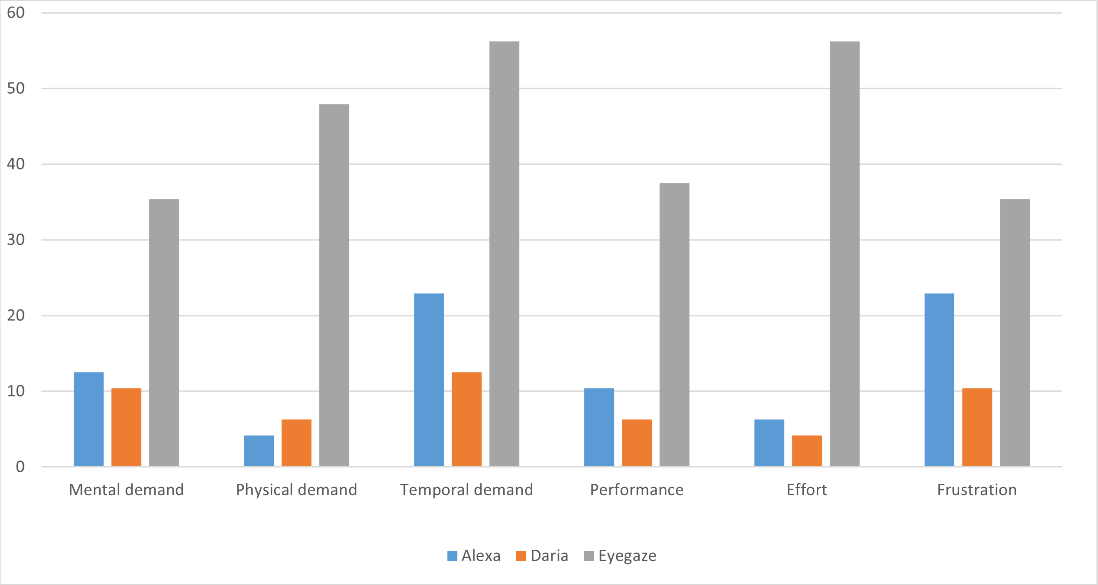


Figure 9.3: NASA-TLX workload

9.3.3 Task success rate

When the interaction with the system was successful, that is, when the action was performed by the SVA, this was considered a successful interaction. Given that it was difficult to control the number of attempts using the eye gaze interaction, we did not fo-

Category	**System**	**Mean rank**	**Chi-square**	**p-value**	**p-value assessment**
Mental demand	Daria	2.25	3.50	0.174	Not significant
	Eye gaze	1.63			
	Alexa	2.13			
Physical demand	Daria	2.31	11.47	0.003	**Significant**
	Eye gaze	1.25			
	Alexa	2.44			
Temporal demand	Daria	2.44	5.85	0.054	Not significant
	Eye gaze	1.38			
	Alexa	2.19			
Performance	Daria	2.31	6.42	0.040	**Significant**
	Eye gaze	1.44			
	Alexa	2.25			
Effort	Daria	2.50	8.96	0.011	**Significant**
	Eye gaze	1.25			
	Alexa	2.25			
Frustration	Daria	2.38	2.82	0.244	Not significant
	Eye gaze	1.69			
	Alexa	1.94			

Table 9.2: Workload: significance between three interaction systems. Friedman's analysis of variance was used for overall comparison. The significance level was 0.05..

cus on the number of attempts; if the device detected the command, no matter after how many attempts, this was counted as a success. Given that there were five commands and eight users, the total number of successful attempts was 40 if all commands were successful for each user. For interacting with Daria, 38 interactions were successful out of 40. For Alexa, 33 were successful, and for eye gaze, 13 were successful.

Category	System	Mean	p-value	p-value assessment
Mental demand	Daria Eye gaze	93.75 43.75	0.223	Not significant
	Daria Alexa	93.75 77.08	0.317	Not significant
	Eye gaze Alexa	43.75 77.08	0.223	Not significant
Physical demand	Daria Eye gaze	93.75 52.08	0.025	**Significant**
	Daria Alexa	93.75 95.83	0.317	Not significant
	Eye gaze Alexa	52.08 95.83	0.024	**Significant**
Temporal demand	Daria Eye gaze	89.58 64.58	0.020	**Significant**
	Daria Alexa	89.58 87.50	0.285	Not significant
	Eye gaze Alexa	64.58 87.50	0.205	Not significant
Performance	Daria Eye gaze	93.75 62.50	0.041	**Significant**
	Daria Alexa	93.75 89.58	0.655	Not significant
	Eye gaze Alexa	62.50 89.58	0.242	Not significant
Effort	Daria Eye gaze	95.83 43.75	0.018	**Significant**
	Daria Alexa	95.83 93.74	0.564	Not significant
	Eye gaze Alexa	43.75 93.74	0.042	**Significant**
Frustration	Daria Eye gaze	89.58 64.58	0.068	Not significant
	Daria Alexa	89.58 77.08	0.461	Not significant
	Eye gaze Alexa	64.58 77.08	0.498	Not significant

Table 9.3: Pairwise workload statistical analysis and significance. The significance level was 0.05..

9.3.4 Preferences

Participants were asked to share their preferences across the three systems. Five participants preferred the Daria system, citing its ability to accurately understand their utterances and the ease of use. One participant specifically appreciated that it did not require pronouncing challenging letters, such as ‘R’.

Two participants favoured Alexa, including one who had moderate and one who had mild dysarthria, explaining that they were comfortable with it and capable of articulating words and sentences using this system. However, for some participants, using Alexa was not preferred because continuous speech was found to be tiring, particularly for those who had more severe forms of dysarthria in which prolonged speaking can be physically demanding.

Meanwhile, one participant (moderate dysarthria) preferred the eye gaze interaction, valuing the option to interact without the need to use voice. However, two participants (P5, P1) reported that the eye gaze system was not their preferred choice because of the discomfort caused by the laser from the tracker, which was uncomfortable or even painful for their eyes. In addition, the effort required to accurately control the eye gaze system was mentioned.

This feedback sheds light on the diverse experiences and preferences of participants for each interaction modality. A breakdown of participants’ preferences and their respective diagnoses is provided in Table 9.4.

9.4 Discussion

This study contributes significantly to the field of HCI and the accessibility of HCI technologies. By examining three distinct HCI methods—direct speech commands, nonverbal voice cues via the Daria system and eye gaze interactions—our study not only revealed their effectiveness, usability and participant preferences but also provided

Table 9.4: Participant preferences for assistive communication systems

Preference	Participant	Diagnosis
Alexa	P1	Mild
	P5	Moderate
Daria	P2	Mild
	P3	Mild
	P6	Severe
	P7	Severe
	P8	Severe
Eye gaze	P4	Moderate

a comprehensive comparison of these methods. These findings are highly valuable for future researchers in this field and contribute to the development of more inclusive and accessible communication technologies.

The diverse preferences expressed by the participants in our study revealed a nuanced picture of interactions with SVAs. Our findings indicate a preference for the Daria system among most participants, which was attributed to its ease of use and adeptness at understanding commands. This preference was particularly notable among participants who had severe dysarthria, suggesting that Daria's design is well suited to users who have significant speech impairments. This finding shows how using nonverbal voice cues, which is within users' capabilities, aligns with Wobbrock's principles, specifically, ability-based design principles [31], for creating systems in accordance with the strengths and capabilities of users, thereby enhancing accessibility. However, Alexa was preferred by participants who had milder forms of dysarthria, indicating its effectiveness for users who can articulate clearer speech patterns. Eye gaze interaction was uniquely valued by one participant who had moderate dysarthria, highlighting its potential as an alternative communication method for those who find voice-based interaction challenging.

These preferences correlate with our findings on usability, which indicate that SVAs which use verbal or nonverbal commands are more usable than those using eye gaze interactions. This increased usability arises from the relative ease of speaking and the ability of the device to understand speech. In addition, prior studies, such as that of [32, 33], have indicated that voice interactions are closely aligned with natural human communication patterns. Further, users who have dysarthria prefer to use their voice to the maximum extent. However, this finding contradicts that of [216], who found that participants rated the usability of eye gaze interactions with SVAs as exceptional, providing an average SUS score of 92.5. However, the limited scope of this study, which focused on a single user who had a disability, raises questions about the generalizability of the findings. Another study, [217], found that users who had a motor disability (but provided no information on their speech ability) gave eye gaze interactions an average SUS score of 78.54, which is higher than our result but lower than that of [216]. A broader participant base in future studies could offer more comprehensive insights into the usability of eye gaze systems.

The alignment of user preferences with usability scores in our study resonates with the technology acceptance model [218] and the unified theory of acceptance and usage of technology [219]. These models emphasise ease of use and effort expectancy as critical factors in technology adoption. This was confirmed by our findings, in which participants gravitated towards systems that offered greater ease of use and less effort, reflecting a natural inclination to technologies that align with their individual abilities and communication preferences.

In addition, interacting through voice is likely to be a more intuitive and natural method [220], even for individuals who have impaired speech capabilities [33]. However, it is important to consider the influence of participants' lack of prior experience with the eye-tracking device and the Daria system. None of the participants had previously used these systems, introducing significant factors that may have affected system usability, including the intuitiveness required to use the system and the learning curve associated with unfamiliar technologies [221]. Although the Daria system, which uses nonverbal voice commands, relies on the inherent familiarity most individuals have with vocal communication, thus making it more intuitive, eye-tracking systems may require a steeper learning curve because of their unconventional interaction mode. Therefore, these factors may influence the overall usability of each system for first-time users [222], underscoring the importance of considering the novelty and intuitiveness of HCI technologies in their evaluation.

These findings are further supported by our workload results, which offer important insights into the experience of users who have dysarthria when interacting with various technologies. The data show that eye gaze interactions involve considerably more effort across several dimensions. This higher level of workload suggests that although eye gaze interactions remain a viable option for individuals with dysarthria, especially those who have severe cases, the extensive demands may affect this technology's practicality and acceptance for long-term use. This has been noted in prior studies [223], suggesting that the burdensome nature of eye gaze interactions may extend to other populations who have similar challenges.

Further, the perceived effectiveness and reduced effort associated with voice-based interactions suggest a higher likelihood of long-term acceptance and use. Their ease of use and lower physical and mental demands position these methods as more sustainable and practical for individuals with dysarthria. This aligns with the broader goal of ATs, which is to enhance quality of life through user-friendly and efficient solutions [224]. Therefore, our findings underscore the critical need to consider workload and user effort as key factors in the design and implementation of HCI methods, specifically, in terms of ATs for individuals with dysarthria.

These findings were also reflected in the success rate, which was higher for interactions that were more usable and required less effort. Participants who had varying levels of dysarthria severity preferred using their voices to interact with SVAs. This finding aligns with that of [32, 33], who found that users prefer using their voices as much as possible.

Our study offers valuable insights into the interaction preferences of individuals with dysarthria. It serves as a foundation for further research in this area. To build on these initial findings, future studies could benefit from exploring a wider range of participant experiences, enhancing the generalisability and depth of the research. In addition, investigating the learning curve associated with various interaction systems, especially for users who are new to eye gaze technology, would provide a more comprehensive understanding of how user familiarity affects effectiveness.

9.5 Conclusion

This study uncovered the effectiveness and user experience of various interaction methods used by individuals with dysarthria. It also revealed users' preferences among these techniques. The preferred interaction for most users was found to be the Daria system, especially for those who had severe dysarthria, owing to its ease of use and effective command recognition. This underlines the potential of nonverbal voice cues in enhancing accessibility for users who have significant speech impairments. In addition, the study identified challenges with eye gaze interaction, especially in terms of usability and workload, and noted that it required more effort than the other two methods.

However, the study revealed that participants who had milder forms of dysarthria favoured voice-activated systems, such as Alexa, indicating their suitability for those who can articulate clearer speech patterns. This preference emphasises the need for HCI technologies to cater to varying levels of speech ability.

Finally, this chapter was instrumental in understanding how individuals with dysarthria interact with various types of SVAs and the factors that influence their usability and user experience. These insights are invaluable for guiding future technology development to better meet the diverse communication needs of individuals with dysarthria. For individuals who have mild or moderate dysarthria and can still use Alexa and for those who have mild, moderate and severe dysarthria and can use Daria, as well as for those who have severe cases and cannot use voice-based systems such as Alexa or Daria, eye trackers emerged as a crucial alternative. In addition, future studies will focus on further investigating these methods of interaction by using larger groups and collecting more comprehensive data. This will not only aid in understanding all methods of interaction more thoroughly but also assist in refining and expanding the systems.

Chapter 10

Discussion and Conclusion

This thesis presented work exploring the accessibility of SVAs for individuals with dysarthria. It did so by introducing the technique of using nonverbal voice cues to enable individuals with dysarthria to interact with SVAs. This chapter concludes the thesis by revisiting the research questions laid out in Chapter 1. The results of this research, including the literature review in Chapter 2 and the studies in Chapters 4, 7, 8 and 9, are combined and discussed to show how they answered the research questions.

This study aimed to determine the efficacy of interaction modalities, including nonverbal voice cues, in controlling SVAs for individuals with dysarthria through four research questions:

RQ1: How do individuals with dysarthria currently use smart voice assistants, and what are their experiences with these devices?

RQ2: Can a standardised vocabulary that aligns with their unique speech capabilities and the range of sounds they can produce be developed for individuals with dysarthria?

RQ3: How does the use of nonverbal voice cue interaction techniques affect the user experience and usability of smart voice assistants?

RQ3.1: How memorable are nonverbal voice cues for users?

RQ3.2: How does usability, user experience and workload differ between the proposed interaction technique and verbal interaction?

RQ4: What is the impact of allowing customisation rather than standardisation on the interaction?

Three studies were conducted to answer the research questions. Chapter 4 described the interview study (Study 1). This chapter aimed to explore the challenges encountered by individuals with dysarthria when using SVAs and to obtain feedback on the proposed interaction modality involving nonverbal voice cues. The results of this study led to (a) the design of the interaction framework presented in Chapter 5 and (b) the design of the Daria system described in Chapter 6. After designing the system, we conducted a preliminary study (Study 2), which served as an initial test to gather early feedback and make necessary adjustments. Following these steps, we conducted our between-subject study (Study 3), described in Chapters 7, 8 and 9. These chapters investigated the usability and user experience of the nonverbal voice cue interaction and provided comparative insights into user experiences with Daria. To further investigate user preferences and compare this interaction method with another interaction modality, we extended the study to compare the usability of Daria and verbal interaction with eye gaze interaction.

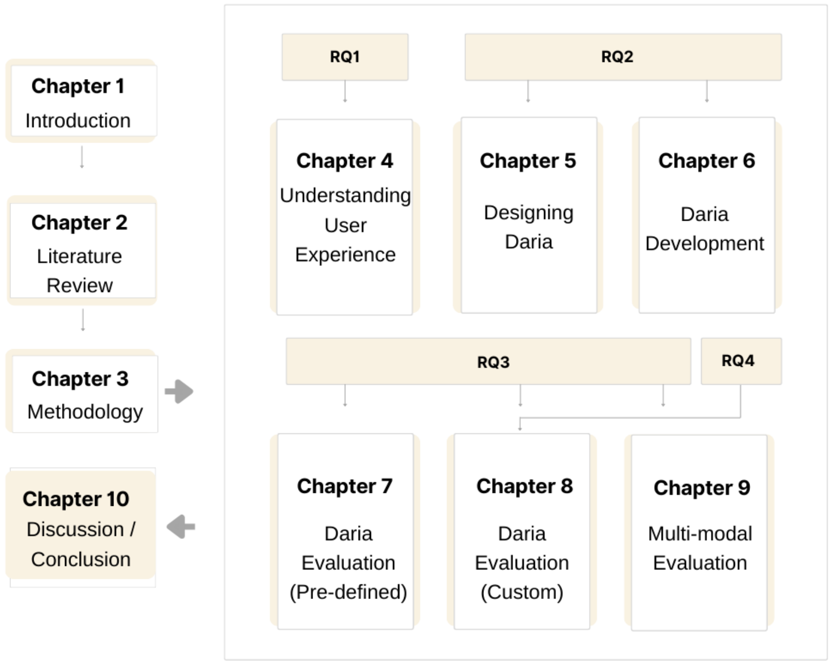


Figure 10.1: Alignment of thesis chapters with corresponding research questions

10.1 Review of research questions and contribution

RQ1: How do individuals with dysarthria currently use smart voice assistants, and what are their experiences with these devices?

To explore the experiences of individuals with dysarthria using SVAs, we initially conducted a literature review. This review, presented in Chapter 2, aimed to identify the current state of the literature and understand the gaps in the field. To address specific questions about how individuals use SVAs, we detailed an interview in Chapter 4. Moreover, we conducted a study to observe individuals' use of off-the-shelf SVAs, presented in Chapters 7 to 9.

In the literature review, we analysed prior studies on this topic, gaining insights into their outcomes. Our findings indicated a limited exploration of how individuals with dysarthria interact with SVAs. In addition, studies on ASR have consistently shown lower accuracy rates for individuals with dysarthria compared with those without. Contributing factors include the difficulties in designing systems that accurately understand dysarthric speech and the challenges in collecting recordings from individuals with dysarthria, since longer sessions can increase their fatigue.

After presenting the findings from our literature review in Chapter 2, which revealed significant gaps in the current understanding of SVA usage by individuals with dysarthria, we designed our studies to directly address these shortcomings. The limited exploration of individuals with dysarthria's interactions with SVAs and the challenges in ASR technology, as highlighted in prior studies, formed the basis of our investigative approach. Our Study 1, detailed in Chapter 4, was specifically structured to delve into these under-researched areas. By conducting interviews with individuals who had dysarthria, we aimed to not only address the research gap but also provide empirical data that could lead to the development of more inclusive and effective voice assistant technologies. This methodological choice was a direct response to the need for user-centred research in this field, a gap identified in the literature review.

The interviews not only addressed the gaps identified by our systematic review but also responded to the inquiries raised by Cave [225]. In his scoping review, Cave highlighted the absence of studies investigating the specific commands used by individuals who have ALS, a condition associated with dysarthria, when interacting with SVAs. Moreover, this review underscored the lack of studies examining the use of SVAs by this particular user group, including explorations of their expectations. One notable point in Cave's review is the omission of detailed information about the condition of users in prior studies—an aspect that our study also addressed.

In Study 1, we identified a need for individuals to independently use these devices, noting that users are restricted to commands they can articulate clearly. Our findings pinpoint major issues, such as timing out, fatigue and the specific nature of speech impairment. However, the insights we gained extend beyond Study 1. The studies described in Chapters 7 to 9 also contribute significantly to answering our first research question (RQ1). In these studies, we evaluated the use of off-the-shelf SVAs, such as Alexa, by individuals with dysarthria. We observed how these individuals interact with SVAs and conducted an analysis focused on the success rate—specifically, how often the device correctly understood their speech out of five attempts.

In addition to the findings mentioned, our study revealed that individuals with dysarthria often rely on intermediary devices to facilitate their use of SVAs. This reliance on additional technology underscores a critical need for independence in the use of SVAs. The intermediary devices, although helpful, introduce an extra step in the interaction process, potentially complicating the user experience and possibly exacerbating issues such as timing out and fatigue. This reliance on intermediary devices reflects the current limitations of SVAs in directly accommodating the unique speech patterns of individuals with dysarthria.

Our studies, particularly those described in Chapters 7 to 9, demonstrate that although off-the-shelf SVAs such as Alexa are capable of recognising speech, their effectiveness declines with the complexity of speech impairments. This restricts users to commands they can articulate clearly. In addition, this requires the use of intermediary devices, which can limit the spontaneity and ease of use, because the additional step of using an intermediary device may affect the user experience.

In light of these observations, future SVA designs may aim to reduce the need for intermediary devices by being more accommodating and sensitive to diverse speech impairments. Enhancing the direct interaction capability of SVAs would empower users who have dysarthria to use these devices more independently and efficiently. This could involve using various interaction techniques. Such advancements in SVA technology would not only enhance autonomy and ease of use for individuals with dysarthria but also represent a significant step towards creating truly inclusive and accessible digital environments.

These results highlighted a distinct experience for people who have dysarthria compared with those without. When comparing success rates, we found that device performance deteriorated in line with the severity of dysarthria. Notably, the success rate was less than 50% for moderate to severe cases. However, there were differences in the results for mild cases between the studies, which had a variance of approximately 30%.

Consequently, we conclude that although Alexa may not be effective for moderate and severe cases of dysarthria, it could be useful for those who have mild cases. This finding is comparable to the Ballatie [22] study, which tested recordings of moderate dysarthria using three virtual assistants and found that comprehension by the device varied in accordance with the severity of the condition; however, moderate dysarthria could still be understood by these devices.

RQ2: Can a standardised vocabulary that aligns with their unique speech capabilities and the range of sounds they can produce be developed for individuals with dysarthria?

A key part of the design of the nonverbal voice cue interaction was developing a vocabulary for interactions with the system. In addressing RQ2, this thesis navigated the design of a standardised vocabulary for individuals with dysarthria, as detailed in Chapters 4 and 5. Chapter 4 laid the groundwork by exploring the unique challenges and preferences of individuals with dysarthria in using SVAs. Being asked about their preferences with regard to voices underscored the need for alternative interaction methods tailored to their capabilities. The insights gained from this qualitative study were pivotal in informing the design choices that followed.

Building on these findings, Chapter 5 introduced ‘Daria’, a nonverbal voice cue system designed specifically for individuals with dysarthria. This chapter exemplified a user-centred design process, in which the chosen voice cues not only were aligned with users' preferences but also ensured ease of production and clear acoustic distinction. In line with ability-based design principles [31], this approach effectively utilised the unique abilities of individuals who had dysarthria. By concentrating on their capabilities, such as producing specific vowels and nasal sounds, we crafted a more inclusive and powerful interaction technique. This approach aligned with Wobbrock's [31] advocacy for creating systems in accordance with the strengths and capabilities of users, particularly those who have diverse abilities, thereby enhancing accessibility and empowerment.

The work in Chapters 4 and 5 focused on the creation of a new vocabulary of non-verbal voice cues. This vocabulary considered the varied speech capabilities of users and offered a customisable yet standardised method for interacting with SVAs. The sound–action mapping framework established in this thesis, which integrated natural mapping principles and everyday life metaphors, represents an innovative approach to intuitive and memorable interactions for users who have dysarthria. This vocabulary could be used as the basis of a command language in which these sounds could be combined.

In summary, this thesis addressed RQ2 and expanded the discourse surrounding accessible technology. By acknowledging and catering to the unique speech capabilities of individuals with dysarthria, this study makes a significant contribution to enhancing their autonomy and engagement with technology. However, it is important to note that this work has only tested and generated a limited vocabulary, which may not encompass the full range of sounds or communication needs of all individuals with dysarthria. The framework and vocabulary developed provide a foundation for future research and development in creating more inclusive and empathetic technological solutions. Further expansion and refinement of the vocabulary is necessary to ensure its broader applicability and effectiveness. In addition, the practical application and effectiveness of this vocabulary were further demonstrated in subsequent chapters, showcasing its real-world applicability and impact.

RQ3: How does the use of nonverbal voice cue interaction techniques affect the user experience and usability of smart voice assistants?

One of the goals of this study was to develop an alternative usable interaction technique. The results of the preliminary study in Chapter 6, and the study focusing on different groups in Chapters 7and 8, demonstrated that nonverbal voice cues are more effective for individuals who have moderate and severe dysarthria. Also, this technique is preferred because of its simplicity and ease of use. This finding suggests a potential re-evaluation of only having verbal commands in SVA systems, and suggests the possibility of exploring new avenues in designing voice interaction technologies that are more inclusive.

In terms of the high success rate, the study highlights a significant finding: nonverbal voice cues had higher success rates in recognising commands from individuals with dysarthria, particularly in moderate to severe cases. This is a notable difference from standard voice recognition models that rely on verbal commands. Nonverbal cues are often shorter for individuals with dysarthria to articulate clearly. This clarity leads to higher levels of recognition accuracy, making the interaction more accessible and less frustrating for the user. The study not only underscores the general effectiveness of nonverbal voice cues but also highlights that certain specific sounds perform better than others. For instance, sounds that are easier to produce and require less articulation, such as vowel sounds or hums, had notably higher rates of recognition. This was attributed to their simplicity and the lower degree of motor control required to produce them, which is particularly beneficial for individuals with dysarthria and who may struggle with articulation. Specifically, our results showed that the highest recognised nonverbal voice cues were /ɑ/ and /ŋ/. However, this study also revealed a limitation in the current system's ability to consistently recognise a wider range of nonverbal cues. Some sounds were less accurately recognised, indicating a need for further system optimisation to enhance recognition capabilities across a broader spectrum of nonverbal sounds. Future studies could focus on optimising the system to better recognise all sounds.

RQ3.1: How memorable are nonverbal voice cues for users?

Results showed that the first group, which interacted using predefined mappings, exhibited significantly higher rates of recall (80%) for the commands and actions than the other group, which created its own mappings (28%). This suggests that when a system offers a set of established voice-command mappings, users may focus more on learning and remembering these associations than on the creation process. This relates to the concept of mental spaces [226], which involves a cognitive process in which users construct understanding by connecting new information to existing knowledge. By designing our mappings to align with mental spaces that were familiar to users, we enhanced the memorability of the interactions [160, 226]. These mappings, systematically developed, had a logical or intuitive structure that potentially resonated more naturally with the users' way of thinking. This could have contributed to the better memorability and smoother interaction experience.

However, customisation required users to create their own mental spaces by merging elements from various domains to create a new, emergent structure. This can be more challenging and less intuitive, leading to lower recall efficiency. Additionally, in our studies, several participants indicated that they approached the mapping process randomly when creating their own voice-command associations. This random approach can exacerbate difficulties in recall because it lacks the deliberate, logical structure that aids in memorisation [197, 198, 227]. The cognitive effort required to establish and remember self-created associations between voice cues and actions seems to have outweighed the potential benefits of customisation. This is a critical insight for designers because it suggests that even if customisation is provided to users, it should not offer too much freedom without adequate guidance or structure to help create systematic and meaningful mapping. For example, users could be provided with guidance or training on how to create a mapping that makes sense to them [197, 198].

Although the comparison of memorability between the two approaches was limited, this finding could contradict other work, such as the study by Nacenta et al. testing the memorability of hand gestures [228]. In their work, user-defined gestures were found to be more memorable than predesigned or random gestures. The study noted memorability differences, as users perceived user-defined gestures as requiring less time to learn. This difference could be attributed to the nature of the interactions, namely, gestures versus voice cues. Moreover, in our study, the groups of participants were exposed to different interaction techniques (nonverbal voice cues) and their responses were compared. This between-subjects design ensured that each participant experienced only one type of interaction, reducing the potential for learning or fatigue effects from one condition to influence the performance in another. However, it also meant that individual differences between participants could have had a more significant impact on the results, because each condition was tested using a different group of people. Nacenta et al.'s within-subjects design involved the same participants experiencing all gesture types (user-defined, predesigned and random). This approach allowed for a direct comparison of each condition within the same individual, controlling for inter-participant variability. Participants in their study could provide a more direct comparison of their experiences and preferences for each gesture type because they were familiar with all of them. However, this design can introduce learning effects, in which the experience in one condition might influence performance in subsequent conditions.

Future work could involve comparing a third approach, which would involve giving users a predefined but randomly mapped list. This would be compared with logically mapped and customised options. Such a comparison would provide deeper insight into what exactly affects memorability.

RQ3.2: How does usability, user experience and workload differ between the proposed interaction technique and verbal interaction?

The studies collectively suggest, according to the SUS results (SUS score = 85.75; see Section 7.3), that the nonverbal voice interaction technique, exemplified by the Daria system, offers a more accessible and user-friendly alternative to conventional verbal interactions for users with dysarthria (SUS score = 71.5). One of the key advantages of this technique is its reduced speech complexity. Nonverbal cues are simpler and easier to use than structured speech, as evidenced in our studies, making them particularly suitable for users who have dysarthria and may face challenges in articulation, breath control and pronunciation.

In addition, the Daria system showcased consistent usability across various levels of dysarthria severity. This contrasted with the performance of systems such as Alexa, which declined according to the increased severity of speech impairment. Such consistency underscores the adaptability and reliability of nonverbal interactions for a broader range of users. In Study 3 with the predefined option, Daria's accuracy rates for users who had mild, moderate and severe dysarthria were 72%, 64% and 66%, respectively (see Section 7.3). Alexa's performance varied more significantly in accordance with dysarthria severity: 82.67% for mild, 33.6% for moderate and a notable drop to 24% for severe cases. The custom group (see Section 8.1.3) reinforced these patterns: Daria maintained similar levels of accuracy across the severity categories (mild: 68%; moderate: 62%; severe: 80%), whereas Alexa's performance is (mild: 57.33%; moderate: 41.6%; severe: 38/%).

Moreover, the study results indicate a preference among users for pre-mapped voice and action associations, given that 12 out of 20 participants favoured this approach. This preference stemmed from the importance users placed on systems that emphasised and leveraged their capabilities rather than focusing on their limitations. The preference for pre-mapped voice and action mappings indicates that users value simplicity and directness in interactions, which are crucial for usability. The results of Study 3 using the predefined group extend these findings by showing the Daria system's consistent usability across various levels of dysarthria severity, contrasting with the varied performance of Alexa. This suggests that nonverbal interactions offer a more uniform and potentially more accessible user experience. The results of Study 3 using the customised group highlight that although customisation is theoretically appealing, it does not necessarily translate into enhanced usability. Users showed a preference for the straightforwardness of pre-mapped systems, emphasising the importance of immediate ease of use.

In terms of user experience, the Daria system was evaluated positively in several dimensions, including speed, habitability and likability. Participants found the system to be fast, habitable and likable, indicating a high level of satisfaction with the interaction process. When compared with Alexa, the likability dimension showed a significant difference (see Sections 7.3 and 8.1.3), indicating that Daria was more likable. In addition, 13 out of 20 participants preferred Daria over Alexa. This preference was particularly notable among users who had moderate to severe dysarthria, suggesting that nonverbal cues were more accommodating to their needs. However, there was a significant difference in the annoyance dimension, in which Daria scored higher. This could be because of a preference for customisation or instances in which the system did not detect commands. When examining this dimension with Group 2, we found no significant difference between Daria and Alexa in this domain, which could indicate that the frustration was due to the missing customisation option. The impact of allowing customisation versus standardisation on interactions is further explored in the later section that discusses RQ4.

The study's comparison of Alexa and eye gaze technologies provides essential benchmarks for evaluating the Daria system. Although Alexa was more effective in mild cases of dysarthria, its performance dropped significantly as severity increased, highlighting the need for more adaptive ASR systems. Eye gaze technology, although a viable alternative, was shown to require a higher level of cognitive and physical workload, making it less suitable for prolonged use, especially in severe cases.

We also covered the workload of the tasks through the NASA-TLX results, offering a comprehensive view of the user workload associated with each interaction method. Interestingly, unlike our hypothesis, there was no significant difference in workload between the two interaction methods (verbal and nonverbal voice cues). Although we hypothesised that shorter commands would require less workload, the equivalence in workload highlighted the effective design of nonverbal voice cue systems . It suggested that these systems have been designed to be as intuitive and user friendly as traditional verbal interaction systems. This observation might also indicate that verbal and nonverbal interactions align well with the natural interaction dynamics of users. It suggests that nonverbal cues, although different from conventional speech, are still within the comfort zone of users, allowing them to interact with the technology without experiencing additional cognitive or physical strain. The similarity in workload may also indicate that the two systems are balanced in terms of cognitive demand. Despite the differences in interaction style, neither system overwhelms the user, suggesting a well-considered balance between functionality and usability.

When comparing the workload with eye gaze interaction (see Section 9.3), we found that there was a significant difference in the physical, temporal, performance and effort dimensions. The reason, as mentioned earlier, is that interacting through voice is likely to be a more intuitive and natural method [220], even for individuals who have impaired speech capabilities [33]. However, it is important to consider the influence of participants' lack of prior experience with the eye-tracking device and the Daria system. None of the participants had previously used these systems, introducing significant factors that may have affected system usability—specifically, the intuitiveness required to use the system and the learning curve associated with unfamiliar technologies [221]. However, there was a significant difference in physical demand between the predefined and customisation groups. This was explained by the additional steps required before using the system. This is further explained in the next section.

RQ4: What is the impact of allowing customisation rather than standardisation on the interaction?

Various findings on customisation emerged from this work, as discussed under the previous research questions. The main finding is the physical demand that accompanies the generation of a user's own custom voice–action mapping, as demonstrated in the significance difference analysis presented in Section 8.2. For instance, the process of customisation could involve speech, fine motor skills and sustained attention. These actions can be particularly fatiguing for individuals with dysarthria who might already be dealing with muscle fatigue and coordination challenges. The physical exertion required in such scenarios is not limited to speech production but extends to the overall interaction with the device, including touchscreen navigation or button presses.

These studies collectively indicate a strong preference among users with dysarthria for pre-mapped voice and action mappings. This preference suggests that these users may favour a more standardised approach to voice interaction in which they are not required to engage in the potentially complex and demanding task of customising their interactions.

In conclusion, although there is a strong indication that users who have dysarthria might prefer pre-mapping because of the reduced physical demand, the benefits of customisation in enhancing user experience, accessibility and empowerment cannot be overlooked. This highlights the need for future studies to explore the balance between providing a user-friendly, standardised system and allowing sufficient customisation to meet the varied needs and preferences of individual users. Achieving this balance is crucial for enhancing the overall interaction experience for individuals with dysarthria using SVAs.

10.2 Recommendations for System Designers

Based on the findings of this thesis, we present the following recommendations for designers of nonverbal voice cue interaction systems, especially for use by those who have dysarthria:

- Meaningful voice cue mapping: Designers should focus on developing meaningful and intuitive mappings, which are built on the natural associations between sounds and actions. This will result in a system that is easy for users to remember and use. For instance, in the mapping presented in this thesis, the humming sound is linked to the music function. This means that when users hum, the system recognizes this sound and plays music. This example demonstrates how associating the sounds with the actions can create an easy-to-use interface that enhances the user experience by leveraging natural sound-action relationships.

- Guidance for Customisation: For systems that are customisable, users should be supported with clear instructions for creating meaningful and intuitive mappings. This could be done by incorporating information that makes users aware of the importance of the logical mapping is likely to significantly enhance user experience. For example, user manuals, tutorials or other forms of instructional content. By offering these resources, users can better understand how to customize the system to their preferences and needs. It could also help the users to maximise the benefits and usability of the system.

- Simplicity in design: Some users prefer pre-mapped systems or plug-and-play solutions that do not require additional setup. Accordingly, designers should develop systems that are ready and easy to use. This means the user can use the system without complex configurations or multiple steps to prepare the system. It needs to be as straightforward as possible, which helps to reduce user frustration.

- Balancing customisation and usability: Although customisation is an important feature, it is crucial to balance this with usability. Designers should aim to simplify the customisation process as much as possible. This could involve creating user-friendly guides that help users through the customization process. An example is providing default settings that users can easily adjust. This means allowing customization without sacrificing ease of use.

- Physical effort consideration: The results highlight that customisation requires higher levels of physical effort. While designing interaction systems for individuals with dysarthria, designers should strive to minimize the physical effort needed to customise the system. For example, the configuration process should be streamlined to avoid multiple steps, making it as straightforward as possible. It should also be designed so that users do not require the assistance of another individual or, if necessary, keep such assistance to a minimum. Additionally, for example, the customization should not rely heavily on voice input, as this may be challenging for individuals with dysarthria.

- Long-term user engagement: Designers should explore methods to maintain usability and user satisfaction over extended periods. Designers should, therefore, study how users interact with a system over time and introduce adjustments as needed to ensure sustained effectiveness and user engagement.This could be achieved through several strategies, such as conducting longitudinal studies to observe how users' interactions and satisfaction levels change over time.

- Testing and feedback: Systems should be tested by diverse users, with different types,levels of dysarthria. This diversity in testing is crucial to ensure that the system is accessible and effective for a broad range of users. Incorporating user feedback should enhance the system's usability and user experience. Feedback can be collected through different approaches, such as iterative design, surveys, and usability testing.

These recommendations will allow system designers to develop more effective, user-friendly and inclusive voice-assisted technologies that better serve the needs of individuals with dysarthria.

10.3 Limitations and future work

This study presents several limitations that pave the way for future research. First, the current study was constrained by a relatively small vocabulary, which may not have fully addressed the wide range of sounds or communication needs of all individuals with dysarthria. Future studies should aim to generate a larger, more comprehensive vocabulary, potentially including multiple voice sounds, to broaden the system's inclusivity and applicability. Moreover, future studies should consider exploring the effect of different languages in the interaction. Second, one of the limitations is the lack of multiple iterations to improve the design. Future work should consider developing and testing new models to enhance performance and better address the needs of individuals with dysarthria. Third, the testing of the system focused primarily on individual voice cues, leaving the exploration of combinations of voice cues as a future area of study. Investigating how users interact with a system using combinations of voice cues could more closely mimic natural speech patterns and offer richer interaction possibilities. Additionally, this thesis has not released the experimental data as there are plans for further analysis and publications, but further work should be careful when considering the statistical tests used in this thesis. Future work should ensure that the chosen statistical methods are appropriate for the data characteristics and research design, and consider the limitations of non-parametric tests in terms of power and complexity handling. For research exploring a new area, as this thesis did, it is arguably legitimate to over-test (also called the multiple comparisons problem) while we look for possible noteworthy but provisional results (see e.g. tables 7.4, 7.5, 7.10, 8.3, 8.8, 8.9, 9.2, 9.3). Subsequently, for future work, one should put the statistical testing on a more rigorous basis now this thesis provides data to better formulate explicit null hypotheses *before* undertaking new experiments. Cairns [229] is an excellent reference on the issue and solutions.

Moreover, most testing in this thesis was conducted in controlled environments, which may not accurately reflect users' everyday contexts. Long-term testing in users' homes is necessary for a more comprehensive understanding of aspects such as memorability and usability in real-world settings. Such in situ studies would provide insights into how systems integrate into the daily lives of individuals with dysarthria, revealing new challenges and opportunities for system refinement over time. Another limitation of this work is that most of our participants were male, which could have an impact on the results.

10.4 Conclusion

This thesis investigated the efficacy of various interaction modalities, including nonverbal voice cues, for controlling SVAs by individuals with dysarthria. It aimed to enhance the accessibility of interactions with SVAs and establish a framework for designing effective interaction methods. The study comprehensively explored how individuals with dysarthria use SVAs, developed a standardised vocabulary tailored to their unique speech capabilities and evaluated the impact of nonverbal voice cues on user experience and system usability. The findings demonstrate a pronounced viability and preference for this interaction technique in individuals who have moderate to severe dysarthria, highlighting its simplicity, ease of use and high rates of success in system recognition. These findings underscore the potential for a wider adoption of nonverbal voice systems to improve accessibility for those who have speech impairments.

Bibliography

[1] Aisha Jaddoh, Fernando Loizides, Omer Rana, and Yasir Ahmed Syed. Interacting with smart virtual assistants for individuals with dysarthria: A comparative study on usability and user preferences. *Applied Sciences*, 14(4), 2024.

[2] Aisha Jaddoh, Fernando Loizides, and Omer Rana. Interaction between people with dysarthria and speech recognition systems: A review. *Assistive Technology*, pages 1–9, 2022.

[3] Aisha Jaddoh, Fernando Loizides, Jimin Lee, and Omer Rana. An interaction framework for designing systems for virtual home assistants and people with dysarthria. *Universal Access in the Information Society*, pages 1–13, 2023.

[4] Aisha Jaddoh, Fernando Loizides, and Omer Rana. Non-verbal interaction with virtual home assistants for people with dysarthria. *The Journal on Technology and Persons with Disabilities*, page 71, 2021.

[5] Statista.com. Number of households with smart home products and services in use worldwide from 2017 to 2025.

[6] Jannette Maciej and Mark Vollrath. Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, 41(5):924–930, 2009.

[7] Wolfgang Minker, Udo Haiber, Paul Heisterkamp, and Sven Scheible. The seneca spoken language dialogue system. *Speech Communication*, 43(1-2):89–102, 2004.

[8] Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. Multimodal interaction in the car: combining speech and gestures on the steering wheel. In

Proceedings of the 4th international conference on automotive user interfaces and interactive vehicular applications, pages 155–162, 2012.

[9] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–23, 2018.

[10] Cathy Pearl. *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc.", 2016.

[11] Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. " siri talks at you" an empirical investigation of voice-activated personal assistant (vapa) usage by individuals who are blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 249–258, 2018.

[12] Alisha Pradhan, Kanika Mehta, and Leah Findlater. " accessibility came by accident" use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[13] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. " 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*, pages 1–12, 2017.

[14] Daniel J Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi. When speakers are all ears: Characterizing misactivations of iot smart speakers. *Proceedings on Privacy Enhancing Technologies*, 2020(4), 2020.

[15] Caroline Lövqvist, Maja Pinter, Montathar Faraon, and Victor Villavicencio. Crafting tasteful experiences: Designing artificial intelligence and voice user interfaces for home delivery contexts. In *Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference*, pages 175–189, 2022.

[16] Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki. End-to-end dysarthric speech recognition using multiple databases. In *ICASSP 2019-2019 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6395–6399. IEEE, 2019.

[17] Fabio Masina, Patrik Pluchino, Valeria Orso, Riccardo Ruggiero, Giulia Dainese, Ilaria Mameli, Stefania Volpato, Daniela Mapelli, and Luciano Gamberini. Voice actuated control systems (vacs) for accessible and assistive smart homes. a preliminary investigation on accessibility and user experience with disabled users. In *Ambient Assisted Living: Italian Forum 2019 10*, pages 153–160. Springer, 2021.

[18] Luigi De Russis and Fulvio Corno. On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, 5(3):163–172, 2019.

[19] Irene Calvo, Peppino Tropea, Mauro Viganò, Maria Scialla, Agnieszka B Cavalcante, Monika Grajzer, Marco Gilardone, and Massimo Corbo. Evaluation of an automatic speech recognition platform for dysarthric speech. *Folia Phoniatrica et Logopaedica*, 73(5):432–441, 2021.

[20] Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840. IEEE, 2019.

[21] Neethu Mariam Joy and S Umesh. Improving acoustic models in torgo dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):637–645, 2018.

[22] Fabio Ballati, Fulvio Corno, and Luigi De Russis. Assessing virtual assistant capabilities with italian dysarthric speech. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 93–101, 2018.

[23] Myungjong Kim, Younggwan Kim, Joohong Yoo, Jun Wang, and Hoirin Kim. Regularized speaker adaptation of kl-hmm for dysarthric speech recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(9):1581–1591, 2017.

[24] Myung Jong Kim, Jun Wang, and Hoirin Kim. Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model. In *INTERSPEECH*, pages 2671–2675, 2016.

[25] Woo Kyeong Seong, Nam Kyun Kim, Hun Kyu Ha, and Hong Kook Kim. A discriminative training method incorporating pronunciation variations for dysarthric automatic speech recognition. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–5. IEEE, 2016.

[26] Feifei Xiong, Jon Barker, and Heidi Christensen. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

[27] Jort F Gemmeke, Siddharth Sehgal, Stuart Cunningham, et al. Dysarthric vocal interfaces with minimal training data. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 248–253. IEEE, 2014.

[28] Jordan R Green, Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, et al. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Interspeech*, pages 4778–4782, 2021.

[29] Brian Luu, Bradley Hansberger, Mandy Chiu, Vinay Kumar Karigar Shivappa, and Kiran George. Scalable smart home interface using occipitalis semg detection and classification. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 1002–1008. IEEE, 2018.

[30] Beth M Ansel and Raymond D Kent. Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research*, 35(2):296–308, 1992.

[31] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)*, 3(3):1–27, 2011.

[32] Rupal Patel, Christopher Dromey, and Hans Kunov. Control of prosodic parameters by an individual with severe dysarthria. *Univ. of Toronto, Toronto, ON, Canada, Tech. Rep*, 1998.

[33] Linda Ferrier, Howard Shane, Holly Ballard, Tyler Carpenter, and Anne Benoit. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175, 1995.

[34] Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269, 1969.

[35] Joseph R Duffy. *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2012.

[36] Rupal Patel. Prosodic control in severe dysarthria. 2002.

[37] Kathryn M Yorkston, DR Beukelman, F Minifie, and S Sapir. Assessment of stress patterning. *The dysarthria: Physiology, acoustics, perception, management*, pages 131–162, 1984.

[38] Donald B Freed. *Motor speech disorders: diagnosis and treatment*. Plural Publishing, 2018.

[39] Ann Nordberg, Carmela Miniscalco, and Anette Lohmander. Consonant production and overall speech characteristics in school-aged children with cerebral palsy and speech impairment. *International journal of speech-language pathology*, 16(4):386–395, 2014.

[40] Frank Rudzicz. Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, 27(6):1163–1177, 2013.

[41] M Dhanalakshmi, TA Mariya Celin, T Nagarajan, and P Vijayalakshmi. Speech-input speech-output communication for dysarthric speakers using hmm-based speech recognition and adaptive synthesis system. *Circuits, Systems, and Signal Processing*, 37(2):674–703, 2018.

[42] Frank Rudzicz. Using articulatory likelihoods in the recognition of dysarthric speech. *Speech Communication*, 54(3):430–444, 2012.

[43] Nancy Pearl Solomon, Donald A Robin, and Erich S Luschei. Strength, endurance, and stability of the tongue and hand in parkinson disease. *Journal of Speech, Language, and Hearing Research*, 43(1):256–267, 2000.

[44] Justine V Goozée, Bruce E Murdoch, and Deborah G Theodoros. Physiological assessment of tongue function in dysarthria following traumatic brain injury. *Logopedics phoniatrics vocology*, 26(2):51–65, 2001.

[45] Kathryn M Yorkston, Edythe A Strand, and Mary RT Kennedy. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, 5(1):55–66, 1996.

[46] Kris Tjaden, Joan E Sussman, and Gregory E Wilding. Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis. *Journal of Speech, language, and hearing research*, 57(3):779–792, 2014.

[47] Kaila L Stipancic, Kira M Palmer, Hannah P Rowe, Yana Yunusova, James D Berry, and Jordan R Green. "you say severe, i say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research*, 64(12):4718–4735, 2021.

[48] Nick Miller, Emma Noble, Diana Jones, and David Burn. Life with communication changes in parkinson's disease. *Age and ageing*, 35(3):235–239, 2006.

[49] Sylvia Dickson, Rosaline S Barbour, Marian Brady, Alexander M Clark, and Gillian Paton. Patients' experiences of disruptions associated with post-stroke dysarthria. *International Journal of Language & Communication Disorders*, 43(2):135–153, 2008.

[50] Marian C Brady, Alexander M Clark, Sylvia Dickson, Gillian Paton, and Rosaline S Barbour. The impact of stroke-related dysarthria on social participation and implications for rehabilitation. *Disability and rehabilitation*, 33(3):178–186, 2011.

[51] Margaret Walshe and Nick Miller. Living with acquired dysarthria: the speaker's perspective. *Disability and rehabilitation*, 33(3):195–203, 2011.

[52] Janice Light. Toward a definition of communicative competence for individuals using augmentative and alternative communication systems. *Augmentative and alternative communication*, 5(2):137–144, 1989.

[53] Kris Tjaden, Deanna Rivera, Gregory Wilding, and Greg S Turner. Characteristics of the lax vowel space in dysarthria. 2005.

[54] Susan E Doble, Cindy Shearer, Julie Lall-Phillips, and Stan Jones. Relation between post-stroke satisfaction with time use, perceived social support and depressive symptoms. *Disability and rehabilitation*, 31(6):476–483, 2009.

[55] David R Beukelman, Pat Mirenda, et al. *Augmentative and alternative communication*. Paul H. Brookes Baltimore, 1998.

[56] Diane Nelson Bryen and Yoosun Chung. What adults who use aac say about their use of mainstream mobile technologies. *Assistive Technology Outcomes & Benefits*, 12(1):73–106, 2018.

[57] Melanie Fried-Oken, Lynn Fox, Marie T Rau, Jill Tullman, Glory Baker, Mary Hindal, Nancy Wile, and Jau-Shin Lou. Purposes of aac device use for persons with als as reported by caregivers. *Augmentative and Alternative Communication*, 22(3):209–221, 2006.

[58] A Moorcroft, N Scarinci, and C Meyer. "i've had a love-hate, i mean mostly hate relationship with these podd books": parent perceptions of how they and their child contributed to aac rejection and abandonment. *Disability and Rehabilitation: Assistive Technology*, 16(1):72–82, 2021.

[59] Rupal Patel. Phonatory control in adults with cerebral palsy and severe dysarthria. *Augmentative and Alternative Communication*, 18(1):2–10, 2002.

[60] JA Clark and RB Roemer. Voice controlled wheelchair. *Archives of physical medicine and rehabilitation*, 58(4):169–175, 1977.

[61] Arnon Cohen and Daniel Graupe. Speech recognition and control system for the severely disabled. *Journal of Biomedical Engineering*, 2(2):97–107, 1980.

[62] Michael Harris Cohen, James P Giangola, and Jennifer Balogh. *Voice user interface design*. Addison-Wesley Professional, 2004.

[63] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17:127–144, 2013.

[64] Michel Vacher, Sybille Caffiau, François Portet, Brigitte Meillon, Camille Roux, Elena Elias, Benjamin Lecouteux, and Pedro Chahuara. Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing (TACCESS)*, 7(2):1–36, 2015.

[65] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 99–103. IEEE, 2018.

[66] Victoria Young and Alex Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010.

[67] Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2-4):225–246, 1999.

[68] MA Anusuya and Shriniwas K Katti. Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*, 2010.

[69] Souheila Moussalli and Walcir Cardoso. Intelligent personal assistants: can they understand and be understood by accented l2 learners? *Computer Assisted Language Learning*, 33(8):865–890, 2020.

[70] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[71] Santiago-Omar Caballero-Morales and Felipe Trujillo-Romero. Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. *Expert Systems with Applications*, 41(3):841–852, 2014.

[72] Christopher J Leggetter and Philip C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech & language*, 9(2):171–185, 1995.

[73] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24, 2018.

[74] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. Investigating the accessibility of voice assistants with impaired users: mixed methods study. *Journal of medical Internet research*, 22(9):e18431, 2020.

[75] Eliseo Sciarretta and Lia Alimenti. Smart speakers for inclusion: How can intelligent virtual assistants really assist everybody? In *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*, pages 77–93. Springer, 2021.

[76] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic. *arXiv preprint arXiv:1708.05044*, 2017.

[77] Iain A McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. On the use of information retrieval measures for speech recognition evaluation. Technical report, IDIAP, 2004.

[78] David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4(1):1–9, 2015.

[79] Meredith Moore, Hemanth Venkateswara, and Sethuraman Panchanathan. Whistle-blowing asrs: evaluating the need for more inclusive automatic speech recognition systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018, pages 466–470, 2018.

[80] Mumtaz Begum Mustafa, Fadhilah Rosdi, Siti Salwah Salim, and Muhammad Umair Mughal. Exploring the influence of general and specific factors on the recognition accuracy of an asr system for dysarthric speaker. *Expert Systems with Applications*, 42(8):3924–3932, 2015.

[81] Kristin Rosen and Sasha Yampolsky. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16(1):48–60, 2000.

[82] Mengzhe Geng, Shansong Liu, Jianwei Yu, Xurong Xie, Shoukang Hu, Zi Ye, Zengrui Jin, Xunying Liu, and Helen Meng. Spectro-temporal deep features for disordered speech assessment and recognition. *arXiv preprint arXiv:2201.05554*, 2022.

[83] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen Meng. Adversarial data augmentation for disordered speech recognition. *arXiv preprint arXiv:2108.00899*, 2021.

[84] Shansong Liu, Mengzhe Geng, Shoukang Hu, Xurong Xie, Mingyu Cui, Jianwei Yu, Xunying Liu, and Helen Meng. Recent progress in the cuhk dysarthric speech recognition system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2267–2281, 2021.

[85] TA Mariya Celin, T Nagarajan, and P Vijayalakshmi. Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):346–354, 2020.

[86] Kristen M Allison, Yana Yunusova, and Jordan R Green. Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *American journal of speech-language pathology*, 28(1):96–107, 2019.

[87] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

[88] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. In *Interspeech*, pages 471–475, 2018.

[89] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Two-step acoustic model adaptation for dysarthric speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6104–6108. IEEE, 2020.

[90] R Sriranjani, M Ramasubba Reddy, and Srinivasan Umesh. Improved acoustic modeling for automatic dysarthric speech recognition. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[91] Davide Mulfari, Antonio Celesti, and Massimo Villari. Exploring ai-based speaker dependent methods in dysarthric speech recognition. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 958–964. IEEE, 2022.

[92] Marco Marini, Mauro Viganò, Massimo Corbo, Marina Zettin, Gloria Simoncini, Bruno Fattori, Clelia D'Anna, Massimiliano Donati, and Luca Fanucci. Idea: an italian dysarthric speech database. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1086–1093. IEEE, 2021.

[93] Mauro Nicolao, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain. A framework for collecting realistic recordings of dysarthric speech-the homeservice corpus. In *Proceedings of LREC 2016*. European Language Resources Association, 2016.

[94] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, 2012.

[95] Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino. Easycall corpus: a dysarthric speech dataset. *arXiv preprint arXiv:2104.02542*, 2021.

[96] Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, et al. Disordered speech data collection: lessons learned at 1 million utterances from project euphonia. 2021.

[97] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[98] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[99] Sunhee Kim, Yumi Hwang, Daejin Shin, Chang-Yeal Yang, Seung-Yeun Lee, Jin Kim, Byunggoo Kong, Jio Chung, Namhyun Cho, Ji-Hwan Kim, et al. Vui development for korean people with dysarthria. *Journal of Assistive Technologies*, 2013.

[100] Jan Derboven, Jonathan Huyghe, and Dirk De Grooff. Designing voice interaction for people with physical and speech impairments. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 217–226, 2014.

[101] Fabio Ballati, Fulvio Corno, and Luigi De Russis. " hey siri, do you understand me?": Virtual assistants and dysarthria. In *Intelligent Environments (Workshops)*, pages 557–566, 2018.

[102] Meredith Moore. Speech recognition for individuals with voice disorders. *Multimedia for Accessible Human Computer Interfaces*, pages 115–144, 2021.

[103] Veton Këpuska and Gamal Bohouta. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl*, 7(03):20–24, 2017.

[104] Mark Parker, Stuart Cunningham, Pam Enderby, Mark Hawley, and Phil Green. Automatic speech recognition and training for severely dysarthric users of assistive technology: The stardust project. *Clinical linguistics & phonetics*, 20(2-3):149–156, 2006.

[105] Foad Hamidi, Melanie Baljko, Connie Ecomomopoulos, Nigel J Livingston, and Leonhard G Spalteholz. Co-designing a speech interface for people with dysarthria. *Journal of Assistive Technologies*, 2015.

[106] Massimiliano Malavasi, Enrico Turri, Maria Rosaria Motolese, Ricard Marxer, Jochen Farwer, Heidi Christensen, Lorenzo Desideri, Fabio Tamburini, and Phil Green. An innovative speech-based interface to control aal and iot solutions to help people with speech and motor disability. In *Italian Forum of Ambient Assisted Living*, pages 269–278. Springer, 2016.

[107] Jort Gemmeke, Bart Ons, Netsanet Merawi Tessema, Janneke Van de Loo, Guy De Pauw, Walter Daelemans, Jonathan Huyghe, Jan Derboven, Lode Vuegen, Bert Van Den Broeck, et al. Self-taught assistive vocal interfaces: An overview of the aladin project. *Proceedings Interspeech 2013*, pages 2038–2043, 2013.

[108] Adam J Sporka, Sri H Kurniawan, Murni Mahmud, and Pavel Slavík. Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 213–220, 2006.

[109] Susumu Harada, Jacob O Wobbrock, and James A Landay. Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In *IFIP Conference on Human-Computer Interaction*, pages 11–29. Springer, 2011.

[110] Perttu Hämäläinen, Teemu Mäki-Patola, Ville Pulkki, and Matti Airas. Musical computer games played by singing. In *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04), Naples*, 2004.

[111] Sama'a Al Hashimi. Vocal telekinesis: towards the development of voice-physical installations. *Universal Access in the Information Society*, 8:65–75, 2009.

[112] Adam J Sporka, Sri Hastuti Kurniawan, and Pavel Slavik. Whistling user interface (u 3 i). In *User-Centered Interaction Paradigms for Universal Access in the Information Society: 8th ERCIM Workshop on User Interfaces for All, Vienna, Austria, June 28-29, 2004, Revised Selected Papers 8*, pages 472–478. Springer, 2004.

[113] Susumu Harada, James A Landay, Jonathan Malkin, Xiao Li, and Jeff A Bilmes. The vocal joystick: evaluation of voice-based cursor control techniques for assistive technology. *Disability and Rehabilitation: Assistive Technology*, 3(1-2):22–34, 2008.

[114] Takeo Igarashi and John F Hughes. Voice as sound: using non-verbal voice input for interactive control. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 155–156, 2001.

[115] Daisuke Sakamoto, Takanori Komatsu, and Takeo Igarashi. Voice augmented manipulation: using paralinguistic information to manipulate mobile devices. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 69–78, 2013.

[116] Yoshiyuki Mihara, Etsuya Shibayama, and Shin Takahashi. The migratory cursor: accurate speech-based cursor movement by moving multiple ghost cursors using non-verbal vocalizations. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 76–83, 2005.

[117] Susumu Harada, Jacob O Wobbrock, and James A Landay. Voicedraw: a hands-free voice-driven drawing application for people with motor impairments. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 27–34, 2007.

[118] Markus Funk, Vanessa Tobisch, and Adam Emfield. Non-verbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[119] Colin Lea, Zifang Huang, Dhruv Jain, Lauren Tooley, Zeinab Liaghat, Shrinath Thelapurath, Leah Findlater, and Jeffrey P Bigham. Nonverbal sound detection for disordered speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7397–7401. IEEE, 2022.

[120] Leela Damodaran. User involvement in the systems design process-a practical guide for users. *Behaviour & information technology*, 15(6):363–377, 1996.

[121] Mark S Hawley, Pam Enderby, Phil Green, Stuart Cunningham, Simon Brownsell, James Carmichael, Mark Parker, Athanassios Hatzis, Peter O'Neill, and Rebecca Palmer. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, 2007.

[122] Donald Norman. User centered system design. *New perspectives on human-computer interaction*, 1986.

[123] John D Gould and Clayton Lewis. Designing for usability: key principles and what designers think. *Communications of the ACM*, 28(3):300–311, 1985.

[124] Patrizia Marti and Liam J Bannon. Exploring user-centred design in practice: Some caveats. *Knowledge, technology & policy*, 22:7–15, 2009.

[125] Martin Maguire. Methods to support human-centred design. *International journal of human-computer studies*, 55(4):587–634, 2001.

[126] ISO HCD definition. https://www.iso.org/obp/ui/en/iso:std:iso:9241:-210:ed-2:v1:en. accessed: 14.09.202.

[127] Karel Vredenburg, Ji-Ye Mao, Paul W Smith, and Tom Carey. A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 471–478, 2002.

[128] Anne Adams, Peter Lunt, and Paul Cairns. A qualititative approach to hci research. 2008.

[129] Ann Blandford, Dominic Furniss, and Stephann Makri. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers, 2016.

[130] Carla Willig. *EBOOK: introducing qualitative research in psychology*. McGraw-hill education (UK), 2013.

[131] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[132] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[133] James R Lewis. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

[134] Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3-4):287–303, 2000.

[135] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[136] Daniela Aidley and Kriss Fearon. *Doing accessible social research: A practical guide*. Policy Press, 2021.

[137] Catherine MacKenzie, Shona Kelly, Gillian Paton, Marian Brady, and Margaret Muir. The living with dysarthria group for post-stroke dysarthria: the participant voice. *International journal of language & communication disorders*, 48(4):402–420, 2013.

[138] Filippo Trevisan. Making focus groups accessible and inclusive for people with communication disabilities: a research note. *Qualitative Research*, 21(4):619–627, 2021.

[139] Vicki Lloyd, Amanda Gatherer, and Sunny Kalsy. Conducting qualitative interview research with people with expressive language difficulties. *Qualitative health research*, 16(10):1386–1404, 2006.

[140] Gail Teachman, Peggy McDonough, Colin Macarthur, and Barbara E Gibson. A critical dialogical methodology for conducting research with disabled youth who use augmentative and alternative communication. *Qualitative Inquiry*, 24(1):35–44, 2018.

[141] Barbara Collier, Donna Mcghie-Richmond, and Hazel Self. Exploring communication assistants as an option for increasing communication access to communities for people who use augmentative communication. *Augmentative and Alternative Communication*, 26(1):48–59, 2010.

[142] Johnna Blair and Saeed Abdullah. It didn't sound good with my cochlear implants: Understanding the challenges of using smart assistants for deaf and hard of hearing users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–27, 2020.

[143] Luke Gelinas, Robin Pierce, Sabune Winkler, I Glenn Cohen, Holly Fernandez Lynch, and Barbara E Bierer. Using social media as a research recruitment tool: ethical issues and recommendations. *The American Journal of Bioethics*, 17(3):3–14, 2017.

[144] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[145] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. sage, 2013.

[146] Guangchao Charles Feng. Intercoder reliability indices: disuse, misuse, and abuse. *Quality & Quantity*, 48:1803–1815, 2014.

[147] GG Landis JRKoch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159174, 1977.

[148] Cliodhna O'Connor and Helene Joffe. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, 19:1609406919899220, 2020.

[149] Alireza Nili, Mary Tate, and Alistair Barros. A critical analysis of inter-coder reliability methods in information systems research. 2017.

[150] Susan Koch Fager, Melanie Fried-Oken, Tom Jakobs, and David R Beukelman. New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science. *Augmentative and Alternative Communication*, 35(1):13–25, 2019.

[151] Shaun K Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. " at times avuncular and cantankerous, with the reflexes of a mongoose" understanding self-expression through augmentative and alternative communication devices. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 1166–1179, 2017.

[152] António Teixeira, Daniela Braga, Luís Coelho, J Fonseca, Joaquim Alvarelhão, Inácio Martín, Alexandra Queirós, Nelson Rocha, António Calado, and Miguel Dias. Speech as the basic interface for assistive technology. In *DSAI 2009-Proceedings of the 2th International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion*, 2009.

[153] Ewa Luger and Abigail Sellen. " 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*, pages 5286–5297, 2016.

[154] Foad Hamidi, Melanie Baljko, Nigel Livingston, and Leo Spalteholz. Canspeak: a customizable speech interface for people with dysarthric speech. In *International Conference on Computers for Handicapped Persons*, pages 605–612. Springer, 2010.

[155] Jeff Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James Landay, Patricia Dowden, et al. The vocal joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 995–1002, 2005.

[156] Joseph M Wepman. Auditory discrimination, speech, and reading. *The Elementary School Journal*, 60(6):325–333, 1960.

[157] Kaitlin L Lansford and Julie M Liss. Vowel acoustics in dysarthria: Mapping to perception. 2014.

[158] Jimin Lee, Emily Dickey, and Zachary Simmons. Vowel-specific intelligibility and acoustic patterns in individuals with dysarthria secondary to amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 62(1):34–59, 2019.

[159] Yunjung Kim, Gary Weismer, Raymond D Kent, and Joseph R Duffy. Statistical models of f2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica*, 61(6):329–335, 2009.

[160] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[161] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

[162] Galina B Bolden. Little words that matter: Discourse markers “so” and “oh” and the doing of other-attentiveness in social interaction. *Journal of Communication*, 56(4):661–688, 2006.

[163] Diane Blakemore. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*, volume 99. Cambridge university press, 2002.

[164] United nations. Sdg goals.

[165] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016.

[166] Byonggon Chun, Beomseok Oh, Chihyun Cho, and Dongyeob Lee. Design and implementation of lightweight messaging middleware for edge computing. In *Proceedings of the 6th International Conference on Control, Mechatronics and Automation*, pages 170–174, 2018.

[167] Rabbitmq. `https://www.rabbitmq.com/`, 2007-2022. Accessed: 2022-06-10.

[168] Valeriu Manuel Ionescu. The analysis of the performance of rabbitmq and activemq. In *2015 14th RoEduNet International Conference-Networking in Education and Research (RoEduNet NER)*, pages 132–137. IEEE, 2015.

[169] Alexander L Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K Evershed. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1):388–407, 2020.

[170] Karl Spencer Lashley et al. *The problem of serial order in behavior*, volume 21. Bobbs-Merrill Oxford, 1951.

[171] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, pages 4277–4280. IEEE, 2012.

[172] Tara N Sainath, Brian Kingsbury, Abdel-rahman Mohamed, George E Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y Aravkin, and Bhuvana Ramabhadran. Improvements to deep convolutional neural networks for lvcsr. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 315–320. IEEE, 2013.

[173] Jui-Ting Huang, Jinyu Li, and Yifan Gong. An analysis of convolutional neural networks for speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4989–4993. IEEE, 2015.

[174] Chandra Kusuma Dewa. Javanese vowels sound classification with convolutional neural network. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 123–128. IEEE, 2016.

[175] Ossama Abdel-Hamid, Li Deng, and Dong Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech*, volume 2013, pages 1173–5. Citeseer, 2013.

[176] Shanqing Cai, Lisie Lillianfeld, Katie Seaver, Jordan R Green, Michael Brenner, Philip Q Nelson, and D Sculley. A voice-activated switch for persons with motor and speech impairments: Isolated-vowel spotting using neural networks. 2021.

[177] Niyada Rukwong and Sunee Pongpinigpinyo. Thai vowels speech recognition using convolutional neural networks. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–7. IEEE.

[178] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das. A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 126–130. IEEE, 2017.

[179] Sayed Khushal Shah, Zeenat Tariq, and Yugyung Lee. Iot based urban noise monitoring in deep learning using historical reports. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4179–4184. IEEE, 2019.

[180] Sanjay Krishna Gouda, Salil Kanetkar, David Harrison, and Manfred K Warmuth. Speech recognition: keyword spotting through image recognition. *arXiv preprint arXiv:1803.03759*, 2018.

[181] edge impulse. edge impulse.

[182] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[183] Gondy Leroy. *Designing user studies in informatics*. Springer Science & Business Media, 2011.

[184] Bennet B Murdock Jr. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482, 1962.

[185] Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

[186] Sultan bin abdulaziz humanitarian city. `https://sbahc.org.sa/`. Accessed: 2023-07-10.

[187] ISO 9241-11. Ergonomic requirements for office work with visual display terminals (vdt)s- part 11 guidance on usability.

[188] Hanbo Cai, Pengcheng Zhang, Hai Dong, Yan Xiao, and Shunhui Ji. Pbsm: Backdoor attack against keyword spotting based on pitch boosting and sound masking. *arXiv preprint arXiv:2211.08697*, 2022.

[189] Lucia Beccai, Stefano Roccella, Alberto Arena, Francesco Valvo, Pietro Valdastri, Arianna Menciassi, Maria Chiara Carrozza, and Paolo Dario. Design and fabrication of a hybrid silicon three-axial force sensor for biomechanical applications. *Sensors and Actuators A: Physical*, 120(2):370–382, 2005.

[190] CS Oon, M Ateeq, A Shaw, A Al-Shamma'a, SN Kazi, and A Badarudin. Experimental study on a feasibility of using electromagnetic wave cylindrical cavity sensor to monitor the percentage of water fraction in a two phase system. *Sensors and Actuators A: Physical*, 245:140–149, 2016.

[191] Ciarán McHale, Robert Telford, and Paul M Weaver. Morphing lattice boom for space applications. *Composites Part B: Engineering*, 202:108441, 2020.

[192] José-Antonio Gil-Gómez, Pilar Manzano-Hernández, Sergio Albiol-Pérez, Carmen Aula-Valero, Hermenegildo Gil-Gómez, and José-Antonio Lozano-Quilis. Useq: a short questionnaire for satisfaction evaluation of virtual rehabilitation systems. *Sensors*, 17(7):1589, 2017.

[193] A Baki Kocabalil, Liliana Laranjo, and Enrico Coiera. Measuring user experience in conversational interfaces: a comparison of six questionnaires. 2018.

[194] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. Understanding and measuring user experience in conversational interfaces. *Interacting with Computers*, 31(2):192–207, 2019.

[195] Nick Feltovich. Nonparametric tests of differences in medians: comparison of the wilcoxon–mann–whitney and robust rank-order tests. *Experimental Economics*, 6:273–297, 2003.

[196] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[197] Stephen A Brewster. Using nonspeech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(3):224–259, 1998.

[198] Marilyn McGee-Lennon, Maria Wolters, Ross McLachlan, Stephen Brewster, and Cordelia Hall. Name that tune: musicons as reminders in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2803–2806, 2011.

[199] Parimala Raghavendra, Elisabet Rosengren, and Sheri Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication*, 17(4):265–275, 2001.

[200] Frank Rudzicz. Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 255–256, 2007.

[201] John Wilson, Bronagh Blaney. Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics & Phonetics*, 14(4):307–327, 2000.

[202] Joseph Perkell, Melanie Matthies, Harlan Lane, Frank Guenther, Reiner Wilhelms-Tricarico, Jane Wozniak, and Peter Guiod. Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication*, 22(2-3):227–250, 1997.

[203] Monideepa Tarafdar et al. Analyzing the influence of web site design parameters on web site usability. *Information Resources Management Journal (IRMJ)*, 18(4):62–80, 2005.

[204] Eugene Cho, S Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. Will deleting history make alexa more trustworthy? effects of privacy and content

customization on user experience of smart speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[205] Eija Kaasinen, Tiina Kymäläinen, Marketta Niemelä, Thomas Olsson, Minni Kanerva, and Veikko Ikonen. A user-centric view of intelligent environments: User expectations, user experience and user role inbuilding intelligent environments. *Computers*, 2(1):1–33, 2012.

[206] Wendy E Mackay. Triggers and barriers to customizing software. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 153–160, 1991.

[207] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130, 2016.

[208] Fulvio Corno, Laura Farinetti, and Isabella Signorile. A cost-effective solution for eye-gaze assistive technology. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 2, pages 433–436. IEEE, 2002.

[209] Helena Hemmingsson and Maria Borgestig. Usability of eye-gaze controlled computers in sweden: A total population survey. *International journal of environmental research and public health*, 17(5):1639, 2020.

[210] Marco Caligari, Marco Godi, Simone Guglielmetti, Franco Franchignoni, and Antonio Nardone. Eye tracking communication devices in amyotrophic lateral sclerosis: impact on disability and quality of life. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(7-8):546–552, 2013.

[211] Petra Karlsson, Abigail Allsop, Betty-Jean Dee-Price, and Margaret Wallen. Eye-gaze control technology for children, adolescents and adults with cerebral palsy with significant physical disability: Findings from a systematic review. *Developmental neurorehabilitation*, 21(8):497–505, 2018.

[212] Mick Donegan, Jeffrey D Morris, Fulvio Corno, Isabella Signorile, Adriano Chió, Valentina Pasian, Alessandro Vignola, Margret Buchholz, and Eva Holmqvist. Understanding users and their needs. *Universal Access in the Information Society*, 8:259–275, 2009.

[213] Ladan Najafi, Marcus Friday, and Zoe Robertson. Two case studies describing assessment and provision of eye gaze technology for people with severe physical disabilities. *Journal of Assistive Technologies*, 2(2):6–12, 2008.

[214] Tobii. Tobii eye tracker.

[215] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 Chi conference on human factors in computing systems*, pages 1118–1130, 2017.

[216] Alexandre Bissoli, Daniel Lavino-Junior, Mariana Sime, Lucas Encarnação, and Teodiano Bastos-Filho. A human–machine interface based on eye tracking for controlling and monitoring a smart home using the internet of things. *Sensors*, 19(4):859, 2019.

[217] Emanuele Pasqualotto, Tamara Matuz, Stefano Federici, Carolin A Ruf, Mathias Bartl, Marta Olivetti Belardinelli, Niels Birbaumer, and Sebastian Halder. Usability and workload of access technology for people with severe motor impairment: a comparison of brain-computer interfacing and eye tracking. *Neurorehabilitation and neural repair*, 29(10):950–957, 2015.

[218] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[219] Viswanath Venkatesh, James YL Thong, and Xin Xu. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems*, 17(5):328–376, 2016.

[220] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. We need to talk: Hci and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2459–2464. 2013.

[221] Yuhui Wang. Gaps between continuous measurement methods: A longitudinal study of perceived usability. *Interacting with Computers*, 33(3):223–237, 2021.

[222] Katarzyna Kabacińska, Kim Vu, Mallorie Tam, Olivia Edwards, William C Miller, and Julie M Robillard. "functioning better is doing better": older

adults' priorities for the evaluation of assistive technology. *Assistive Technology*, 35(4):367–373, 2023.

[223] Kang Wang, Shen Wang, and Qiang Ji. Deep eye fixation map learning for calibration-free eye gaze tracking. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*, pages 47–55, 2016.

[224] Sajay Arthanat, Stephen M Bauer, James A Lenker, Susan M Nochajski, and Yow Wu B Wu. Conceptualization and measurement of assistive technology usability. *Disability and Rehabilitation: Assistive Technology*, 2(4):235–248, 2007.

[225] Richard Cave and Steven Bloch. The use of speech recognition technology by people living with amyotrophic lateral sclerosis: A scoping review. *Disability and Rehabilitation: Assistive Technology*, 18(7):1043–1055, 2023.

[226] Manuel Imaz and David Benyon. *Designing with blends: Conceptual foundations of human-computer interaction and software engineering methods.* 2007.

[227] Haiwei Dong, Ali Danesh, Nadia Figueroa, and Abdulmotaleb El Saddik. An elicitation study on gesture preferences and memorability toward a practical hand-gesture vocabulary for smart televisions. *IEEE access*, 3:543–555, 2015.

[228] Miguel A Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1099–1108, 2013.

[229] Paul Cairns. *Doing better statistics in human-computer interaction.* Cambridge University Press, 2019.

Appendix

Appendix A

Interview Questions

1. About Dysarthria
   (a) Can you tell me a brief history about your case with dysarthria?
   (b) Could you describe your speech after having dysarthria?
   (c) How do you feel when you are speaking?
   (d) How does dysarthria affect your daily life?
2. Coping with dysarthria
   (a) What technologies are you using to cope with dysarthria?
   (b) Do you have/use a virtual home assistant? (If yes: What do you use these devices for? If No: Why you do not use it?)
3. The proposed system
   (a) (after explaining the concept of non-verbal interaction) Do you think using nonverbal voice cues will be convenient?
   (b) What non-verbal voice cues can be convenient for you to utter?
   (c) Would you prefer having a predefined list of commands or having the ability to program your own commands?
   (d) Are we willing to record non-verbal sounds to be used in building our system?