Pwy: Designing a Discrete Speaker Recognition App for Conversational Support on Smartwatches

Osian Smith

September 2019

Theses submitted to Fulfilment for the Degree of Masters Of Research

Swansea University
Prifysgol Abertawe

Future Interactive Technology Laboratory
Department of Computer Science
College of Science
Swansea University

Page removed for security
reasons

Acknowledgements

Firstly I would like to thank my supervisor Stephen Lindsay for his advice and guidance throughout the project. His advice and guidance have been invaluable to the project.

I would also like to thank Joss Whittle for his support for the Machine Learning side of this theses. Without his help, we would not have been able to have an working machine learning algorithm.

I would also like to thank Darron Scott for his support during studies and assistance with writing for CHI. I would also like to thank Zoe and Carole from Traumatic Brain Injury Service, Morriston Hospital Swansea, for helping to recruit participants with Traumatic Brain Injury for our design study.

I would also like to thank Dr Matt Roach for our discussions and his reccomendations into reviewing the tipping points of technology.

I would also like to thank the class of CSC349 2019 for their input during their assignment for this project. I want to thank all participants and actors that took part in my studies and everyone who has given me feedback on the research and its development.

I want to thank my friends and the Postgraduate Lab in Computational Foundry, especially Cameron Steer and Gavin Baily for their support during my research. I would also like to thank my parents, Dilwen and Peter Smith for their loving support throughout my Masters. Thank you, Diolch

Abstract

We investigate conversational speaker-recognition systems, inferring identity from any spoken phrase, to support people who find recalling names in conversation difficult by discretely providing them with speakers' names and other relevant personal information via a smartwatch. We ran participatory design sessions with expert designers, people who self-identify as finding socialising difficult and people diagnosed with Traumatic Brain Injury. Sessions addressed social attitudes, privacy and adding new people to the system for future recognition. We discuss significant differences the process uncovers between groups. We train a speaker-recognition algorithm based on spectrogram feature extraction and classification. However, the implementation had delays of two to eight seconds between the start of conversation and recognition of speakers. Consequently, we ran studies to understand how delays in alerting users to speaker identity impacted on the perceived usefulness of the application.

Contents

Definitions

In this section, we will define keywords that will use within this theses. We will also refer to a more formal definition within this theses.

Accessibility Tools (AT)

Accessibility Tools (AT) are tools that are used to support users with imparements to assist them. AT can be hardware (such as hearing aids, wheel chairs) or software base (colour inversion, text to speech).

Bystanders

Bystanders are people who captured by the application, even though they may have no desire or be unwilling to interact with the user [39] .

PVI

People with Visual Impairments

Prosopagnosia

Prosopagnosia is the condition that inhibits the ability to recongise faces. A further explination can be found in chapter chapter 2 section 2.1.

Speaker recognition

Speaker recognition is identifying individuals from the characteristics of their voice. The use of speaker recognition is not for authentication, which is defined as speaker verification. A further definition can be seen in chapter 5 section 5.1

Tramatic Brain Injury (TBI)

Tramatic Brain Injury (TBI) is an injury that has been sustained through a brain injury such as a or impact. A further explination can be found in chapter chapter 2 section 2.1. While TBI is usually used to define patients who have had a tramatic event such as a fall, Morriston Hospital Tramatic Brain Injury Service refers to all brain related inuries as TBI and as a result we will use TBI to refer to any brain injury in this document.

Situationally-Induced Impairments And Disabilities (SIIDs)

Situationally-Induced Impairments And Disabilities (SIIDs) is an event where a user is impaired using technology by their situation, for example, using your phone in direct sunlight.

1 Introduction

The ability to recognise faces and link them to names is fundamental to functioning within society with evidence suggesting this trait is evolutionary [55]. The ability to link faces and names allows us to understand many important components of a conversation such as allowing us to determine our relationship to whoever we are in discussion with, know where a conversational partner is looking or infer a stranger's gender, age, health, and mood [44, 157]. However, several health conditions inhibit the ability to recognise faces including brain injury, Alzheimer's disease, Autism Spectrum Disorders (ASD) [26] and Prosopagnosia which can prevent facial recognition entirely. The World Health Organisation estimates that 4.23% of the global population are blind or have a visual impairment [119]. Approximately 2.5% of the adult US population has prosopagnosia, a condition where they are unable to recognise a face. These conditions do not necessarily overlap, and further conditions such as Alzheimer's, Autism Spectrum Disorder (ASD) and Traumatic Brain Injury (TBI) can all lead to the ability of facial recognition to be diminished.

People with prosopagnosia have a smaller social network compared to those who do not, with many citing traumatic social experiences [26, 157]. The ability to socialise is a crucial part of human psychology with smaller social networks increasing the risk of depression [45]. Maslow's hierarchy of needs suggests that humans require high social standards and values a sense of belonging with others, to be loved and have strong relationships with people or have a healthy self-esteem [100] all of which can be impaired by the inability to hold conversations. The consequences of an impaired social life can be far-reaching. Arthritis is a painful condition that inhibits the physical movement of the joints, which has a significant impact on the livelihood of people with arthritis. However, evidence shows that people who live alone have a more significant impact on their quality of life than a diagnosis of osteoarthritis does [45], the most common type of arthritis in the UK [112]. This demonstrates that the inability to socialise is worse for a person's wellbeing than decreased mobility.

People who are sensitive to social rejection engage in social withdrawal, leading to social avoidance and distress along with depressive symptoms [150]. Compunding this, depressive symptoms are significantly more problematic for people with prosopagnosia as they are likely to face social embarrassment. While medication can support people with depression, many patients do not respond to medication alone [66]. Many of these conditions do not have any treatment, although therapy does exist to support people with these conditions. In addition, for conditions such as prosopagnosia, treatment does not always result in any improvement [29, 109, 140] due to the cognitive load. For example, learning peoples mannerisms and visual strategies such as the colour of people's hair. Furthermore, many of these treatments are unsuitable for people with Traumatic Brain Injury (TBI) due to their added cognitive load [20].

The current work of the CHI community to offer support for people living with conditions like prosopagnosia has focused on supporting them using facial recognition through video capture [96, 154]. However, video-based recognition relies on specific circumstances so can be inhibited by, for example, poor lighting, a long or short distance to the face or partial or full obfuscation of the face. Furthermore, the use of cameras introduces serious, complex privacy concerns for both users and those being observed[83, 124]. These constraints limit the practical support video-based recognition can offer in the real world. Furthermore, much of the research to date has not utilised the input of participants in the design process [56, 96, 103, 137, 148, 154]. This lack of involvement of the user regularly leads to product abandonment [105]. Many people have already developed their support strategies, and if the technology does not fit within their needs, or harms

their support strategies, they are counterproductive to users. Research estimates for the reasons mentioned above that 75% of all accessible technology users will abandon their accessibility tools [105].

Within the last decade, the speech processing area has developed due to the rise of smart assistants such as Apple's Siri and Amazon's Alexa. Traditional approaches have focused on recognition from pitch-contours [12]., Hidden Markel Models (HMM) [24] and Gaussian Mixture Models [16], recently speech processing has moved to the use of Neural networks [89, 129]. Recently mobile phones have started to ship with neural networking processors allowing neural networks to be run on the device efficiently allowing the use of neural networks in new novel applications.

In this work we explore the design of a smartwatch wearable to discretely tell you who you are talking to in order to support the wearer in conversation. The smartwatch records and analyses voices in the environment and presents best guesses as to the identity of the person you are in conversation with derived by machine learning ran entirely on a companion smartphone. This work is timely as today's smartphones are starting to include dedicated neural networking processors that allow us to run machine learning frameworks such as TensorFlow Lite[1] [TFLite] and CoreML [2], which allow us to run new machine learning algorithms locally [65]. Video-based facial recognition is a computationally cheap process which explains its use to date in prototypes that recognise people in social situations [154]. However, using machine-learning frameworks such as TFLite and a smartphone neural network processor, we can execute novel machine learning algorithms such as speaker-recognition from short snippets of conversation to identify the person that one is in conversation with in real time. Speaker-recognition has several advantages over facial recognition, such as improved perceptions of privacy intrusion and practical reductions in the data being stored for recognition's security risk coupled with a reduction in social stigma.

Wearable technology such as smartwatches give us a new method of interaction with apps without the need for a mobile phone in the hand of the user and so are more discrete than a mobile phone. Consumers perceive smartwatches as fashionable [25] with fashion brands such as Fossle[3] and diesel [4] producing smartwatches. Many smartwatches look commonplace on users, for example, in figure 1.1 where a person is wearing an Apple Watch. Many smartwatches contain a large array of sensors such as step counters, heart rate sensors, touchscreens and microphones. Smartwatches enable us to run applications on the device and allow us to access these sensors.

Figure 1.1: An person wearing an Apple Watch, a popular smart watch

However, ethical and legal questions arise with the use of speaker recognition. Previously, smart glasses failed due to concerns surrounding bystander privacy with wearers of Google Glass (GG) were named "glassholes" by the popular press [27]. Research into GG found that bystanders do not know whether they were filmed on GG [133] and would have liked to have given permission [31] to be filmed or not. Many people do not like being filmed and with GG ability to discreetly record it made people uncomfortable around glass wearers leading to the term coined above. Further questions surround the legal aspect of capturing someone using wearables. While it is legal to record an individual in

[1] https://www.tensorflow.org/lite
[2] https://developer.apple.com/documentation/coreml
[3] https://www.fossil.com/UK/en/wearable-technology/smartwatches/smartwatches.html
[4] https://uk.diesel.com/en/man/smartwatches/?t=m

public in the UK due to privacy is not guaranteed by law [34] there are restrictions on private property that can restrict recordings [69]. In this research, we will explore the ethical and legal challenges that are using wearables to support the recognition of people.

In this research, we used speaker-recognition to aid social interaction. We developed Pwy, which is Welsh for "Who" as in "Pwy ydych chi?" or "Who are you?", to use in participatory design workshops and to run evaluation studies with. We analyse audio data captured from the Apple Watch to understand the quality of the audio from the watch. Our work also shows how HCI research, through participatory design workshops, can support choices made by the machine learning community and drive decisions on trade-offs between speed, features and privacy.

1.1 Project scope

For this MRes, the project investigated previous work in the HCI community to support people with social interaction. While previous work in the HCI community has focused on supporting people with recognising others, these approaches also present serious privacy concerns. While we will review machine learning approaches to understand the trade-offs involved in their use, contributions to the performance of the machine learning algorithm are outside the scope of this research.

The project scope is to focus on user requirements with machine learning algorithms and, as the research progresses in both areas, how the trade-off between features, speed and privacy can be made. Through running participatory design workshops with different potential users, we arrived at our own set of trade-offs between feature richness and privacy concerns. In the evaluation study of delays, we gained an understanding of how a delay in presenting a name or prompt can affect the acceptance of the application and furthered our understanding of the balance between speed and privacy that users require.

1.2 Research Aims

In this project, we investigate the design of speaker-recognition to support people with difficulty interacting with other people and evaluate elements of its performance.

We aim to address the following:

Understand Design Requirements and Specifications For Speaker-Recognition From Stakeholders

In this project, we will run design activities such as participatory design workshops with design students and expert designers along with people who lack social confidence and people with TBI. By employing participatory design workshops, we can engage with users to allow their inputs in the design process. From these sessions, we will derive specifications and requirements for developing a system to support people with inferring the speaker through speaker-recognition.

Develop A Speaker-Recognition Algorithm Using Machine Learning

We will develop a speaker-recognition algorithm to understand the performance of such an algorithm on a smartwatch along with its limitations arising from current machine learning knowledge and from the quality of audio data that the smartwatch captures. By researching current algorithms, we will then derive our approach that places the privacy of users and bystanders at the forefront.

Evaluating The Use of Speaker-Recognition To support Users

We will run evaluation studies to explore how speaker-recognition can support people with inferring a speaker. We will evaluate our speaker-recognition algorithm, along with running studies with participants to evaluate

the use of speaker-recognition in conversation.

1.3 Ownership of work

This project has involved other researchers who have supported Osian Smith during his MRes. Dr Stephen Lindsay, as a supervisor, has supported this project through literature, ensuring that all design activities with participants were suitable and provided writing support on the paper submissions. The candidate is the lead author, wrote the first drafts of all sections and held responsibility for accepting or rejecting edits to the final paper.

Darren Scott, a PhD candidate in Swansea University, also supported the candidate in design activities and was a co-author of 'Pwy?: Designing a Discrete Speaker Recognition App for Conversational Support on Smartwatches' which we have submitted to CHI 2020.

Dr Joss Whittle, a Tutor at Swansea University, worked with Osian Smith to gather training data and then worked with the candidate to produced the speaker-recognition algorithm. The candidate then attempted to convert the model to a mobile model, which was unfortunately unsuccessful due to compatibility issues.

1.4 Arising Publications

From this MRes, we have submitted two publications to peer reviewed destinations. These papers are as follows:

1.4.1 Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury

This paper was accepted to a CHI 2019 Workshop on Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction. This paper can be found at `arxiv.org`. We discuss this paper further in Chapter 3.2. The paper is viewable in the appendix under section A. A slide deck from a presentation is viewable in the appendix under section B.

1.4.2 Pwy?: Designing a Discrete Speaker Recognition App for Conversational Support on Smartwatches

We have also submitted a paper to CHI2020 in which is derived from this theses. At the time of writing, this paper is currently under review. The paper that was submitted to review on the 20th of September 2019 and can be viewed in the appendix under section C.

2 Literature Review and Background to current work

In this chapter, we will present the background previous work within this area. We will present the work mainly focusing on Human-Computer Interaction where the work of the HCI community has focused on two ways to support people in social situations through recognising people around them: biometrics and alternative identifiers. Biometrics are the identification of individuals on their anatomical and behaviour characteristics such as fingerprints, facial patterns or characteristics of their speech [96, 154]. Alternative identifiers work by identifying things associated with a person such as digital devices they carry or items they wear. Examples of successful alternative identifies include clothing identification [148] and mobile phone signal logging [56]. In this chapter, we will also explore the legal and ethical concerns surrounding the recognition of people and bystander privacy and how different demographics feel towards being tracked. We will also explore the tipping point of acceptance for machine learning algorithms such as virtual assistants.

2.1 Conditions that Inhibit the Ability To Recognise Faces

Humans recognise faces in several stages. When the brain detects a face (through vision or mental imagery), the signals are analysed for individual (such as eyes) and non-individual information (such as emotional expression) [55]. The brain processes this in the fusiform gyrus [76]. If the fusiform gyrus is impaired, this results in the inability to recognise faces. Numerous conditions can inhibit peoples ability to recognise other people and even themselves in photos and mirrors. All of these are cognitive conditions, although certain conditions are as a result of an injury. In this section, we will give an overview of some of these conditions that can affect the ability to recognise faces.

Learning difficulties

Children with Autism Spectrum Disorder (ASD) may find it challenging to infer peoples' faces. While a significant amount of studies do suggest that people with ASD struggle with facial recognition, there is also a significant amount of studies arguing that this is not the case [132]. Evidence does suggest that young children with ASD perform worse with recognising faces compared to other children similarly aged, however, while they grow up they develop coping strategies [82].

Alzheimer's Disease

Alzheimer's Disease can also impair the ability of people's ability to discriminate faces, but still remember familiar. Although Alzheimer's patients are unable to recognise people's faces sometimes, they can distinguish facial emotion, demonstrating that facial recognition and emotion recognition are separate cognitive systems [130]. Losing the ability to recognise people is an extremely upsetting situation for people with Alzheimer's and their families.

Genetic Conditions

Genetic conditions such as Turing Syndrome in Women [87] can also lead to an inability to infer facial

recognition. Turing syndrome can also affect the ability to recognise facial emotions, which suggests that an anomaly causes this within the workflow of recognising faces. However, come genetic conditions can actually improve the ability to recognise faces, conditions such as Williams syndrome increase the ability to discriminate between faces [43].

Traumatic Brain Injury (TBI)

TBI, sometimes also called Acquired Brain Injury, is brain injury that is caused by a traumatic event such as a fall or a stroke that was not present at birth. Brain haemorrhage, tumour, carbon monoxide poisoning and meningitis are among the conditions that can cause TBI [58]. The severity of TBI and specific location of the injury will determine whether it causes difficulty with faces depending on what part of the brain is affected [55, 109]. When socialising, this issue can be compounded by other factors such as memory problems, increased irritability or reduced ability to regulate emotions.

Prosopagnosia

Prosopagnosia (the phrase prosopon meaning face and the phrase agnosia meaning non-recognition or non-knowing) is used to refer to people with facial blindness [55, 157]. Prosopagnosia may be present in patients due to developmental factors, genetic or acquired through a Traumatic Brain Injury (TBI) [29] and is distinguished from factors such as impaired vision as a specific cognitive disorder of face perception. Current estimation suggests that prosopagnosia affects 1 in 50 people [55]. However, many people with prosopagnosia, especially people with genetic or developmental prosopagnosia, may not realise that they have prosopagnosia [109]. Medical professionals only diagnose prosopagnosia when no other medical condition that can explain why an individual cannot recognise people [55, 157].

2.1.1 Treatment for facial blindness

Many of the conditions that lead to prosopagnosia do not have treatments; however, many of these conditions can be improved with therapies which can demonstrate tangible benefits such as the being able to recognise colleagues after intense training [29, 109, 111, 113, 140]. Medical literature is still developing regarding facial blindness with a greater understanding of what treatments work on different conditions [20] but this section will give a brief overview of therapies where they exist.

There is no treatment for ASD. Nonetheless, there are a concerning amount of fake treatments circulating, many of which pose a significant risk of harm to the patient. Many of these treatments are extreme, which include injection, hormonal treatments, "Mineral Miracle Solution" which requires drinking diluted industrial bleach at unsafe levels [145] or hyperbaric oxygen therapy [111]. These treatments offer no benefit to the patient and in many cases, poses a significant risk to the patient's health and wellbeing. Currently, the NHS states that there is no treatment for ASD [111]. Therapy can help people with ASD recognise faces, but Kiln et al. stated that many people with ASD would develop their coping techniques so it is likely the ability to recognise faces will improve with age [82].

There are no treatments available for Turner Syndrome and Willams Syndrome as these conditions are genetic and affect the development of the person with these conditions. While treatments do help to mitigate the majority of symptoms, treatments place a strain on the person due to constant medical attention.

Not everyone with TBI will develop facial blindness but will depend on the location and severity of the injury. People with TBI may have developed acquired prosopagnosia. While cases do exist of spontaneous recovery, current medical literature suggests that this is a result of medical misdiagnosis [29]. However, for most patients, a CT scan will take place to understand the severity of the injury and to determine the best route of action [110].

Most of the therapies available for prosopagnosia relies on training of facial processing such as recognising blemishes and birthmarks, or verbal strategies such as identifying people from colour of their hair or from

freckles [18, 22, 40, 123]. However, many participants will demonstrate limited to no improvement [29, 140] and may be unsuitable for people with TBI due to their diminished cognitive ability [20]. Currently, patients rely on therapy for prosopagnosia, however not all patients will demonstrate any improvement. While historically some people have historically demonstrated a spontaneous recovery to acquired prosopagnosia, current literature states that this was likely down to a misdiagnosis [29].

2.2 Applications of Recognition for Social Support

The work of the HCI community has focused on two ways to support people in social situations through recognising people around them: biometrics and alternative identifiers. Biometrics is the identification of individuals absed on their unique anatomical and behaviour characteristics such as fingerprints, facial patterns or characteristics of their speech [96, 154]. Alternative identifiers work by identifying things associated with a person such as digital devices they carry or items they wear. Examples of successful alternative identifies include clothing identification [148] and mobile phone signal logging [56].

2.2.1 Biometrics

During our literature review, we found that biometric recognition is the primary approach within the HCI community to recognise people. Biometrics is defined as the "automated recognition of individuals based on their anatomical and behavioural characteristics such as fingerprint, face, iris, and voice" [73]. Biometrics is made up of biometric identifiers that are the measurable characteristics of an individual, such as the distance between the eyes, ears and nose [72].

Biometrics are widely used within the security field as a form of authentication [127], with many phones containing a facial recognition system or fingerprint sensor [28].It is one of the most accurate methods to infer an individual's identity as it does not rely on an external factor such as carrying a mobile phone, but it is the most intimate type of recognition requiring the processing of personal data such as facial data or voice data that cannot be changed if it is compromised [73].

A vast array of open-source libraries for biometrics and computer vision such as OpenCV[1] have allowed HCI researchers to study automated recognition of people through biometrics collaboration with a machine learning expert, allowing novel approaches ways to support people with facial blindness.

2.2.1.1 Smart Glasses and Facial Recognition

Smart glasses are a subset of wearable technology that sits on the users head. Many smart glasses contain an array of sensors and output's such as microphone, GPS, holographic lenses and cameras and communication to apps allow for the possibility of using facial recognition to identify people [154, 96]. The benefits of Smart glasses is that most glasses have a display which allows information present without having to look at a smartphone or smartwatch to infer whom you are in discussion with, however certain smart glasses such as Snapchat Spectacles[2] do not contain any output. A pair of smart glasses can also be constantly monitoring their surroundings for a person to enter its field of view through a camera without any input of the users, allowing users to just glance at the smart glasses screen to identify a person. Many smart glasses such as the Microsoft Hololense [3] contain processors onboard that allow them to run facial identification, although the computational cost we found in the literature review were expensive and as a result significantly decreased the user experience [96]. Smart glasses can also feed information to smartphones, which allows other novel algorithms to be used to infer the speaker, such as looking at clothing or walking signatures. However, there are significant user and bystander privacy concerns with smart glasses resulting in a backlash against wearers of the devices with users being called "Glassholes" which lead to widespread public backlash and rejection of

[1] https://opencv.org
[2] www.spectacles.com
[3] https://www.microsoft.com/en-us/hololens

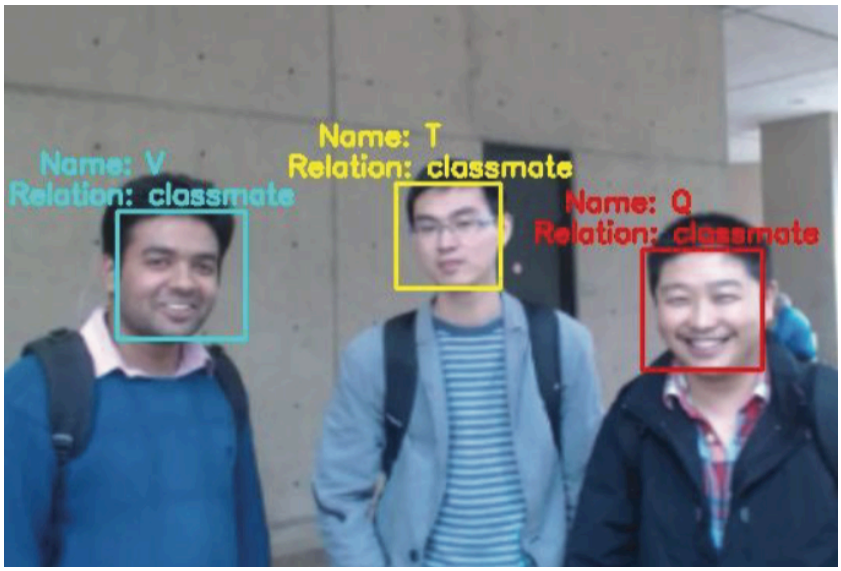


Figure 2.1: An example of an application of smart-glasses by Wang et al. In this example a name and the relation to the individual is displayed using an AR screen. Image source: [154]

Google Glass in 2013 [11, 27, 79]. Research by Wang et al. [154] explored using the Vuzix STAR 1200XL[4] smart glasses with augmented reality (AR) screen to highlight faces and overlay the name of the individual. A camera was attached to smart glasses which captured images continuously and utilised facial recognition to detect faces. Once a face was detected, it was analysed to test to see whether it was a known face. The glasses also displayed information on the relationship of the person to the wearer, such as whether they were family or a classmate. By utilising AR glasses, information was displayed directly in the sight of the user, meaning that the user did not have to look elsewhere for information as can be seen in figure 2.1. Wang et al. were able to achieve 99% accuracy for detection of faces and 95% of recognition of known faces on the Yale dataset which consists of 165 grayscale images of 15 individuals [155] along with Wang et al. own database which we could not access [154]. The choice of using an undisclosed dataset suggests to us that this data set consisted of ideal images.

We consider this to be a very high rate of success for the system as this is more accurate than someone with facial blindness [29]. Under this system, for every 100 recognitions carried out, five people would be recognised incorrectly by the system which we consider to be acceptable as this is high accuracy for an algorithm, and this system is possibly more accurate than someone without facial blindness recognising faces [29]. However, Wang et al. did not document the testing conditions of the system, which means the system was possibly only tested under ideal conditions such as clear lines of sight, good lighting, low levels of visual clutter etc. It is crucial to understand how the system would work under poor conditions because this is where the system may be needed by the user the most and in fact might be the majority of someones experiences.

Using AR is a novel approach, which has its benefits for group situations where the user can locate the speaker. Wang highlighted this use case in the paper by utilising an image with faces highlighted with their name and relation above. However, AR is computationally expensive, with the work in Wang et al. requiring the inference and AR placements to run on the mobile phone as the Vuzix STAR 1200 smart glasses did not contain any processors capable of running graphical software. Wang et al. did not state the battery cost of using facial recognition on the phone or the Vuzix STAR 1200, which makes it difficult for us to judge the cost of AR and weight up the merits and detriments for utilising it to support people with facial blindness.

[4] https://www.vuzix.com/Products/LegacyProduct/6

While the Vuzix STAR 1200 is discontinued and no longer for sale, Vuzix current offering, Vuzix blade current smart glasses do contain the ability to run applications on device allowing facial recognition to run on the device, making it suitable for on-device inference. However, a review of the Vuzix blade by Mike Prospero, writer for the technology news and review site Tom's Guide[5], stated that while the battery should last for 8 hours, this was dependent on the application and with five minutes of using the application Accuweather resulted in a battery depletion of 14% with several other apps resulting in poor battery life [126]. We consider this to be unacceptable, though, we must consider whether optimisation would help increase battery life by, for example, running the facial recognition algorithm on an external device such as a phone.

Battery life is a significant problem for wearable technology as wearable technology is generally smaller than smartphones, thus contain a smaller battery than a phone, similar to comparing a phone to a laptop. The users of these devices will require these devices throughout the day, and do not want the device dying on them.

Mandal et al. [96] extended the work of Wang et al. by using Google Glass (GG) a wearable headset available to consumers[6] to run facial recognition and compare on device inference and off device inference using Youtube Face Database. By comparing performance between the processor on smartglasses compared to a proccessor on the phone phone, we can understand how the user experience would be effected and whether on device inferance is suitable. After 100 on device facial recognitions, GG battery was at 33%. When Mandal et al. used GG as a client to send photos to a smartphone for inference, after 100 facial recognition tasks GG battery was at 70% while the mobile phone was at 97% battery. Mandal et al. also note that GG performance degraded very quickly due to the heat produced running the algorithm [96] with an accuracy of 97%. The work that Mandal et al. highlight that processing on wearables are not comparable to the performance of a smartphone.

The difference in battery life is significant, with off device recognition causing a battery improvement of more than 212% in battery life with no significant impact on the phone battery. Off device recognition used on average 0.3% of GG battery. On device recognition resulted in 0.67% of GG battery being used for facial recognition. By running a computation on the phone, we can process 333 facial recognitions; however, runing on GG would only allow us to run 149 recognitions. The majority of adults in the UK already smartphone users [116] and by utilising their current equipment, there is no tangible benefit of using on-device recognition. On-device recognition could still be retained when a phone is not present; however, it is clear that for this task should resort to the phone for inference.

2.2.1.2 Walking Signatures

Facial recognition is not the only biometric method that can be utilised to infer someones identity. Work by Wang et al. [148] has utilised walking signatures to infer individuals. In this research, Wang et al. were attempting to create "temporary fingerprints." Temporary fingerprints allowed the system to recognise an individual, but the individuals could also turn off their fingerprints through an app. Wang et al. also used clothing signatures to support the recognition of an individual. In this work, a video from GG was transmitted to a remote server where it extracted motion from the video for analysis. The system analysed footage when the person was walking for step-duration,phase of step the step along with the direction that they were walking. Once the system has inferred someone, it would then extract their clothing data and refer to clothing data to stop the system reanalysing walking signatures again on the same person [148].

The work by Wang et al. demonstrates that walking signatures is a viable but computationally expensive method. Beyond computation, it does have specific drawbacks; walking recognition requires a person to be walking, which can be problematic for users. We must consider whether the person we want to infer

[5] www.tomsguide.com

[6] Since Mandal et al. has published their work, Google stopped producing a consumer version of GG[95] and opted to provide an enterprise only edition [84]. Due to this decision, consumers and researchers cannot access GG.

is sitting, whether waiting for a participant or whether they are in a wheelchair. Here the user will be unable to identify the person if they are in different clothing. People usually change clothes once a day, leading to a clothing signature only working for that particular day, which can lead to further difficulties.

Utilising camera technology in general also has drawbacks. In a study by Stals et al. [137] to explore peoples emotions with places in a city using emotional recognition through a camera, they found that several factors affected the performance of their system. Poor weather, poor lighting conditions, clothing such as hoodies, other faces along with the person not looking directly to the camera limited the usability of Stals et al. System. Stals et al. stated that these were difficult to mitigate due to the lack of control that the researchers had on situations. While methods do exist to mitigate certain situations, such as using thermal infrared images to mitigate low light situations and glasses [75], these require a broader array of sensors along with more power requirements, resulting in a larger battery.

2.2.2 Alternative Automated Methods of Identification

While biometrics have been the focal point of research within the HCI community, there has been other research that has focused on other methods of recognition. Alternative identification focus on identifying the individual through other means such as recognising the devices that they carry and by following a predefined gesture, and not considering any personal traits.

2.2.2.1 Recognition by detecting smart phones

Smartphones offer an exciting possibility to allow identification of an individual. Within the UK, there are an estimated 53.7 million smartphone users [116]. Phones emit signals, many of them unique to a specific device and these can be used to identify a user of the device.

Research by Halperin et al. has focused on using WiFi signals from identifying people. The system would listen to WiFi address that phones transmitted as part of the WiFi hot-spotting standard 802.11. Halperin et al. used auditory mapping of the distance between the phones and gave audio feedback based on the location of the phone to the individual. Halperin et al. taught participants to identify a phone from the tones that their device had emitted [56]. Halperin et al. require that bystanders have hotspots to be switched on. Hotspots do consume battery power and can incur extra charges on users phone plans, and as a result, particular phone's hotspots will turn off when not in use [131]. Halperin et al. could not rely on standard WiFi addresses for a satisfactory result because of MAC Randomisation.

A MAC address is a unique address that's issued to every device that connects to the internet. Phones while scanning for a WiFi address broadcast their MAC address to allow access points to detect them. However, due to privacy and wardriving concerns, iOS, the operating system on Apple iPhone and iPad's now randomise their mac address while in their discovery phrase and only reveal their MAC addresses when connecting [8]. Current versions of Android do not offer MAC randomisation [50]; however, Android Q, the next version of Android scheduled for public release in late 2019 will contain MAC randomisation as part of its core stack [5].

Tracking someone through phone signals raises ethical and legal concerns. As tracking MAC addresses do not require a line of sight of a person, there is the possibility of an unscrupulous person driving around in a city at night when people are sleeping to try to identify where individual lives based on where their smartphone is broadcasting information. In unverified posts online, people have claimed that they logged over 30,000 WiFi hotspots and were able to correlate them to certain individuals just by using standard equipment [151]. The specific act of tracking someone through wireless signals is known as wardriving. Wardriving has questionable legal precedence, with it being illegal in some parts of the EU [94, 136]. People can wardrive without bystanders being aware that wardriving is taking place. Within the Hack a Day Article, Mehdi was able to detect whether they were sharing a commute with and that someone was looking for a

Dominos Pizza WIFI hotspots and were able to infer relations between commuters [151]. We assume that this technology can also be deployed for calculating where people live and work by wardriving. The ability to track an individual in this way leads us to question whether wireless communication technology endangers vulnerable people such as those at risk of domestic abuse. While open-source implementations do exist[7] with concerns about the legality and ethics of such an application, we felt that using a similar approach would be unsuitable.

2.2.2.2 Recognition by detecting gestures

Wearable technology can also be used to identify someone without having to wardrive. Bâce et al. [13] looked at using gestures for sharing an identity with a "HandshakeAR." Here both parties would wear a wearable such as a smartwatch and would detect when they perform a predefined movement, such as a handshake or a high-five. When the wearable detects this, it can then scan for the other device to see if they both made the same gesture. If this has occurred, they both share and display a business card [13].

While Bâce et al. work did not focus on supporting people with social interaction, this can be extended to assist people with social interaction and show less formal information such as "This is Tom, he sits with you on the bus into university, and he has a cat." Gestures can be something less distinct than a shaking of a hand as this may not be suitable for all contexts. A basic gesture could be the user waving their hands, followed by the bystander waving followed by the name displaying on a watch.

2.3 Legal And Ethical Issues Surrounding The Recognition Of People

2.3.1 Legal Concerns

In this section, we will cover the legal and ethical concerns of audio recording, covering consensual and non-consensual recordings of audio. We will seek a general overview of the legal field; however, due to the complexity of the laws, this section does not contain legal advice.

This section was written while current affairs had placed Google, Apple and Amazon practice of audio recordings by their virtual assistants (Google Assistant, Apple Siri and Amazon Alexa) under question in regard to consumer consent. Data from Whistle-blowers and media reported that Google Assistant was listened to by contractors and employers to analyse whether the AI system correctly processed this audio. Employees had access to raw and intimate audio that may be recorded by the user accidentally. As a result, Google has been ordered under the General Data Protection Regulations (GDPR) to suspend all recordings of audio data within the EU by The Hamburg Commissioner for Data Protection and Freedom of Information [139]. Apple's Siri has suspended all audio samples globally [60] while Amazon's Alexa is under review by the National Data Protection Commission in Luxembourg [78]. A class-action lawsuit has also been brought forward to Apple within California surrounding a recording of a minor [144].

The UK does not have a guarantee of privacy within the law [34], but people can expect privacy in private life such within the home, through correspondence and a private relationship with family and private life under Article 8 of the Human Rights act 1998 [153]. However, no law covers the recording and capturing of people in public. According to police documents by the Association of Chief Police Officers of England, Wales and Northern Ireland Communication Advisory Group obtained by Institute Of Amateur Cinematographers (IAC) there are "no powers prohibiting the taking of photographs, film or digital images in a public place." Police can use the Terrorism Act 2000 to question reasons for filming, however, cannot request deletion without a court order. However, private property owners can impose limits on photography and filming [69]. The only time recording someone in public is illegal is if it breaches the Human Rights act 1998.

[7] Work by Medi: https://github.com/mehdilauters/wifiScanMap

The only time it is illegal to capture and record audio in public is from phone calls under The Telecommunications (Lawful Business Practice) (Interception of Communications) Regulations 2000 except for national security, prevention and detection of crimes or detecting unauthorised use of a telecommunication system [88].

However, the processing and storing of data raise questions about its legality. Under GDPR and guidance issued under the UK Information Commissioner Office (ICO) any data that can be used to identify an individual, even if it requires combination with other information, is classed as personal information and is covered under GDPR [68]. So while the act of recording is legal, the act of storage may be illegal under GDPR.

Any individual has the right to be informed if an organisation carries any information on that individual along with privacy information explaining the purpose for the storage of information [67]. They must offer services to view and retrieve all data that they possess on a person and the ability to delete all information [68].

If data is stored on a device and does not get sent to a central server, the user of that application becomes responsible for the information. Theoretically, anyone who has their voice data within the application can request their data and its deletion from the device. While guidance from the European Commissioner states that " [GDPR] gives individuals the right to ask for their data to be deleted and organisations do have an obligation to do so [38]." There is no guidance given on individuals of a service carrying data, making it difficult to understand how and whether individuals do have to follow GDPR.

In theory, an individual can request users of a service to remove them from the system or view any information that they possess on them, similar to an organisation would have to. However, the individual will have to approach all users as they are separate entities. As a result, the user must provide the individual privacy information about the processing and storage of audio as voice data is personal information. Nevertheless, the act of capturing audio if done publicly with no information that violates the Human Rights Act 1998 it is perfectly legal.

2.3.2 Ethical Concerns

The general public have traditionally disliked having biometrics captured. The US authorities first widespread use of facial recognition was at the 2001 Superbowl games was perceived to be controversial by the general public. However, after the September 11th attacks on the US, authorities believed that widespread facial recognition alerted them to the attack, and the general public has started to accept the use of biometrics tracking [19].

Surveillance has also grown in other domains such as within organisations; however, professional guidance has struggled to keep up. The current ACM Code of Ethics and Professional Conduct guidelines discuss the use of information when only required and for legitimate needs; and, current ACM guidelines do not give guidelines surrounding bystanders [54]. While concerns about data handling are within the remit of the ACM Code of Ethics and Professional Conduct guidelines, the general population are still concerned with their imagery, with patients becoming concerned that sensitive images may become available online [86].

Work by Nebeker et al. demonstrates that not all research ethics committees fully consider the ethics of bystander capture in studies. As Nebeker et al. demonstrated in ethical applications to several Institutional Review Boards surrounding imagery and location studies, they did not consider potential concerns regarding how the data was stored or the ethical concerns of bystanders [114].

With recent events surrounding Facebook's handling of the Cambridge Analytica scandal where information on Facebook was being utilised to target voters in the 2016 US elections [23], the ethics of data handling should be more scrutinised, and where it might have one, its impact on subjects wellbeing highlighted. Work by Ball et al. has demonstrated that workplace surveillance has consequences for employees,

affecting their wellbeing along with productivity [14] which is supported by further research by Jeske et al[74].

Researchers have argued for ethical requirements for over twenty years of [99]. However, individual researchers state this highlights a significant ethical issue with how data is handled at present within the field of Computer Science [159]. These ethical concerns are evident with smart glasses, where bystanders demonstrated concerns of being filmed by individuals with smart glasses [31, 83, 133].

2.4 Bystander Privacy

The privacy of bystanders is a crucial factor in the acceptance of any technology. If people do not feel comfortable surrounded by technology, they are unlikely to engage with the technology.

Smart glasses have previously been available to the general public through Google Glass [GG]. When Google launched GG in 2013, bystander privacy became a significant concert with the general public and the popular media's covering the lack of bystander privacy that came from GG [11, 79, 98, 149] Many GG wearers were given diminutive nicknames such as "Glassholes" [27]. Google attempted to combat the bystanders concerns with an etiquette guide issued to glass wearers [51] but this did not lead to a change of public opinion. Many of the privacy concerns derived from the lack of visual cue that stated whether GG was filming or not. Participants in a study by Singhal et al. we're not aware when they were being filmed by researchers using GG [133].

However, when compared to a researcher filming on a smartphone, participants were aware and altered their direction of movement, speed up and avoided contact with the camera. When interviewed by researchers, participants were more comfortable being recorded by smartphone. Participants stated they found it challenging to identify a GG wearer while participants were more comfortable with mobile phones citing their prevalence [133].

Research by Denning et al. also demonstrated that bystanders might be uncomfortable with smart glasses. In an interview of 31 bystanders, there was a significant split between people who were indifferent or demonstrated an adverse reaction to smart glasses. Participants found that smart glasses were more subtle compared to other methods of recording and a set of participants stated it was relatively easy to start a recording of an individual. Participants also stated that they wanted to give their permission first to people recording with smart glasses, although they accepted that this might be challenging [31].

Work by Koelle et al. demonstrated that the usage of smart glasses were perceived critically by bystanders, although users were more likely to perceive smart glasses positively. However, the research also found that female participant were more likely to express negative feelings towards smart glasses [83]. Koelle et al. believed that smart glasses are perceived critically due to the unfamiliarity and time of exposure was a factor similar to the Walkman effect, where negatives attitudes diminished over time [62]. They thought that as smart glasses become more prevalent in society, these concerns will also diminish.

The work by Singhal et al., Denning et al. and Koelle et al. highlighted that various bystanders were concerned with their privacy around smart glasses, and might not feel comfortable in the presence of wearers of smart glasses if they were aware that they are in the vicinity of a user. However, their perceptions seemed to alter once they were aware that people were wearing smart glasses as AT, especially for those People with Visual Impairments (PVI).

Research by Profita et al. demonstrated that bystanders made a more positive assessment of people with GG once they became aware that the wearer utilised GG as an AT device to support people with PVI. However, participants highlighted that they did not consider GG as positively if GG wearers used the data in a way that participants perceived as being for non-assistive purposes, especially for a photo album. Participants felt that this was not required as the wearer was blind, therefore not needing a photo album [124].

Research by Ahmed et al. further demonstrated that participants are willing to share more information

with PVI compared to sighted people. Participants were also willing to share further data with PVI if further access control and with assurances that data is for Assistive Technologies' purposes only [3].

However, it is crucial to consider that not everyone wants to disclose their disability. Many users of AT feel that their devices carry a "social weight" that drastically impacts the adoption and the use of devices [30, 124]. Consequently, many people who require AT may not use it when needed, for example, individuals who require a white cane due to vision impairment may abandon it to avoid drawing attention to themselves and be perceived as PVI [118, 124]. McNaney et al. also highlighted that patients with Parkinson's disease who had worn GG for a period of time were concerned about their privacy. Participants were worried that relatives might abuse the video linking features as a way of monitoring what they were doing [103].

It is crucial to consider if or how users will accept AT technology developed to support them and whether it could lead to abandonment by the user. If the user perceives themselves as disregarding bystander privacy, they will likely abandon the AT. Currently, AT abandonment rate is estimated to be as high as 75% with many user citing stigma of using AT as the driving force for abandonment [105].

2.5 Research into Speaker Recognition

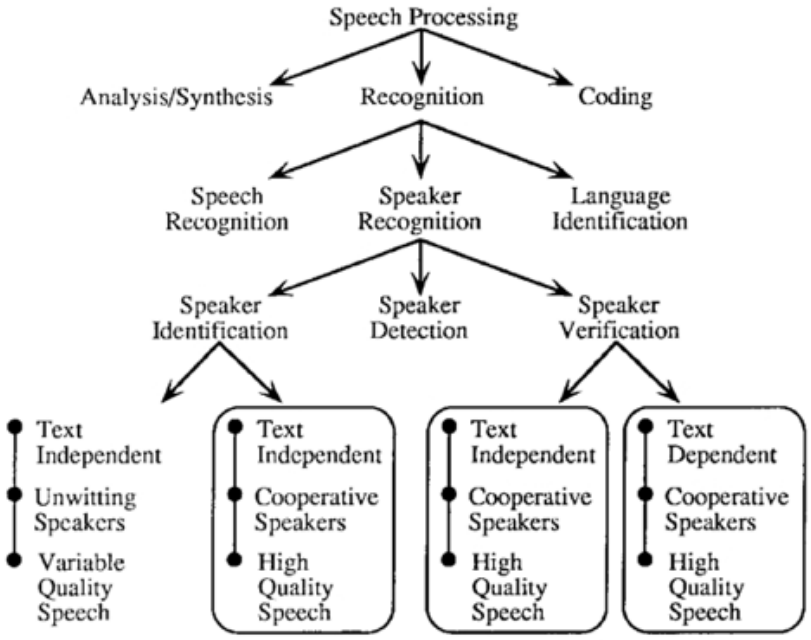


Figure 2.2: Breakdown of the area of Speech Processing by Cambell [24]

This subsection relies on the taxonomy of speaker recognition which we develop in 5.5.1.

Speaker recognition is a research area within speech processing which is a subsection of Machine learning. Speaker recognition's task is to answers the question of "who is speaking." It is different from the task of speech recognition which answers the question of "what is the user saying?". While these algorithms can be used together, they achieve different tasks. In image 2.2, we can see a breakdown of the fields of speech processing by Cambell.

The first commercial use of speaker recognition was found in Julie the Doll, a child's doll that was re-

lased in 1987 which was capable of primitive speaker and speech recognition. Julie's approached relied on digital signal processing that monitored pitch contour trends [41]. Further work within the area of Speech recognition focused on access control [24], however, it is worth noting that there were concerns about the precision of these applications. The main area of application for speaker recognition today is working in tandem with speech recognition (what text is being said) and natural language processing (what does the user mean by what was said) in virtual assistants such as Google Now and Apple Siri [134].

The first work within speaker recognition was by Pollack et al. in 1954 [122]. However, the first research of note is by Atal in 1974 that relied on Pitch Contours. However, Atal's work is an interesting starting point for the field of speaker recognition as it demonstrated that computers were able to identify who was speaking accurately. Atal's work was a text-dependent closed set algorithm that analysed patterns in a 20 dimension vector achieving an accuracy of 98%. Atal's algorithm was a lab-based study on 10 participants with 20 phrases[12]. A closed set algorithm can only identify people based on a set phrase.

The first text-independent speaker recognition was carried out by Markel and Davis, which could identify 17 speakers. Markel and Davis achieved an accuracy of 98%. Nevertheless, it is worth noting with a large amount of data per speaker was required, with 30 hours of audio required for training making it difficult to reply in real world situations [97].

Around 1991 speaker recognition started to utilise traditional machine learning networks such as Hidden Markel Models (HMM) and statistical models. Traditional machine learning networks are based on frameworks which give the researchers the ability to adjust how the algorithm works by configuring what information is fed to the algorithm and the parameters of the algorithm. Work by Cambell noted that researchers had started to develop pipelines that contained feature extraction and pattern matching, [24] which is still present in speaker recognition today [134]. However, at the time of writing (1997), the field of speaker recognition did not utilise neural networks [24].

Further work by Reynolds in 2002 found that much of the research within speaker recognition surrounded template matching, nearest neighbour algorithms (such as K-means), HMM and neural networks. Reynolds stated that current neural networks were identifying whether the voice belonged to the set of known voices as algorithms such as K-means performed poorly with this task and that neural networks were not the primary recognition algorithm. It is worth noting that Reynolds stated that many of the features and techniques used for the task of speaker recognition are being used in speech recognition at the time of writing [128].

Work by Beigi in 2011 stated that GMM algorithms was the most prevalent approach to speaker recognition, however a significant amount of work into using SVM as the primary algorithms. SVM approaches presented optimisation problems due to the high dimensionality of the audio data utilised, which resulted in high computational and memory requirements for training. Neural networks are briefly mentioned as a means of recognition, but neural networks was not a primary topic [16]. Beigi stated that the area of speaker recognition has expanded from access control and children toys to finance, legal, security, audio/video indexing and diarization, teleconferencing, surveillance, along with many other areas [16].

Neural networks as a primary recognition algorithm were first used by Lei et al. in 2014. Lei et al. replaced the GMM component with a neural network and achieved a 30% improvement over traditional GMM algorithms [89]. Further work by Richardson et al. found that neural networks gave significant potential for speaker recognition [129]. Today commercial applications are utilising neural networks such as Apple "Hey Siri" and Google Voice Match, which we develop in in chapter 5.5.2.

2.6 Tipping Point Of Acceptance Of Technology

In this section, we will develop the narrative for the acceptance of machine learning technology and its accuracy, specifically the acceptance of Automatic Speech Recognition (ASR), commonly known as speech-to-text. ASR became an area of research in the 1980s for dictation of documents through speech to text. In

the last eight years, ASR has become a vital part of virtual assistants such as Apple Siri, Amazon Alexa and Google Assistant. ASR is utilised to convert raw audio data to strings which are fed into a Natural language processing algorithm to infer and execute commands [49].

Munteanu et al. evaluated the acceptance of transcripts of internet broadcast videos generated by an ASR at different Word Accuracy Rates (WAR) and a human transcription at 100% accuracy. Munteanu et al. also evaluated the acceptance of accuracy with undergraduate students on text to speech algorithms on transcripts generated from internet broadcast videos. Munteanu et al. found that transcripts of 75% Word Accuracy Rate (WAR) were acceptable, but a WAR of less than 55% was unacceptable [108].

Work by Stalk et al. further demonstrated that people who had access to high-quality (above 84% WAR) transcripts were less likely to abandon them and people with moderate to low transcripts (less than 70% WAR) were more likely to abandon transcripts. However, it was noted that only half of the participants found that perfect transcripts were very readable with a further 57% stating that they were very comprehensible, although the researchers commented that this could be a result of missing text and or grammatical errors which are not present in speech [138].

From the work of Munteanu et al. and Stalk et al. we can conclude that there is a tipping point that exists between 75% and 84% where people perceive technology as being accurate, however, people did not perceive these results as very accurate. We consider this to be the benchmark of any solution to in supporting people to recognise others.

Research into ASR has led to a significant increase in accuracy within the last ten years. Traditionally, state of the art ASR has consisted of combined hidden Markov models and Gaussian Mixture Models (GMM) [49]. TIMIT, an algorithm that was used as a benchmark for phoneme recognition, received a 26% phoneme error rate [46].

Researchers in ASR did not see a compelling reason to switch to neural networking until the late 2000s where much larger models replaced GMMs. Further, more massive labelled datasets were being used by researchers to train the models, such as Mozilla Common Voice [107] along with pruning pretraining, which led to a significant improvement of 30%. Within two years, most of the industrial products had moved to incorporate deep neural networks replacing the traditional models [49].

Labelling of speech data by people became prevalent, and this led to a further improvement of ASR.

Figure 2.3: Google Trends are showing interest over time for the search term "call dad." Google operates its virtual assistant, Google Assistant, which uses the Google search engine. Note the significant drop that occurs in 2016 (Above "note") which was the result of Google changing how it collected trend data. Image source: [53]

Much of this data came from virtual assistants such as Siri and Alexa. Google trends for the search terms "Call Mum" and "Call Dad", as seen in Figure 2.3 showed a significant increase from June 2012 until 2016 when Google changed how they tracked trends [53]. Google trends demonstrate that people are using virtual

assistants.

Further analysis of the usage of virtual assistants found that people aged 45 and older were most likely to use virtual assistants, with the average user of virtual assistants being a 52 year-old woman using 1.5 hours a month with virtual assistants [121].

This led to Sundar Pichai, CEO of Google stating at Google I/O 2017 Developer Conference, that they had achieved "significant breakthroughs" and that Google speech recognition was nearly on par with humans. The popular press stated that Google had achieved 95% human accuracy [1].

Industrial products have stopped publishing their results on accuracy. However, Google has published that 41% of people who own a smart speaker found it was similar to talking to another person [81]. Despite this lack of data sharing, we do know that in section 2.3 Google, Apple and Amazon labelled audio data. At the time of writing, it is uncertain whether this had any impact on the acceptance of virtual assistants however, this will likely set a precedent for the acceptance of virtual assistants in the future.

2.6.1 Design changes based on the literature review

From reviewing the literature, we can see that previous work in wearables demonstrate that the medium is suitable; however, there is very research regarding into smartwatches, the main focus these studies focusing in smartglasses. We can also see that the use of biometrics to identify an individual is suitable, but privacy concerns surround the use of biometrics.

Reviewing the literature surrounding speaker recognition and the tipping point required for technology, we feel that the use of speaker recognition offers the potential as a means of recognising an individual. From this literature review, we have added the following to the specification:

- The application interface should be accessible through a wearable device (for example, interference can take place off devices such as on a phone or server).
- The application must respect the privacy of the bystander and the user.
- The application should achieve high accuracy.

2.7 Conclusion

In this chapter, we have explored the previous research within the HCI community to support people in recognising other people through biometrics and other identifiers. We have also explored the legalities and the ethical concerns for filming people, which demonstrates a grey area within the law and also ethical considerations. Further, we explored the tipping point for people to accept the accuracy of algorithms and found that the number lies around 75% and 84%, currently the industry is targeting 95% of human accuracy.

3 Design Process

In this chapter, we will discuss our design process. We first ran a session with the third-year computer science students studying user experience and usability to understand the critical points of user interaction with our proposed system. We then submitted a workshop paper to CHI'19 addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction. We later developed a high fidelity prototype application, SR1 (speaker-recognition V1) to emulate the design flow that the design students produced and to correct any design flaws. From S1, we then produced Pwy, (Welsh for 'who') an application to use within our participatory design workshops as a design critic and in evaluating how the application use in a conversation.

3.1 Design Sessions with Students

To begin the design process, we wanted to understand how people less familiar with the basic concept of social support through voice-recognition felt about the concept. To this end, we worked with around seventy third-year students to develop user journey maps in order to understand critical points of the user interaction with the application and to understand possible pain points that users may face. Journey maps are a growing tool with User Experience (UX) researchers as it allows focus on the UX of a product over time [63].

User journey maps are widely used within the industrial design sector as it allows organisations to understand the interaction of the system from a users point of view. The information that user journey maps capture provides us with a clear set of challenges that our users may face [77].

User journey maps typically display the significant phases of a user interaction along the top horizontal axis. The vertical axis will display actions with links between each action stating how users interact and feel between actions allowing an extra dimension displaying how users may interact with a system [63, 92].

For the first part of our design, we approached an third-year UX class aimed at computer science students, asking them to design user maps of the exploration and the use of our application. These user maps then formed part of the assessment [1]. students received the following guidance as their design brief:

You have been tasked with designing a smartwatch interface that will capture audio and display the names of people speaking in the environment around you Using the system requires you to capture a short audio clip of a speaker talking (4-5 seconds) and label it with their name Users of the application will include:

- *Older adults with memory problems (Dementia, MCI [2])*
- *Younger adults with memory problems (TBI [3])*
- *People who have problems remembering names and need to (Teachers, salespeople)*

End goals should include capturing and labelling a speaker and using the watch in conversation.

[1] This activity was marked as binary - 1 mark for user maps being present, 0 otherwise.
[2] Mild Cognitive Impairment
[3] Traumatic Brain injury

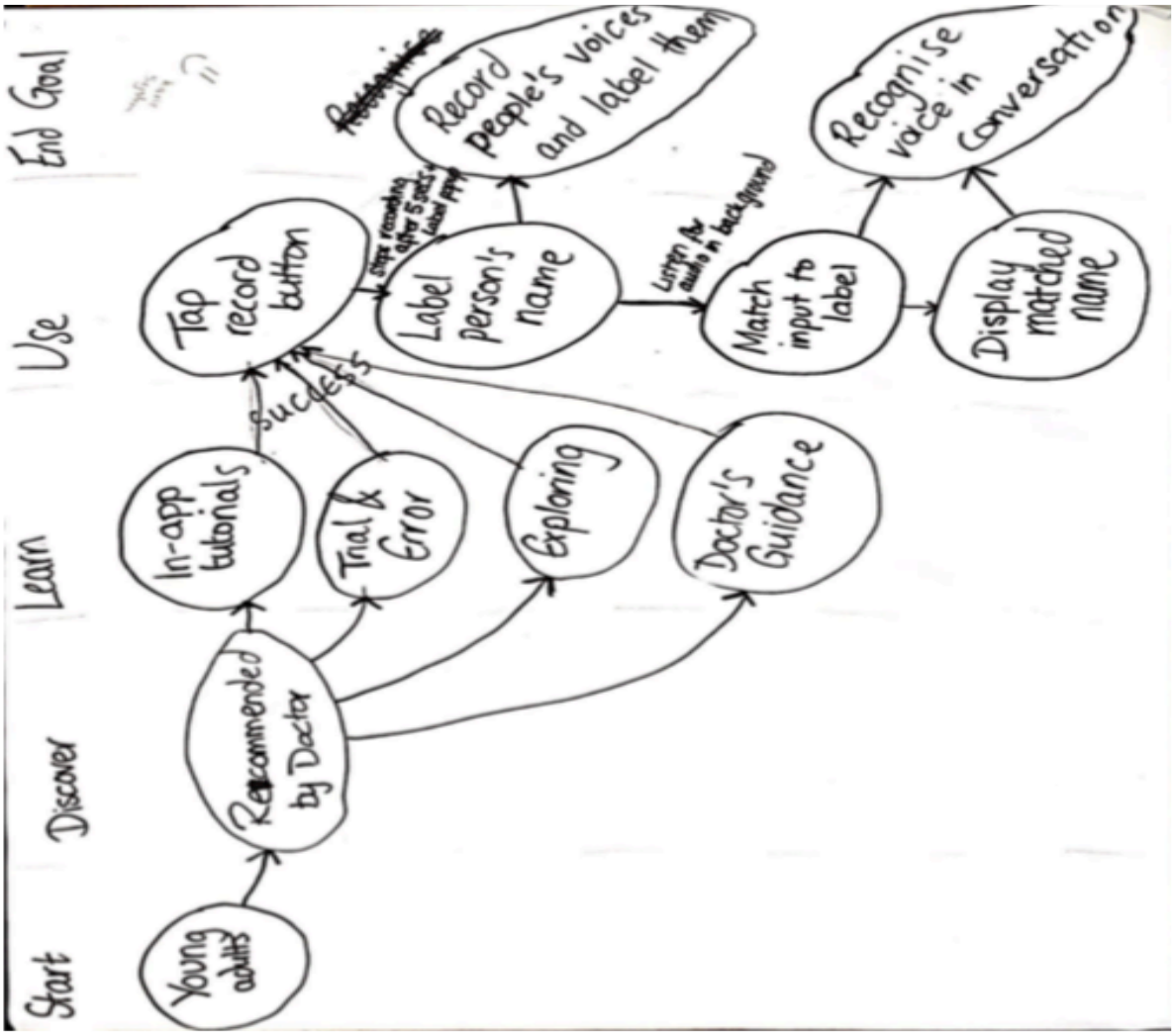


Figure 3.1: An Example Jorney map that was drawn by a student in CSC349 in Swansea Univerity as part of their assignemt

Students worked through five significant phases: where would users start the process, the discovery of the application, learning to use the application, using the application and end goal. For the start phase, many students focused on young adults or people with memory difficulties. We expected that this group of users to be predominantly discussed by students by their demographic.

Students submitted this work along with other workshops as part of an assignment. Once the course-work was available to us, we extracted our workshop work and sorted it into completed user maps and non-completed user maps. We then subdivided the work into the demographic category and thematical analysis noting all the key points that were brought up by participants.

Discovering the system In the discovery phase, the majority of students stated that users would discover the application through an advert or a recommendation of a medical expert. Students felt Students suggested that the application could become standard and became commonplace, and people would discover the application through word of mouth.

We hypothesised before this activity that this application would be recommended to participants if it became commercially available by medical professions. However, we did not consider prior to the students input that

the application could become commonplace.

Learning the system. Within the learn section, students suggested that users could be able to learn to use the application through an in-app tutorial such as a video along with trial and error. Students highlighted that trial and error could be difficult for individual users, and this was possibly a weakness within the system. A potential resolution for this would be allowing users to review tutorials or given guidance from a medical expert. Another solution which another group suggested was to place a help button that could contact a family member who understood the app and could give support to the user.

Using the system. For using the application, students suggested that the application should always listen for conversations within a conversation environment for a known person to reach identification. Students suggested that using the application should require as little effort as possible on behalf of the user. They suggested keywords such as "hello" could be utilised to trigger listening for known voices, which could also add a level of privacy to the users.

Students stated that when a user wanted to add a person to the system the user would press a button or remembering a time where the app sampled the new voice and could return back to the app after the conversation. If the user presses a button, this allows interaction with the speaker being involved themselves by helping the user, adding them to the application.

Failure of the system. Students also noted the chance of failure of the system, and how the application can overcome these by having the system displaying confidence scores. Other students stated that the application could keep trying to identify voices until the application was confident whom the speaker was, by testing several voice samples.

Further discussion with the students highlighted the need to make users feel confidence and accomplishment, however, many noted that if the system failed the users would feel frustrated, annoyed and confused. Students were aware that this application should not be a replacement for remembering names, although none of the students we spoke with were able to come up with a viable solution to this concern.

3.1.1 Non User Journey Submissions

Figure 3.2: An example design idea by a CSC349 Student

In this session, some students ended up creating other design artefacts instead of the User Journey map.

These included a design of the application and a flow diagram. We thought these designs were attractive; however, did not fit into the above section.

One student developed designs for an application that included two buttons, listen and add, where one button listened to the conversation then alerted them to a speaker that can be seen in figure 3.2. This screen would then display notes on the speaker. The user would be given conversation guidance (such as their name and hobbies). This conversation would then allow the app to them infer the name of the speaker saving the user typing.

Another student created a flow diagram of the application where the user holds down a button informing the app to start listening then the user releases of the button when the watch has enough audio. The watch then works out the name of speaker and alerts the user and tells the user to say the name of the person or whether it was a new person. Due to the anonymity of the coursework submission, we could not follow up with the students as we did not see there work in class. We believe that the student theory is that the device controls the listening to ensure that it identifies the speaker's name accurately or alerts the user to a new person.

3.1.2 Design changes based from student feedback

From reviewing the literature, we can see that previous work in wearables demonstrate that the medium is suitable; however, there is very little research regarding into smartwatches, the main focus these studies focusing on smartglasses. We can also see that the use of biometrics to identify an individual is suitable, but privacy concerns surround the use of biometrics.

From the design process study with design students, we found that for the design, the app should focus on being intuitive and straightforward to use, and we have added the following to our specification:

- Students stated that the discovery phase and learning the application should be intuitive
- The use of the system should be simple relying on as many small operations to the user as possible. For example, we should investigate the use of keywords or merely using a single button press
- Further investigation is required to understand how failure can affect the use of the application and how end-users would want to handle failure.

3.2 Paper Submission to CHI'19 Workshop: Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction

As part of our design process, we submitted a paper to CHI 2019 Workshop on Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction titled *Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury.* This paper can be read in the Appendix in section A.

In this paper, we wanted to discuss how, by supporting the work of people with SIIDs, we could remove the stigma from people who needed to use an application for medical reasons, such as those with TBI. We wanted to highlight how the work of the SIIDS community has benefited from supporting accessibility technology and vice versa. We highlighted work by Trewin [143] on how users in a SIIDS who require sticky keys found it challenging to configure correctly and may not change it, however people with motor disability would spend time to adjust it correctly [143]. We stated that a more proactive approach is required for people with SIIDS, which can then support people with motor difficulty.

We also highlighted the work by Wobbrock that stated that once the underlying cause of SIIDS is understood, research should focus on supporting people with disabilities [152]. From this examination, we can

establish relationships between these conditions and understand a solution which encompasses both SIIDS and people with impairments.

We also highlighted work by Tigwell et al. found that design clients often find accessibility overbearing and that current well-known frameworks, such as Apple Human Interface Guidelines not supporting SIIDS with Apple's low contrast fonts being difficult for people with no history of vision impairment challenging to read [117, 141].

We further highlighted during our paper presentation that a large amount of the populations had mobile phones. Many of these phones contain an array of sensors which people already use as AT. We highlighted how the cameras are used to help people view small text by zooming in. We also highlighted the current high abandonment of AT which many users cite stigma and that we wanted to use standard devices such as mobile phones to help patients.

As part of the workshop, we carried out a 4-minute presentation which we discussed our paper briefly. Our slide deck is available to view in the Appendix under Section B. We focused on the high abandonment rate of AT technology and how mobile phones can be used as AT and how our application (SR1 as development for Pwy was at an early stage) was able to support SIIDS.

We also jointly carried out scenario card exercises to develop tools to support people with SIIDs to understand their situations and what they wanted to develop. Scenario cards stated a situation, the impairment (such as having an umbrella in your hand) and the task that the user wants to carry out. In our groups, we then developed a concept of a tool that could support users with SIIDs. The scenario cards inspired us to use scenario cards as situation cards in our own participatory design workshops to prompt participants.

3.3 Development of an Application to Research Using Speaker-Recognition to support Social Interaction

As part of this research, we developed two applications, Speaker-Recognition 1 (SR1) and Pwy. We developed SR1 to understand the core interface of the application as a high fidelity prototype. SR1, however, was mainly based on Storyboarding features in Xcode, an integrated development environment by Apple for iOS, WatchOS and macOS, and did not consist of any watch to phone communication. For the application to work as a tool for our participatory design workshops and evaluation studies, it would require a complete rewrite of the codebase, which leads to Pwy, an MVC application that allowed for communications to the Apple watch and a working interface. In this section, we will discuss both SR1 and Pwy development.

3.3.1 Speaker-Recognition 1 (SR) Prototype

SR1 is a storyboard based high fidelity iOS prototype that was developed to run on the iPhone X, and Apple Watch 42mm screen sizes. It consisted mostly of storyboards and showed certain functionality, such as the listening screen on the watch and playing audio which is programmed in Swift, an object-orientated programming language[4]. SR1 did not consist of any communications between the iPhone and Apple Watch and any machine learning.

We developed SR1 to demonstrate our work to potential stakeholders early on in the development life cycle. We also developed SR1 to allow us to give a demonstration of the application at the above mentioned CHI2019 workshop. These designs allowed us to design Pwy, which contained databases, watch communications, microphone access and machine learning.

[4] https://swift.org

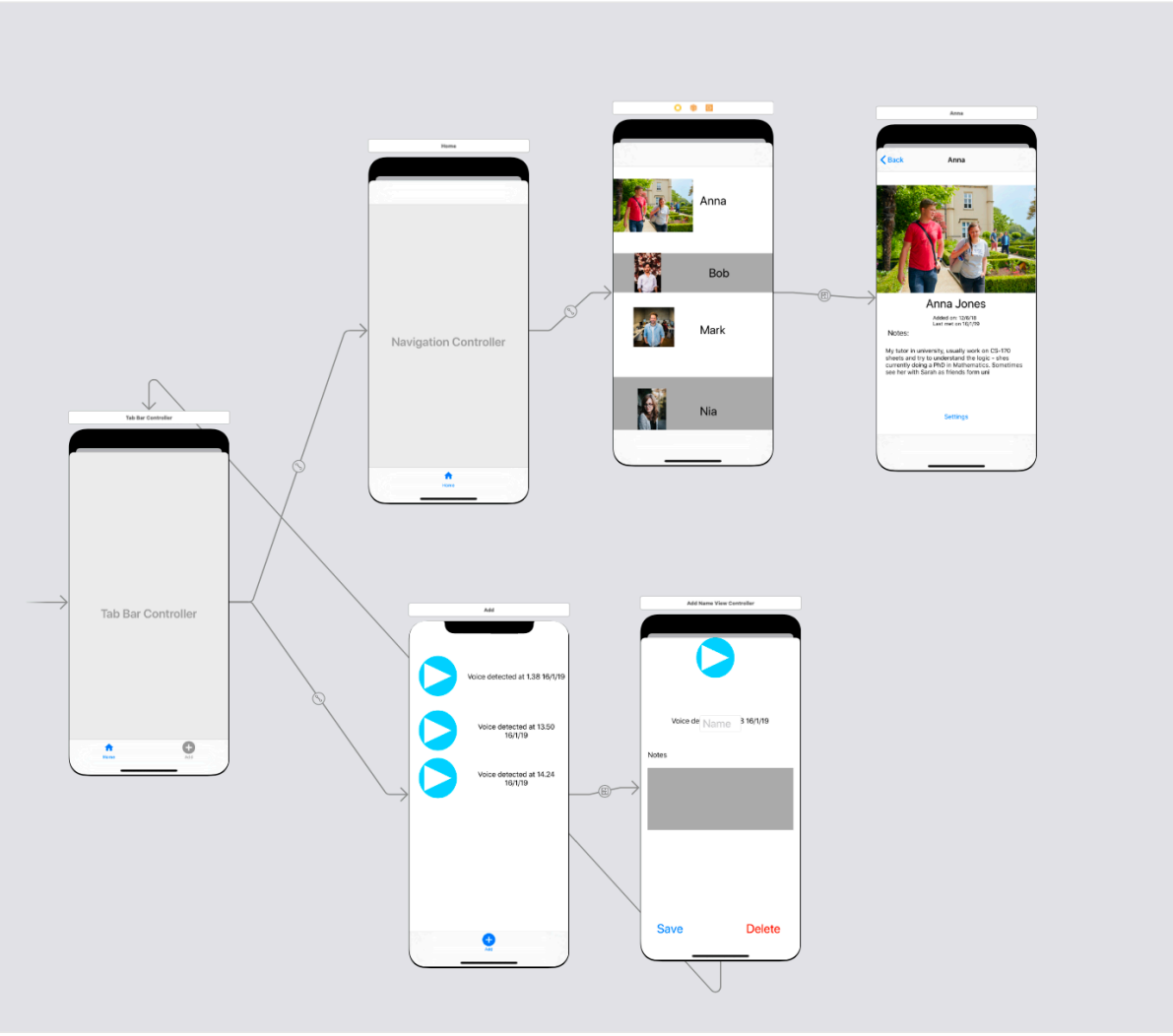
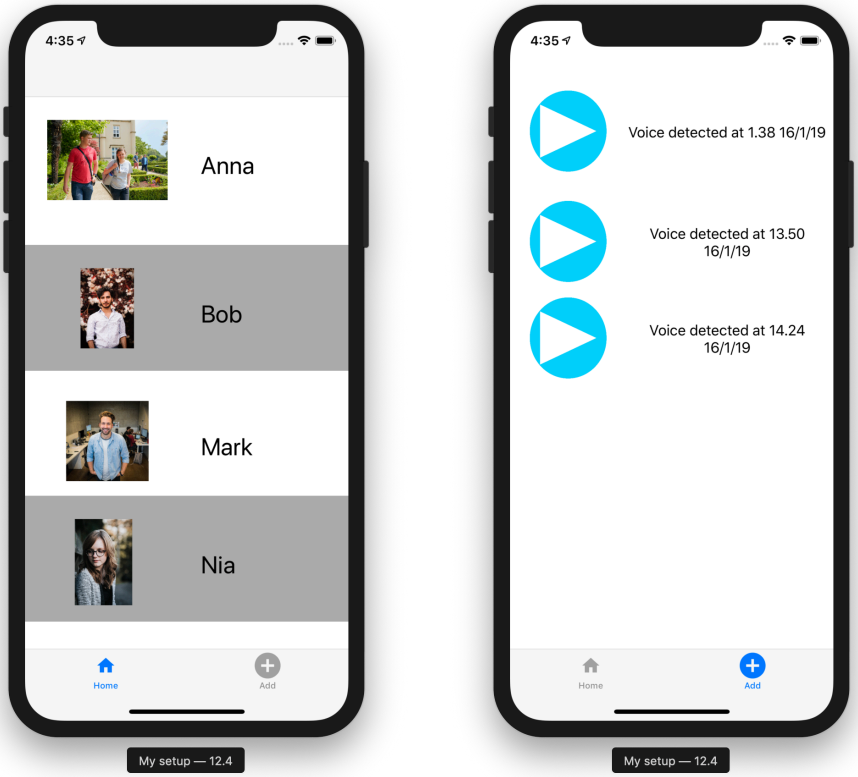


Figure 3.3: An image of SR1 storyboard in Xcode

3.3.1.1 iPhone application

We developed the iPhone application as a voice manager interface to allow them to link new voices to names and not as the main method of interacting with the application. We did not want users to use the phone in conversation as we felt it was too obvious that the app was being used in conversation compared to using a watch. The iPhone application consisted of four primary screens:

- Home Screen - this displayed a list of people that are known to the application. Tapping a cell would take the user straight to the More Details screen corresponding to that person. This can be seen in Figure 3.4a
- More Details Screen - this displayed more information on a person from the Home screen. This contains a picture of an individual along with their name and notes. This can also be edited by the user.
- New Speakers Screen - this displayed a list of people's voices that have been detected but not been added to the application. This can be seen in Figure 3.4b
- Add Person Screen - this display allows people to add information to the application such as a name and notes and also listen to their voice.



(a) SR1 Home Screen (b) SR1 Add person list

Figure 3.4: The iPhone application of SR1.

SR1 did not store or record voices. Adding people to the application did not result in anyone actually being saved and also did not result in the Home Screen displaying the new person.

3.3.1.2 Watch Application

We developed a Watch application for SR1, which we designed to be the primary application used by users during conversations. This application consisted of three screens which were:

- Tap to listen - This screen allowed users to trigger the watch to listen and would display Listening as seen in Figure 3.5a
- Listening - This screen gave the users an illusion that the screen was listening and would wait 10 seconds before a segue to "Name Found." This screen is Figure 3.5b
- Name found - This screen would display the name of the person along with a picture and notes. In SR1, this screen will always display "Anna" along with notes about her being a university tutor, as seen in Figure 3.5c.

3.3.1.3 Findings from SR1

As previously discussed, there were no communications between the apps occurring, and all the details were hardcoded, meaning nothing could update without a new source code compile would be needed for change to occur. We only developed SR1 as a prototype which, as a result, lead us to abandon its codebase for the development of Pwy. However, some interesting findings did emerge from this prototype.

(a) SR1 Home Screen

(b) SR1 listening screen

(c) SR1 Person has been found - an image along with notes is also displayed

Figure 3.5: The Watch Application for SR1

The most significant alteration for this application was the removal of the ability to listen to the audio. We felt that this posed a privacy concern, and we felt stating the time of the discussion would be sufficient information to remember who the person was.

We also removed images displaying on the watch as it led to the user having to scroll on the watch, which we felt made the user appear rude to the person they are in a conversation with. We wanted to make it clear whom the user is talking to and while seeing a picture of the person would help against calling someone by the wrong name, we had concerned from where they would have gathered the picture from, and it could lead to stalking.

3.3.2 Pwy

The Pwy system is envisioned as being a primarily smartwatch-based app. When in conversation with someone, the wearer will be alerted by vibration on their wrist when the Pwy system has established whom they are talking to, and their name displayed on the screen as a discrete form of support. Pwy uses the smartwatch microphone to capture audio as it does not require the phone to be out of the user pocket

We designed Pwy intending to protect bystander privacy. We wanted to utilise a Machine Learning algorithm that extracted specific features from people's voices to stop and did not store audio in order to stop any possible data breaches and ensure that no reverse engineering could occur. We also wanted inference to occur on the phone using it's neural network chip rather than send data off-device for processing. Through on-device inference, no external server would receive any of the voice data. If a breach of information occurs, only information relating to the bystander's interaction with the user would be compromised and not all users. On-device inference, however, makes all users a data controller for all data on their phone under GDPR law so this still introduces some complexity to the system while respecting privacy rights more than, for example, simple video capture and processing.

Pwy is an iOS application that is designed to work on all iOS 12 devices and watchOS 5 devices or later. Pwy is primarily designed for the iPhone as it requires an apple watch. Currently, an Apple Watch requires an iPhone to activate.

Pwy design elements are derived from SR1 and also utilises a Realm object database[5] that is an alternative to Core Data, Watch communication and microphone permissions. Pwy is designed to be easily modified based on feedback from future design works and to allow us to add activities such as Wizard of Oz.

3.3.2.1 Design

Pwy's main design elements are derived from SR1. However, parts of the application's interface (Table view cells, Information screens) came from HomeBP, an iOS Blood pressure application that was developed for the NHS and is awaiting approval.

Similar to SR1, the Pwy iPhone application is envisioned as a voice manager and not as the main interface to be used during conversations. Pwy iPhone consists of three main screens, with more details and Add person screen merged into one screen in the UI to reduce code repetition. We removed images in this build to increase privacy. Watch communications in this build were also implemented. The iPhone application consisted of the following:

- Home screen - This screen shows a list of known people to the application along with the date they were added to the system and when the user last met this person. Tapping on the cells takes the user to More information. Swiping right on a cell presents the option to delete a person from the database, thus losing their voice data requiring them to be added again if needed.

- More information - This screen shows the person's name along with notes and allow the user to save changes to the application. The user would visit this screen either from the home screen or adding people screen.

[5] https://realm.io

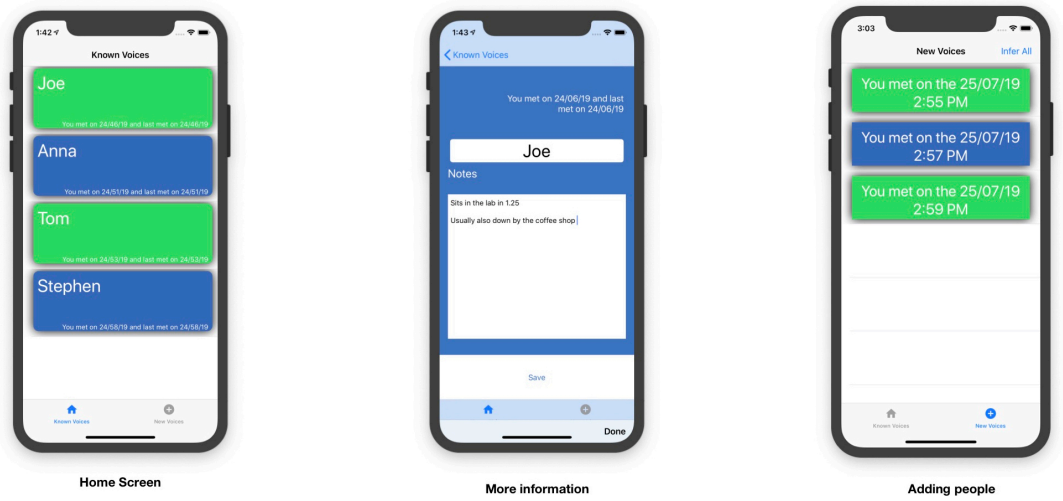


Figure 3.6: SR1 Home Screen

- Adding people - This screen shows the date and time the user meets an individual and tapping on the cell saves the person to the database and navigate the user to More information screen. Swiping right on a cell presents the option to delete a person from this list, thus erasing their voice data.

The iPhone application can be seen in 3.6.

Similar to SR1, the Pwy Watch application was designed as the main point of interaction during conversations. While we tried to keep the screens similar to SR1, alteration between listening to an added screen for sending data was required to allow the application to work. The Pwy watch application consists of four screens:

- Tap to listen - tapping this sends the user to Listening screen
- Listening - Listening screen displays the system's listening screen that allows the user to tap to listen and record, before pressing "processing" that then navigates the user to Sending data
- Sending data - This screen is required to send information to the iPhone to wait for a response
- Person found. This screen shows the name of the person and relevant information about them and then allows the user to navigate back to Tap to listen to enable them to listen to another voice

3.3.2.2 Implementation

In this subsection, we will give a brief overview of the key components in Pwy's implementation.

3.3.2.2.1 Database For our database, we used a Realm database[6], a non-SQL open source database in Swift. We utilised realm due to prior knowledge of the framework in HomeBP. All database work was on the iPhone application with no database on the Apple Watch. Our database consisted of two models - SpeakerEntry, which contained information about the person such as notes, along with temporal data and database lookup integer produced by the machine learning. AudioEntry was our second object that contained information on when the first meeting took place and a link to the WAV audio in File Manager.

SpeakerEntry was the main model that contained all the information that was required about a person within the database itself. This consisted of entities such as the name and notes of the person, however it is also able to carry information that was related to the database such as the WaveNet temporal results along

[6]https://realm.io/docs/swift/latest/

with the integer that a second model which we will discuss in Chapter 5

AudioEntry contains a pointer to the WAV file along with data of when the user met the other person, to help aid the user to recognise who they are speaking to. AudioEntry was designed to contain temporary data, and when a SpeakerEntry derived from AudioEntry, AudioEntry would be deleted along with the WAV file once the SpeakerEntry had the temporal data from the WaveNet.

3.3.2.2.2 Watch Communication To permit the iPhone and Apple Watch to communicate with each other, we needed to implement a platform to allow communications between both. We decided to utilise WatchConnectivity[7] to allow communications between the Apple Watch and iPhone. As the Apple Watch would send an M4A file, which is a standard audio file that is output by the Apple Watch Microphone, to the iPhone where it would be converted to a WAV file, we needed to send a large amount of data very quickly. This framework was very limiting on audio collection as it restricted the size of immediate transfers [10].

While WatchConnectivity permitted unlimited amounts of data to transfer between the watch and the phone, large files can only transfer as a background process. Apple documents say that this is to preserve battery life, but these documents did not state the possible delays [9].

transferFile() is a method that allows for unlimited file size transfers but may throttle them to improve power and performance of the watch. transferFile() also does not support native reply which would require the watch application to continuously monitor for the file to arrive before triggering a sendMessage() to retrieve a reply [9].

We utilised a sendMessageData() function. SendMessageData() offers immediate sending of data along with a reply. In our experience with sendMessageData() this occurs within one to two seconds, depending on whether Pwy was present in the iPhones memory or not. However, there is a limit to how much data can be sent in sendMessageData(). Apple does not publicly announce this data [7]although , discussions on Stack Overflow point sendMessageData() at 65.5 KB [6]. Sending 10 seconds of audio was below this threshold of16khz audio, which meant that this was suitable for our implementation. However, any increase to audio quality or length of audio recording time will likely require us to take another approach.

3.3.2.2.3 Microphone access For microphone access, we used WatchKit WKAudioRecorderPreset narrow band present at 16kHz. We set a maximum duration of 10 seconds, which would start to listen immediately once the present AudioRecorderController method had been displayed. and then utilised a narrowband to ensure that the audio could transfer through sendData() method for immediate transfers. Files were saved as an m4a, the default output of the Apple Watch microphone before being sent in sendMessageData() function.

3.3.2.3 Altering Pwy for Wizard Of Oz (Pwy WOZ)

To run an evaluation study on the application with participants in order to understand the delays of the application, we altered the application to create a Wizard Of Oz (WOZ) interface. The Apple watch Record screen was replaced and led to minor adjustments to the UI on the Apple Watch by removing and Sending data screen was altered to print "Listening" on the screen. A further WOZ screen was added to the iPhone application to control the WOZ. On the Apple Watch, participants would activate listening by tapping "Tap to listen" as seen in figure 3.8a. Once tapped, the user is transferred to the Listening screen. At this screen the watch would call SendData() every second and await a response from the phone. This interface can be seen in 3.8b 2 possibilities occur: the phone would not reply (not ready to trigger) or would reply with a string which contained the name and notes on the individuals. When a reply was received, the application would navigate to Speaker found as seen in Figure3.8c, which showed the correct information of the person that they were in discussion with. Once Speaker found was called, a vibration and a chime would then occur to notify the user that the watch has accessed the information on the person who is speaking.

[7] https://developer.apple.com/documentation/watchconnectivity

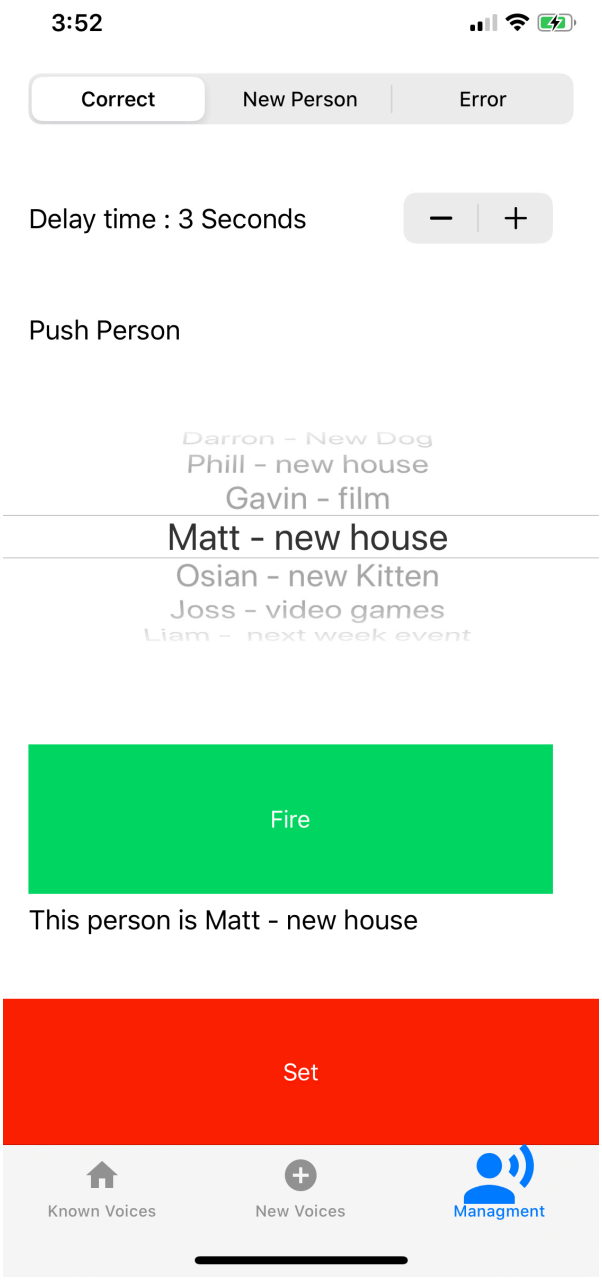


Figure 3.7: The options that were available to researchers during the WOZ study

On the iPhone, an extra screen was placed with options that allowed the WOZ tests to take place, which can be seen in figure 3.7. The following options were available to researchers:

- Type of reply - This option was able to inform the watch whether it was a new person or whether it was a known person. The new person reply was used to give an illusion that the application was active.
- Delay length - This allowed the researcher to state how long the watch could delay for before firing a notification stating who was talking. The minimum delay time available was 0 seconds, and the maximum delay time was 10 seconds.

- Actors Name - This allowed the researcher to select whom the participant was talking to along with notes.
- Fire - This triggered a reply to the watch once a delay had occurred.
- Save - This saved the settings in memory, allowing for the app to close or be reset if required.

(a) Tap to listen button

(b) The watch passively listening. Here the application is looping over SendData() every second until a response from the iPhone is displayed

(c) Speaker has been detected

Figure 3.8: Above are the displays that the participants interacted with on the Apple Watch. Display 3.8a was the screen that participants tapped before entering into conversations. A trigger was utilised to preserve battery life. Display 6.2b was displayed until the watch displayed a speaker in 3.8c. When display 3.8c (Speaker had been detected) was presented, the watch would sound a chime and vibrate.

4 Participatory Design Workshops

In this chapter, we will present a background for the participatory design process we engaged in and discuss the three participatory design workshops that we ran. We ran three different participatory designs workshops with three sets of participants: expert designers, people who find social interaction difficult and people with TBI. In Table 4.1 you will find the basic demographic data of participants from each session. Each workshop was altered to make it more suitable for each set of participants and to encompass feedback from previous sessions as we moved forward. For example, in our study with people with lack of social confidence who were not designers, we supplied them a choice between different design along with more guidance within the session compared to our expert design group. However for our Traumatic Brain Injury (TBI) design group, due to time constraints, we merged the design theatre, and scenario cards along with removing the design critic as this would allow us to generate the most original ideas within the short time frame.

People who find it difficult to infer faces are classed as vulnerable people, resulting in this research project requiring three ethical approvals. For our participatory design workshops, expert designers, people who have difficulty identifying faces required us to submit an "Application For Ethical Approval Of Projects Involving Human Subjects" Form to Swansea University College Of Science Ethics Committee. These applications are in the Appendix under the Section D.

Session	Gender
Expert designers	3 males aged 20 to 35 currently working towards a PhD in the design field
People with difficulty socialising	3 males and 2 female aged 18 to 25 who are currently students and not reciving any formal support
People with TBI	4 males and 3 females who are outpatients with TBI however were attending a TBI support sesssion

Table 4.1: Above states a basic outline demograpical data of participants in each of our participatory design sessions

4.1 Background to Participatory Design

Participatory Design is a group of design and research practices that emphasises on the need of users and designers actively working together in the design process to improve the design of the system to improve the daily life of the end user [57]. Participatory Design emerged in Germany and Scandinavian countries in the 1970's as computers were being introduced into the workplace and concerns arose surrounding the effects that these systems would have on workers. [80]. Participatory design has become a tool in the HCI community for developing technologies with users, allowing for sharing control of the development process of technology, but also sharing expertise which is essential for people with difficulties [147]. Many user groups consist of people with specific usability needs that may not be apparent to the researchers, family members or caseworkers. By employing the end-user in the design process, these end-users essential needs can become apparent and considered within the design process.

Participatory Design Workshops require different approaches depending on the context of the participants, and understand that not all activities are suitable for each set of participants. As Lindsay et al. [93] discussed, designing with people with dementia requires long term relationships with participants with several sessions

with little progress, whereas McCarthy et al. gave participants with diabetes tasks before their Participatory Design Workshops as they were experts in the field [101].

While it is common for Participatory Design Workshops to follow a similar format for all sessions, based off the iterative changes from previous sessions, this is not always the case, and Participatory Design Workshops may be changed depending on the participants. Not all activities within a Participatory Design Workshop may be suitable for all of the participant's groups, and unique insights may not become apparent. For example, a workshop with a group of participants with severe social anxiety may not respond well to taking part in design theatre however may respond well to a design crit. Frauenberger et al. developed tailored elements of their participatory design workshop around the characteristics of the participant, allowing the participant to be comfortable [42].For example, for one participant, Mia found unknown materials and topics overwhelming and preferred routine. Frauenberger et al. generated a set of ads that described her interests and characteristics along with her no-go's allowing them to engage with Mia effectively. By tailoring the experience, we can encompass a more extensive range of stakeholders and users which may not be suited to only one iterative user study.

4.2 Participatory Design Workshop With Expert Design

We ran an expert design session with three participants. The participants were currently working towards their PhD within the Swansea University Future Interaction Lab (FIT lab). The expert design workshop aimed to identifying potential design problems and concerns with the implementation and to provide prompts for future participatory design workshops. This workshop ran for a hour and a half.

This session consisted of the following:

- Design theatre - 20 minutes
- situation cards - 20 minutes
- design critic session - 20 minutes

The design theatre and situation cards were a group activity, whereas the design critic was an individual activity with a discussion between actitivities allow comparison between pariticipants. We compensated participants for their time with refreshments.

4.2.1 Design Theatre

For this participatory design workshop, the design theatre was acted out between researchers. We did not film the design theater as this would allowed us to make changes based off feedback from the designs.

Design theatres allow for the communications of ideas between designers and users of technologies to explain how each party envisions the user flow of the application. It also allows for stimulating discussion between users and designers [115]. In this study, we used design theatre to demonstrate to participants how we envisioned the user flow of Pwy.

In this design theatre, three scenarios were played out. These scenarios were:

- App working as intended
- Meeting someone and adding them to the app
- The system failing

4.2.1.1 App working as intended

The first design theatre was to show the system working as intended on meeting a known person. In this situation, it was a student meeting a lecturer in a cafe. This was to illustrate a possible situation that the design experts could find themselves in. Here the participants were concerned that the user was continually looking at the device and would be perceived as being rude by bystanders. There were also concerns regarding the reaction of the person engaged in conversation when they recognise their name on the device. Some felt that the person engaged in the conversation could be concerned that they were being stalked or spied on by the user. Currently, it is not a social norm to take photos of a person after meeting them for the first time. Participants would also be anxious over the possibility of the user looking on social media for a photo of the individual and then downloading it.

4.2.1.2 Adding a new person

The second design theatre focused on adding an individual to the system and meeting them a second time, similar to design theater one. In this situation, the user met someone in the pub and then again outside. Participants here questioned how peoples voices are saved on the system - should this be done automatically through text to speech algorithms, after the conversation when the individual might unaware that their voice signature is on a phone of an individual that they met once and may never meet again.

Participants raised concerns about consent to save people's voices. Participants felt that not everyone would be happy having their voice signature on a random person phone without realising that this had happened. However, once the individual was aware of why voice was captured, they were more likely to consent to this.

Participants felt that this work could also be extended to have a social network of voices where users did not have to add their voices but instead voices added by individuals. The voice was processed on a server and searched on a central database. Users could also send "requests" to users for their voice - similar to friend requests on Facebook.

4.2.1.3 Demonstrating failure

For the third design theatre, we demonstrated a failure of the system. Here the user met Stuart, and the application called them "Lauren" which was the wrong name and wrong gender. The wrong gender was used to clearly demonstrate that the name was incorrect.

From this design theatre, participants considered that the simplest solution to this situation was to explain to the person why they had their name wrong. For example a participant in a situation like this could play out with *"I have memory problem - people may think it is a bit weird however I have been trying to use this app; this app told me that you were Lauren isn't that weird, followed by "bloody technology" and both have a laugh about it."* While this would result in the user stating that they had a condition, it would also show that they were combating it through the application and that it was the technology that got their name wrong.

Another possible solution to this problem was to use a tree-like structure to question the person, for example asking the user if they had a dog. If yes it is Joe, otherwise its Tom. However, the use of a tree-like structure would require more time on the watch, which bystanders may perceive as the user being rude. However, it was suggested that the tree-like structure could be utilised to display discussion prompts such as "How is the dog?" or an escape route if they get their name wrong, such as "I'm sorry I got your name wrong, I struggle with names, and I was using this app, but it's got your name wrong." Many participants agreed that an escape route would be useful for people who may not be sure what to do if they call someone by the wrong name. They may get overwhelmed and knock their confidence. An escape route may offer a method of explaining why it failed and that they have difficulty recognising people and that they used the watch to aid their communication.

Our expert design team also suggested running an extensive survey of perceptions around the product and how different groups felt about including their voice data on the system. Our expert design panel felt that by gathering this data, we would be able to understand social constructs better and understand how people would use the application and how comfortable is the rest of society to the application.

4.2.2 Scenario Cards

Scenario cards allow for participants to develop situations by giving them information labelled information about a scenario such as "who, " "what," "where" and 'when." From these cards, participants can think about these scenarios and come up with problems and ideas [32].

For the second design activity, participants worked as a group to understand how the application would work in scenarios given by scenario cards. Here participants chose a random situation, person and discussion topic and then analysed how users could use the application. A full list of possible scenarios can be found in table 4.2.

Scenario cards allow for participants to develop situations by giving them information labelled information about a scenario such as "who, " "what," "where" and 'when." From these cards, participants can think about these scenarios and come up with problems and ideas.

Situation	Talking to a:	About
Home	Friend	Last nights football score
Shop	Group	Your new assignment
Pub	New person you think you recognise	A night out
Bar	Partner	Planning for a party
Beach	Family	A new Job offer
Nature Reserve	Boss	An exam
Bus		Your cool new watch
Train		

Table 4.2: Each group was given a random situation, talking to and about card for the scenario cards activity. From this each group was free to consider the finer details of the situations such as the formality of the conversations and whether the situation was busy.

The first situation that participants randomly chose was:

Talking to your boss about planning a party in the bar.

Here the participants identified that it is essential for the user to know their bosses name and that they should have a regular contact with their boss. However they also identified that a good boss should be aware of the condition and as a result be understanding. Party planning would be a difficult situation for the user of the system as they would struggle to identify who is who and as a result, who to issue an invite to the party.

Participants suggested temporary prompts that could be added to the application to give to the user such as clothes and location of seats that was only available for that night. These prompts would automatically delete once the user left a location or after a particular time and new prompts would be generated when they next met the participant. Participants also felt that it would be less intrusive for the user to look at their device as it was going to be busy, and as a result, people would be more forgiving.

However, a potential issue that was spotted was that peoples voices change as they drink alcohol. Voices become more slurred the more intoxicated someone becomes and as a result, this may change their voice

signature and could cause issues for the user if they got names wrong. This situation is especially challenging as people might be less likely to be forgiving when under the influence.

The second situation that participants randomly chose was:

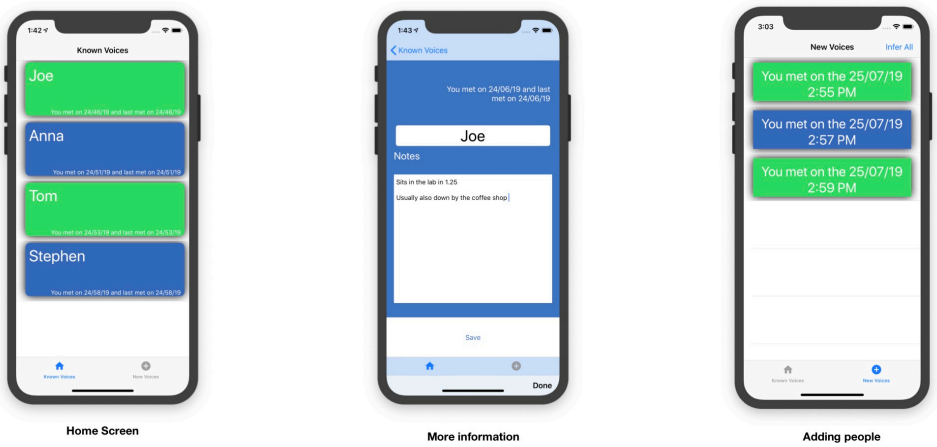
Talking to a family member about your new university assignment in the shop.

Here participants decided that in this context they would use someone whom the user does not regularly interact with such as an aunt. They noted that in this situation, the user had not told their aunt about their assignment but instead someone such as their mother had stated this to their aunt without the knowledge of the user.

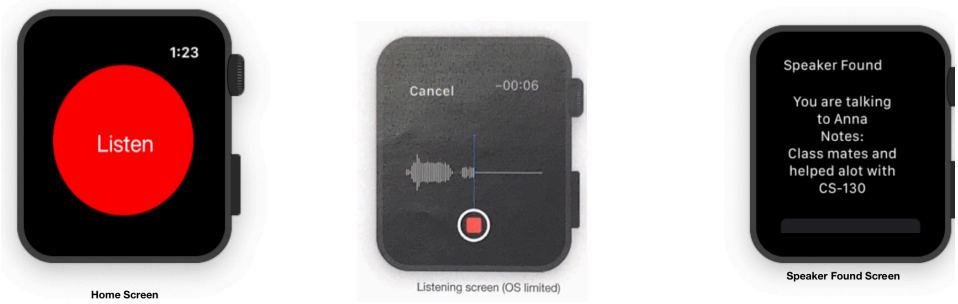
Participants stated the situation might already be awkward if they had walked past the aunt without recognising them and as a result may not have time to start the watch, resulting in them being unable to recognise their aunt. Participants noted that this is a difficult situation for many people and that the application could start listening as soon as the application hears the name of the user name, similar to how "hey Siri" works as discussed in section 5.5.1

Participants noted that an aunt not being aware of the extent of the user's condition could also be offended, and their behaviour inferred as rude and upsetting.

4.2.3 Current Application Design Critic Session



(a) Handout given to the participants containing the phone interface design



(b) Handout given to the participants containing the watch interface design

Figure 4.1: Above shows the handouts that participants recevied during the design ciritc

Design critic is an activity carried out when designers want to improve designs. Participants evaluate ideas that have been developed by the designers and can comment on current designs. These are usually carried out early on in the design process to allow alterations to the systems to occur [17].

In this section, each participant was given a page of A3 paper with three screenshots from the current application. Participants were asked to critique the current design of the application. Once participants had critiqued the design of the phone application, they received another page of A3 with screenshots from the smartwatch application. The participants then had the app flow along with an explanation of each screen presented to them. For the smartphone application, participants received the following screenshots: Known People, Detailed View, and Adding People that can be seen in Figure 4.1.

Once participants finished, we reviewed each participant's design critic and compared them with each other noting key similar themes. Because participants discussed their thoughts behind each of their points we were able to compare themes to the transcript of the session and drew out themes.

4.2.3.1 Customisation

Within the Known People, participants would have liked to have been able to have the colours of each of the cells within the table to display colours of their relation to the participant such as blue for family, red for co workers and pink for friends. Participants also suggested that they could also see when they last met people and felt that it would be nice to reorder the cells in the TableView.

Participants felt that by ordering people by frequency, it would allow users to train themselves to recognise people whom they often see. Users could state that they would want to learn to recognise the top three people they meet, for example, Aled, Bethan and Chris and every time the user meets Aled they would try to learn a unique feature of their face, for example, Aled always wears glasses.

4.2.3.2 Further information.

For a detailed view of known voices, participants felt that it would have helped have seen a picture diary of the individual to aid them to link faces to voices, along with a map of where they have met previously and possibly a dictation of the conversation. They felt that having a name and notes was not sufficient and that individual users might struggle to link names to names to people, similar to linking the faces to names of people with prosopagnosia. However, with a photo diary, participants felt that this could use photos already on the phone could used only for friends and family.

The ability to add photos was a contrast to the design theatre discussions where participants felt that photos might encourage stalking behaviour by the user for photos, and here participants did not feel that images would cause an issue. They felt that this could be utilised from the photos already on the phone and that could be used only for friends and family. Nevertheless, there is a possibility that some of the systems intended users such as people with ASD may not be aware of the social norms, and as a result, may end up searching for people online.

Expert participants suggested developing a social media for sharing voices to allow users to share voices with each other, along with information that the system would generate. Each person would be responsible for placing their voice on an online server along with their name, and a photo and users could then send requests to other people for their voice. They can then also revoke access to their voice. The idea for a social media of voices is similar to how Facebook works with requests and access to information

4.2.3.3 Adding people.

Participants felt that adding people's voices to the application was controversial as it was difficult to differentiate who the speaker was from time and date and it might be difficult to recall who the user was speaking to at that time. Participants felt it might be more suitable to give users the ability to listen to voices to understand who the person is and where they met. Participants did not understand the order that the app displayed new voices and suggested being able to see the time they met as they may have met several people in one day. Participants would have also liked a labeled location, such as "Computational Foundry, Swansea", of where they met that individuals to help prompt the user to work out who they were talking to

4.2.3.4 Watch application.

For the design of the smartwatch, participants received an interaction flow of the following: trigger listening, listening screen; and the results. We stated that the the listening screen design on the watch interface is limited by the operating system that requires a specific view controller to be displayed to make users aware that they are recording. We felt it was important to add this, as it was a vital feature of this user flow within the watch application

Participants felt that a large red circle in screen trigger listing screen saying "listen" was intrusive at a glance and that bystanders may feel that users are recording them to listen back at a later time. Participants felt that this screen should display another colour or something more subtle and not mimic a record button.

When participants viewed the inference screen, they raised concerns that the name of the individual was too small on the screen and that users might find it difficult to view at glance. Participants had concerns that users might also be concerned if there was more than one person in the system with the same name which could lead to confusion on this screen. Some participants suggested that they could view notes quickly in a card style format by scrolling down on the screen or the crown of the watch.

4.2.4 Conclusion of Participatory Design Workshop with Expert Designers

From this study, it became clear that participants found the idea and concept exciting and that there were potential use cases in the real world to support users with social confidence. However, participants highlighted several critical issues along the way with privacy and how failure of the system could be handled, which may influence the experience of using the application. However, participants felt that specific issues could overcome, such as guiding the user through failure, such as a conversation guide if the application failed. Decision trees and discussion prompts could also be utilised to support the recognition of an individual to ensure that it was in discussion with the correct person.

Participants also brought new ideas to the application, such as photo diaries and social media of voices. Participants felt that photo diaries might be useful for some participants depending on their situation. Nonetheless, participants voiced concerned that this could lead to stalking behaviour. Participants further extended the functionality of the application by encompassing features such as social media where every person becomes responsible for their voice, which also removes the need for users to append their data to the application.

However, participants felt that changes were required to the user interface to make it more suitable for participants. Participants felt that UI tweaks were required along with adding more information such as labelled locations to help the user identify whom that person is when adding to the application.

4.2.5 Changes made based on the expert design session

We made several changes based on the feedback, of the expert design session and to orient the studies to be more suitable for further groups.

For the design theatre, we produced 3 videos based on of the same script as the acting within the design theatre. By recording the design theater, it allowed us to compare design theatres between sessions better as we would have removed any errors that may have affected the outcome of the session. Further, it allows us to repeat the video as many times as is required, which is beneficial when working with the TBI design group due to short term memory issues.

Situation cards were unchanged. However, a change to the procedure was made to allow the participants to change cards if they felt that they were unable to come up with a reasonable scenario for this situation.

We also made design modifications to the design critic on the phone and watch. Although the user interface of the phone was unchanged, flow arrows to demonstrate how each screen interacts with each other were added to show the participants the flow between the screens and how they would achieve this. These were not included in the previous stage as we assumed that the participants would be able to infer this, however, this was not the case.

Flow arrows were also added on the watch interface however this was extended to allow 3 options of the possible listen button. For example, option A on trigger was the original button that was shown to the expert design group as a control, while 2 others represented a listen button, one with text and without. We also removed the listing screen as we were unable to change this due to watchOS limitations.

4.3 Participatory Design Workshop With People Who Lack Social Confidence

As the second part of the participatory design workshop, we ran a workshop with participants who lack social confidence. For this study, we focused working participants who self-reported their lack of social confidence. We decided to go for self-report to allow us to encompass participants who may not have received a diagnosis however felt they lacked social confidence. Within this study, participants were aware that they were part of a series of studies, and that this workshop was named "General User Workshop." We chose the phrase General User Workshop to ensure that there was no stigma during the recruitment and that participants were not labelled. We did, however, during recruitment make it clear that it was to help design a system for social interaction.

We recruited 5 participants whom we recruited through social media, mass mailing lists within the university and through personally approaching participants. As compensation for their time, participants had refreshments during the study along with a £5 Amazon gift voucher for their time. This workshop ran for a hour and a half keeping to the same times as the expert desgin study.

4.3.1 Design theatre

For the design theatre, participants viewed the same three scenarios, as shown in the expert design study; however, they were pre-recorded to allow repetition if needed. These videos are available to view via an unlisted link on YouTube. These links are:

- Scenario 1 System working as expected: https://youtu.be/eIjshNYqxDY
- Scenario 2 Adding person: https://youtu.be/Phgpx3Vt5YU
- Scenario 3 Failure: https://youtu.be/FOoXFPm57pA

After viewing the design theatre, participants stated that they thought it would be helpful for the application to continually listen and only alert the participants when the app has detected someone speaking. Participants further wanted the system to automatically add a voice to the app based on pronouns such as when they hear the application say "Hey, I am Tom, what is your name?"

4.3.1.1 Privacy and legality

Participants quickly highlighted concerns that the person that the user is talking to could see their name on the screen and were unsure how they would react to this. Participants felt that they were worried about being perceived as a "stalker."

Participants extended the above concern stating they were unsure what to do if someone asked them to delete themselves from the app. Participants said that they could remove someone but they were also concerned that they would accidentally add them in the future, unaware that they were breaching their trust. There are also legal concerns that participants mentioned surrounding the General Data Protection Regulation (GDPR). Participants were concerned that they would be breaking the law. Although we discussed GDPR in 2.3, participants said that there could be a central blacklist for the voices of people that did not want to be included the application.

4.3.1.2 Edge Cases

Participants did question how the system responded to twins, people with a cold or people who are bilingual. We were unable to state how the app would react to these concerns, as we did not have the results at this stage. However, participants suggested that group activity could be carried out on adding someone to the system on repeating "Hey my name is Joe" for example.

Participants were also uneasy about the reverse engineering of the application and social engineering. They were concerned that someone could extract the neural network from the application and then reverse engineer it to extract the voice of an individual within the system. Participants felt that data could be reverse engineered to produce audio which in turn could be used without the speakers permission. One participant said that the application could be used to benchmark a neural network for faking people's voices. Participants extended the above concern by voicing sophisticated concerns regarding whether a person could also take this neural network and reverse engineer it in a way to trick others who use the application to think they are someone else, such as a friend and get them to cooperate with them and divulge information.

4.3.1.3 Generation of notes

Participants went further by stating that they would like notes on the individual in a way that ensured that they were talking to the right person such as "This is Lauren, she is in your computer science course. She has a dog called Spock." Users can use these notes as probing questions allowing the user to ask questions such as "how is Spock?" or "you going to class later?" This information would not be readily available to someone trying converse with another person without having to require them to compete research beforehand. Generation of notes is also similar to the expert designers with the use of notes and short prompts; however, in this case, the participants wanted the application to generate notes automatically.

4.3.2 Scenario Cards

Following the expert design study, our study with people who lack social confidence also consisted of an the use of scenario cards. Here participants randomly selected a situation, who are talking to and in what

context. For the first situation, participants chose:

You are in a pub talking to a group about a new university assignment.

The participants here assumed that the discussion was about coursework and they had just been handed out a challenging assignment and that they were talking about the lecture unfavourably.

Participants suggested two features that they felt would assist them in this situation. The first feature would be for the system to group the people and detect that they were classmates to give the user prompts about the conversation, such as the lecturer's name and module and course deadlines. Participants wanted notes to be generated automatically based on the conversation suggesting that participants are willing to have their conversations captured and processed.

The second feature that participants suggested was the ability for the watch to calculate where the speakers were and gave a compass-like an interface to show the direction of the speaker to make it clear where they are. Participants felt that this feature would help them when they wanted to direct a message an individual.

However, participants had two concerns about the app in this situation. The first potential problem was how would the app work with background noise. One participant said that she struggled to hear friends in noisy environments and was unsure how the app would then cope and whether the change of error would be significantly higher than in a quiet situation.

Secondly, participants were concerned about the automatic note generation and whether this could then make the app think that the lecturer is unpleasant as a result of this assignment. A series of complaints may make the application state to the user that the lecturer is unpleasant when, in fact, they are pleasant and as a result altering the relationship between the user and lecturer. Any future work that adds automatic notes to the app must consider the effect that the application could have on the user and bystanders and the negative consequences that the user may face.

For the second scenario, participants randomly selected the following scenario:

You are on a train talking to your boss about the football score.

Participants assumed that the relationship between boss and user was friendly and informal.

Participants first suggested that the application could be used to alert the user if they were spending too long on a specific topic such as football. Participants noted that they sometimes end up spending more time than they were planning on a topic than was necessary. They felt that the application could detect when the other person was starting to lose interest in the topic and alert the person to change the topic and possibly suggest a tangent. Alerting the user that they are spending too long on a single topic than necessary can be utilised by users to aid with coaching on changing topics which can improve the users communication tools.

Participants were also concerned that the application could end up learning the voice of the train conductor or another person on the voice as the bosses voice and increase the likelihood of the application getting names incorrect. Participants felt a possible solution for the issue above is for the new person in the conversation to speak into the watch saying their name, resulting in adding an individual and becoming a group activity.

This problem highlighted to the researchers that the application could have hundreds of voices added to the system over a period, such as a year. Train conductors, shop assistants, call centre staff and waiters are individuals that a person may come into contact with frequently; however, no interaction occurs after their initial meet. A self-deleting system may be required for the application to save the system slowing down as a result of the number of comparisons that are required. We discuss this issue further in Chapter 5.

4.3.3 Desgin session

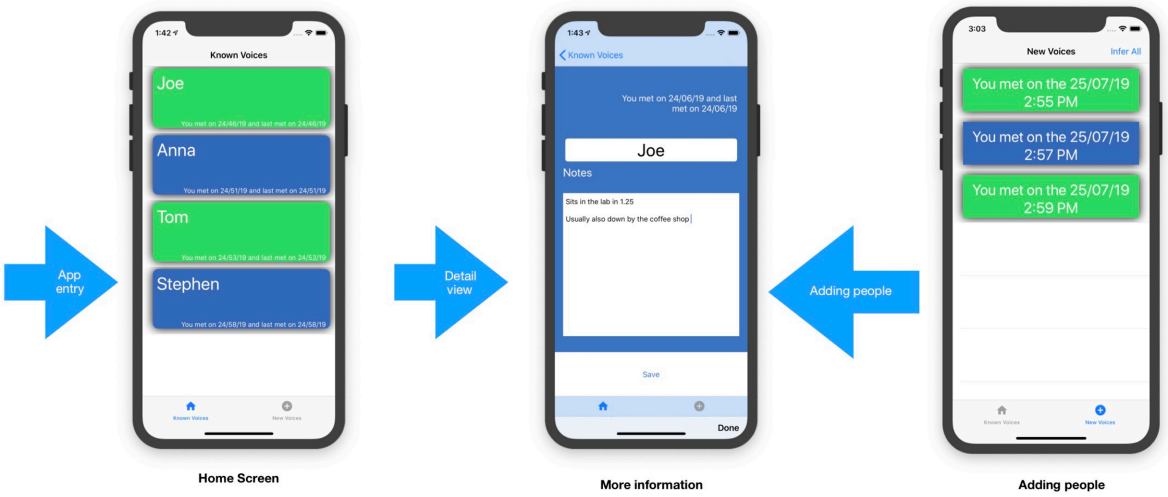


Figure 4.2: Handout given to the participants containing the phone interface design including flow arrows

In this section, participants received an updated design document that contained three screenshots of the mobile phone application and five screenshots of the smartphone app as seen in figure 4.2. The smartwatch app had options for the record button as during the expert design study. During that study, we found that participants did not like having a record button. We still gave participants the record button as an option to act as a control, we also gave a listen button with text to see if participants preferred a button with text.

We analysed the workings of the participants using the same method as described in expert designers, however we encouraged them to engage by providing starting points, such as, asking them what features they wanted and if they liked the colours. The main reason we provided a starting point for the participants was due to the fact that many had not taken part in a design critic before. we did not find any evidence that the prompts influenced the results significantly.

4.3.3.1 Phone Application

4.3.3.1.1 Protection from unauthorised users and accessing data. Participants felt that they wanted a password lock on the app which they could also unlock with their fingerprint if possible. This would allow users to give their phones to someone else knowing that they would not be able to get into the phone.

Participants wanted to listen back on conversations as a method to remember the discussion. They felt that if the smartwatch was saving conversations, it would be useful for them to access the data.Participants felt that it would be beneficial for them to listen back to conversations to remind them of the discussion they had and to possibly reflect from previous conversations. By reflecting on previous conversations, participants can understand how to improve conversations such as what topics to discuss. Reflection of discussions is a skill that people with a lack of social confidence are taught to help them understand what went well and what did not.

4.3.3.1.2 Customisation. Participants suggested being able customise the colours and order of people in the app. They wanted to be able to colour people based on their relationships such as blue for friends, red for family and green for work colleagues. Participants also wanted to change the order of cells on the home screen (table cells), such as the frequency of how often they talk to each other like the expert designers

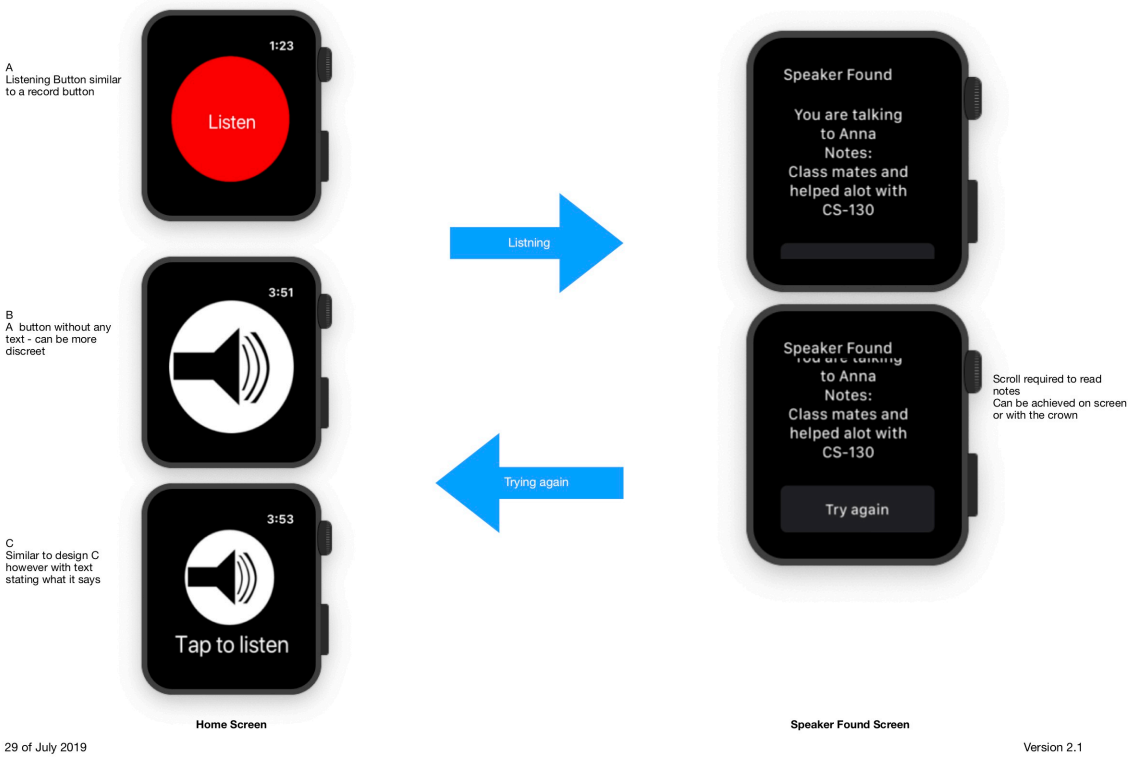


Figure 4.3: Handout given to the participants containing the watch interface design including flow arrows

suggested. Participants stated that they wanted the ability to edit the notes that the application could generate and customise it to what they wanted information on as the notes could be incorrect. They also wanted to change the names of the individual.

4.3.3.1.3 Third-party integration. Participants suggested that they wanted to be able to link to other services such as the contacts app so they can text and call people. Participants wanted the app that does everything with contacts in one application instead of 3 or 4 different apps.

4.3.3.1.4 Accessibility. Participants voiced concerns that the text on the screen was too small for people with visual problems to view and that they found it difficult viewing it on the document. After further review, text was dynamically sized based on the phone setting, resulting in the user having bigger or smaller text that iOS handles this automatically. As the screenshots the screenshots on the researchers phone, here, the text was small as that is how that device was configured.

4.3.3.2 Watch Application

For the watch designs, (seen on figure 4.3) we gave the participants three options for triggering listing along with a scrolled view of the notes screen. We also showed navigation between the the watch applications as

seen in Figure 4.3

(a) The original listen button from the expert design study

(b) The listen button

(c) The listen button with text

Figure 4.4: The designs that participants were given for the design activity

4.3.3.2.1 Trigger listening We found that the large red record button that we used (as can be seen in figure 4.4a) was unpopular as participants felt being red was too attention-grabbing. They also felt that it was too similar to a record button which participants may consider that they were recording with the ability to listen back.

We still included the red record button as a control. However, we also included a speaker icon (figure 4.4b) and one including text (figure 4.4c).

Participants disliked all options. They felt that each option was too intrusive and too apparent to bystanders. Participants felt that strangers might interpret the watch as listening to them. These findings were in complete contrast to the comments participants had stated earlier in the design session. Participants felt that they might be perceived as being creepy with one participant stating that they felt "100% creepy doing this."
Participants said that they found that the icon used in figure 4.4b could be confused as an audio sign, however they found that figure 4.4c was the most suitable as it stated what it did; but, they felt it was too attention grabbing.

4.3.3.2.2 Getting caught using the application Once participants discovered that they would have to trigger the listening themselves and inferred a risk of being caught out by bystanders this lead to a significant change of the application. Participants now felt uncomfortable with the application and disliked all designs along with the idea of the application. Participants did not want bystanders knowing they were using the application, highlighting their condition. As discussed in previous work, users do not want to be considered as unwell [105], and as a result, participants may have thought that bystanders may perceive them as being unwell from using this application.

4.3.3.2.3 Known person screen Participants stated that they wanted the name of the person that they were talking to, to be larger and more pronounced, allowing them to see quickly at a glance whom the individual is. They felt that having notes on the same screen as the name was too much to see at a glance and they would instead prefer to scroll for notes. Participants felt that they were seeing too much information at once and could not process it quick enough.

Participants further stated that they would want the screen not be viewed by bystanders. Participants also supported earlier concerns of that the application could get the speakers name wrong and causing offence to the person they are trying to engage in conversation.

4.3.4 Conclusion from the design workshop with people who lack social confidence

From this workshop, we found that participants were less concerned over privacy compared to the participants from the expert design study. These findings are in contrast to previous research by Ahmed et al. that found that people are concerned with trading privacy for accessibility [2]. Nonetheless, we were surprised in contrast between the two participant groups.

While participants did state that they were concerned with identity theft and social engineering. Participants were also concerned about their legal rights and the legal rights of people stored by the system.

However, participants did also suggest that they wanted more features in the application, such as detecting proximity to the speaker and the automatic note generation. Some of these features are currently not technically possible on our current hardware; but despite this could form our future works.

We were surprised to find how participants reacted once they found that they may have to trigger the listening on the application, completely changing the dynamic of the session. Participants became very quickly aware that bystanders can perceive the user as being rude by disregarding their privacy. Due to time constraints, we were unable to address possible solutions to their concerns. We hypothesise that a trigger button has to be less obvious to make users comfortable or that the user makes it clear that they are recording the bystander on their watch.

4.4 Participatory Design Workshop With People With Tramatic Brain Injury (TBI)

For our final design workshop with potential users, we ran a workshop in Morriston Hospital Traumatic Brain Injury Service in Swansea. In this session, we worked with six participants in a session under the supervision of one of their caseworkers as agreed upon in ethical approval. Each participant had a traumatic brain injury, which affected their ability to interact with others.

Due to constraints in our ethics approval, we could not record participants. To compensate for no audio within the session, two researchers lead the session. We tasked one researcher with talking notes based on themes and exciting concepts, ideas and quotes were recorded on a tablet computer. We did not capture any participants details to ensure the anonymity of participants.

Once the session was completed the notes captured by the researchers were analysed for themes. These themes were written down along with the participants reasonings and were shared to the researchers present. This report was reviewed by the researchers to ensure it was an accurate verdict of what happened in the session.

The Traumatic Brain Injury Service invited us to run a 45-minute workshop as part of their social interaction workshop. Because of the lack of time, we significantly changed the structure of the workshop to combine the design theatre and scenario cards into one activity and removed the design critic. The design theatre was played first to participants followed a set of scenario cards to help foster ideas leading onto a discussion on how participants socialised and how they would want the application to work. We modified these cards to reflect the different demographics of the participants. These cards can be seen in table 4.3.

We first played the design theatre followed by having participants randomly selecting a scenario card. The scenario that the participants generated was:

Meeting your boss on the beach and talking about your hobby.

From this situation, participants came up with the scenario based on one of the activities of the service;

Table 4.3: For the session with participants with TBI, we adjusted the cards to better reflect on the demographics of the session. We also removed Bar as in the workshop with people who lack social confidence "bar" came after "pub" leading to a change of cards.

Situation	Talking to a:	About
Home	Friend	Last nights football score
Shop	Group	Their pet dog
Pub	New person you think you recognise	An event both of you are going to
Restaurant	Partner	Planning for a party
Beach	Family	A new Job offer
Nature Reserve	Boss	An exam
Bus	Care worker	Your cool new watch
Train		A new hobby
Hospital		
This activity		

surfing. Participants came up with the scenario where they were just about to go surfing, meeting their boss on the beach and explaining why they were going surfing.

4.4.1 Safe topics and alert when stress detection

Two of the participants started to act the scenario out quickly with one participant being the surfer and the other being their boss. Their boss asked them when they were going to finish off their work. This lead to the surfer replying "I'm self-employed, and I've already done by 40 hours this week." This comment leads to a discussion with the participants and the caseworker about safe topics. Following this discussion participants said that they would like to have a colour coded screen to alert them to be careful during a conversation.

Participants suggested that the watch could also track heart rate for the detection of stress as heart rate rises with stress[1]. Including heart rate tracking would allow the app to detect that the user was getting nervous and then give them safe topics to the participants so that they would not get into difficulty. Participants also stated that they felt that the detection of nerves could be used to reassure the user that everything was okay.

4.4.2 Failure to recognise voices

Participants stated while they accepted that failure was likely, as they found names difficult already. If the system was more accurate than what they were able to achieve, they felt that the application would be beneficial. Participants had already developed coping techniques when they got a name wrong by explaining that they were not good with names.

Participants said that they did want a confidence score displayed to them. They felt that this would allow them to judge whether to mention a name or keep the conversation going until the system was more confident with recognising with whom they were in discussion. Participants did also state that they would keep the conversation going and to aid the application in identifying who they were conversing with.

Participants felt that having a recovery path would help them to explain when they got confused useful, similar to an SOS button that gave them a script to read that gave them a recovery path. Participants

[1]Participants had already taken part in a session about mindfulness which explained that being stressed caused their heart rate to rise

wanted the script to say that they were sorry and maybe to explain their condition and then carry on with the conversation.

Participants were also concerned about background noises. Participants voiced concerns that the app might pick up someone else who was not involved in the interaction; not the person they want to know. Participants sated that they have heard someone they knew before in the supermarket, but they were on another aisle.

4.4.3 Privacy and legal concerns

Participants were unsure about the legality of recording someone as they were very aware of the privacy laws and data protection but once participants were made aware that this is legal, they felt more comfortable with the system. Participants, however, stated that they did not think someone would sue them for using speaker-recognition.

When we probed participants on how they would respond when someone asked them what the watch was doing and why they were using it, one participant quickly replied stating "This is my communication buddy" with another participant stating "I have a brain injury and this is part of my brain." Participants were quick to explain that they would say that the watch was a communication aid and that they had a TBI, and as a result, they found names difficult.

When further probed regarding people enquiring about how the technology worked, they stated they would like to give them a business card or give them a quick explanation. Participants stated that while they felt that people found the technology interesting they would become frustrated having to continually how it worked or explain their condition. One participant stated that "I have had a brain injury, this is my communication aid, I'd rather not talk about it, let's talk about puppies." From this discussion, it became clear that people with TBI do want to continue with their lives and while are happy to receive support, they do not want TBI or their accessibility tools to become the focus of attention.

4.4.4 Third-party integration

Participants felt that if the application was listening to their discussions, it could also analyse the discussions for events and to allow the user to create events and manage conflicts through speech. While in this workshop, the caseworker explained that they advice patients to use calendars on their phones to remove any mental work and remove reliance on other people. Participants would like the application to speak to many apps, especially calendar and give them feedback during discussions if they were free or not, and make the event if they can. Example participants stated that if there were a party on Wednesday, they would like the watch to vibrate if they were free or not and if they were, to press a button that added it to their calendar.

Participants were not interested in photo diaries as such, as they felt that it was removing people's privacy. Participants felt that it could add context to the discussion, such as showing people photos of the football with their football partner, however, participants only had a couple of seconds to look at the watch without the danger of being perceived as rude.

4.4.5 Customisation

Participants wanted the ability to customise the application to state how much information they would see along with what features were enabled. Participants were keen to stress that each of their needs was different and as a result that they felt that a one size fits all solution would not be suitable for people with TBI. While some participants felt that the use of photo diaries would aid conversations, others felt that it might simply confuse the user with a different context or consume too much time. Participants were also concerned about stalking. Participants stated that they want to be able to customise the watch interface as for some of them, they only need the see a name while other participants needed more information to infer whom they are in discussion with.

Figure 4.5: Example of using a wireless bluetooth earphone would look like. In this example AirPods are used

4.4.6 Use of other modalities

The researcher probed the participants on other modalities, such as sound that they would like to use. Participants were asked whether they would prefer feedback in an audio modality by using a Bluetooth earphone which listens and then alerted the user auditory if it detected something. Participants favoured this idea as they felt it was more discrete than a watch and removed the need to look down at their watch. Participants were interested firstly in an earphone that made it clear that they were being recording and found it was exciting. Smith placed an AirPod in his ear (similar to figure 4.5) where the AirPod was discreet. Participants prefered this due to the discreteness of the AirPod and Participants felt that a small headphone similar to the AirPod was fashionable and did not look out of place.

The discussion on Bluetooth headphones led to a conversation on whether participants felt that they were giving the impression of being rude wearing headphones. However, participants stated that many people today wear earphones while engaging in a conversation. They said that that they know whether a person is listening or not. The caseworker also agreed with this citing her son also leaves his headphones in when he interacts with others.

However, one participant stated that they might find the audio input distracting and that it may just give them information that they already knew or potentially cut them off before saying something leading to them to confusion. These concerns were further inspired by the facilitator when his Airpod started playing music halfway through a sentence due to an autoplay feature. Participants felt that auditory was while suitable for them it would not be for all people with TBI.

4.4.7 Conclusion of design session with people with TBI

In this workshop, we found that participants were less concerned about failure and privacy concerns and instead were more focused on appending features that would help them. Participants saw the app more positively compared to the other design groups that we have worked with in this chapter.

Participants felt that if the watch could detect nerves and guide them to safer conversations it would be a massive asset to them. Participants felt too that an SOS button that could be triggered if the name was wrong; it would help them explain why it had failed. However, many participants already produced coping strategies when they got individuals names incorrect, so this was less of a concern to them.

Participants saw the watch as an extension of their brain and wanted it to track events on people without the direct input from the user. Participants wanted to confirm events but however, they did not want to input them manually. Participants reacted similarly to photo diaries and notes. However, customisation was vital, with participants stating that some user would like to see more information than others, stating that their conditions were unique.

Participants were interested in making the system more discreet and using other modalities. As participants saw the device as a communications aid, and while they were happy to explain why they used the watch, they did not want this to become the centre of attention. Participants had already had to explain their injuries routinely and wanted to continue their lives without interference wherever possible. Nonetheless, participants felt that other modalities such as using Bluetooth headphones were another method of allowing them to be alerted of who was speaking.

We found it interesting how the participants accepted their condition and the use of the watch as an aid. Participants wanted the application to be completely discrete however participants had no concerns if they were caught using the watch, and would happily explain it as a communication aid, unlike the study with people who lacked social confidence. While the group with people who cited they lacked social confidence were self-reported and the TBI group professionally diagnosed, both groups could find the application useful. However, as the participants of the TBI group already attending support sessions, these participants are more willing to accept professional support.

4.5 Conclusions of Participatory Design Workshop

In the series of participatory design workshops, participants consistently raised several themes such as failure and privacy; however, different groups had different views on each of the problems. In this section, we conclude all three of the participatory design workshops we ran.

4.5.1 Failure

All three groups accepted that failure was inevitable, though, participants with TBI were more accepting of failure, citing that any assistance is better than no assistance. While the expert designers and participants who lacked social confidence wanted a tree-like structure to ensure that they were talking to the correct person, participants with TBI were happy with a confidence score of the system to allow them to make their judgments.

People who lack social confidence are likely to have the ability still, however, may find it challenging to have confidence that they have the correct name, resulting in this system, causing them more anxiety. Participants may have a name in mind, but they may be unsure whether it is the correct name. However, for people with TBI, many of them have impaired cognitive ability, which none of our participants with difficulty socialising declared. Participants with TBI may not be able even to infer a face.

4.5.2 Privacy

Privacy was a concern that all three sets of participants highlighted with the main concern being bystanders privacy. No group was concerned about their own self-privacy. The lack of self-privacy concern was likely a result of being aware of the recording taking place knowing that they could stop if they wanted privacy, such as during a sensitive topic. Furthermore, each set of participants had a different approach to discreetness and how they felt that they would accept the application.

Concerns for bystander privacy were present in all sets of groups, and each group had a different approach to this privacy problem. The expert designers felt that the best approach was a social media of voices, where users could request access to peoples voices. Participants with TBI focused more on explaining that they were recording and explained the reasoning behind it; however, they felt that most people would be accommodating and that participants would rather talk about other topics.

Participants with TBI were also interested in using other modalities such as audio feedback through Bluetooth devices such as AirPods. Participants with TBI were more willing for participants to know that they were using the watch as a support for their brain injury. However, our expert designers group and participants with social difficulty felt that the interface was more discreet and should not show to bystanders as such to stop bystanders believe that the user is stalking them.

Participants with social difficulty showed minimal regard for bystander privacy. Participants wanted all audio recorded and processed without the knowledge of bystanders and to be as discreet as possible. Nonetheless, when it became apparent that bystanders could catch the participants using the application, they quickly rejected it, finding it was creepy. We hypothesise this is a result of participants did not want to be seen using an accessibility device.

4.5.3 New Features

Each group presented a feature that they wanted added to the application. These being social media, continues listening to notes and third-party integration. A social media platform that the expert designers proposed would allow bystanders to control their data better than the user, though, for a system to be successful, it would require a large amount of the population to sign up. It is unlikely as the general public may be unwilling to share data to services, especially since the Facebook and Cambridge Analytica scandal [23].

The ability to listen to conversations to generate notes that the participants with social anxiety proposed was a novel concept which can support people with short term memory issues or unsure about the person that they are in discussion. Participants with TBI further extended this by wanting notes to create calendar events then or work with other third-party apps and support people during a conversation.

These features demonstrate that participants want the application to support them in real-life conversations and to become a conversation aid to support them. They feel that while capturing this data, this data should also be used to support the users during conversations further.

4.5.4 Design changes based from participatory design workshops

This chapter highlighted that a one size fits all approach is not suitable. However, due to time and resource constraints of this MRes, we decided to continue with a single application. This single application would attempt to encompass the results from the group who find socialising difficult and the TBI group.

For the single application, privacy and reduction of failure must be considered in the application. Participants perceived that privacy was critical to the application, and participants wanted to be seen as respecting privacy when recording.

Participants also wanted the application to have high accuracy. For high accuracy, it would require using a high accuracy algorithm which can be challenging. While this is something that we have encompassed into our specifications, in chapter 5 we will further discuss more the challenges of a high accuracy algorithm.

4.6 Discussion

In this section, we have learnt that participants needs vary depending on their condition. Participants with TBI were more accepting of using the device compared to those with difficulty socialising. We can hypothesise this is a result of people with TBI are more willing to receive help because they have already had

assistance with rehabilitation, while people with social anxiety have not received any support.

However, these workshops have identified that each group have a unique set of requirements which result in many conflicts. For example, a tree-like question system was a feature that the exert design and participants with difficulty socialising would have liked. However, participants with TBI wanted to look at the watch as little as possible and did not want the distraction of the watch when not required.

While all groups supported passive listening, there were major conflicts over privacy. The group concerns about discreetness were substantially different, with the group with people with difficulty socialising wanting the most discreetness while the expert design group was wanting a social network to share voices. Participants with difficulty socialising wanted a conversation aid which guided them during conversations. Participants with TBI wanted a memory aid that helped them through calendar reminders, reminding of safe topics and giving guidance on how to talk to a bystander.

These results show that to build a system to support people infer faces, using a one fit approach would not be possible due to the differences between each of the group's requirements. Each of the group's requirements is different from each other, which results in any technical solution for one group may not be suitable for another group to use. For example, for social media, as the expert designers wanted would require identification to occur on a server which would not be discreet as people with social difficulty wanted.

These differences present new challenges to producing an application to support people with inferring faces because each set of users would require an application that is specifically designed and produced around their needs. While this opens up further work, evaluating between different user groups becomes significantly challenging as it also requires evaluation of three designs instead of a single design. The technical evaluation also becomes difficult between all three designs as each of them would require different altered algorithms which may produce different results.

5 Speaker-Recongtion In Machine Learning

This chapter outlines our approach to recognising individual speakers using an open-source machine learning model trained on a dataset gathered from Librivox. We will also discuss how we modified the machine learning approach to improve efficiency to place onto a mobile phone application. We will develop a taxonomy of machine learning along with an overview of other speaker-recognition systems. We will then review machine learning algorithms that will allow us to develop a real-time speaker recognition algorithm to run on mobile phones using some offline training.

We utilised Machine learning due to the complexity of the voice data. Each second of audio data contains 16,000 samples. To calculate basic parameters of audio such as timber requires analysis of surrounding samples to produce a wavelength, resulting in it being too complicated for humans to calculate. Machine learning can also detect patterns and data invoices that a programmer may not have recognised as an identifying trait.

However, producing a machine learning algorithm is not a trivial undertaking as it requires a theoretical understanding of the algorithms, along with the production being computationally expensive. As a result, we reviewed several approaches, some prebuilt and trained while others required training from scratch.

5.1 Taxonomy of Speaker-Recongtion

In this section we will discuss the taxonomy of speaker-recognition. Speaker-recognition is part of the area of voice recognition that covers speaker-recognition, that is used to answer who is speaking, Speaker verification (otherwise known as speaker authentication) is the identification of a person along with another form of ID and speaker diarisation that recognises when the same speaker is speaking [85].

Speaker-recognition and Speech recognition. Speaker-recognition is the ability to identify a person from speech [**?**] and is used to answer the question "Who is speaking?" This is different to speech recognition that focuses on turning what a person has said into text or "what is said". Speaker-recognition is broken up into several subcategories, each with their features which separates them from other categories. While specific applications do utilise both speech recognition and speaker-recognition, such as voice assistants on smartphones for example Siri, Google Now and Amazon Alexa [134, 135, 52, 4], these require two separate algorithms in order to function.

Text-Dependent and Text-Independent Speaker Recognition. We can categorise speaker-recognition into text-dependent speaker recognition or text-independent speaker recognition. Text-dependent requires that the same phrase used in training a model for voice recognition must be the same phrase used within inference [59]. Text-Dependant speaker-recognition is currently primarily used within the financial sector, with organisations such as the UK Government HMRC using speaker-recognition with the phrase "My Voice is my Password." [61]

Meanwhile, text-independent recognition allows for training and inferring phrases to be different with no effect on the accuracy of the system [48], allowing the use of this model in real-world conversations. Text-dependent speaker-recognition is good at inferring the identification of an individual over an phone call and is utilised by phone banking, while text-independent is has the advantage of recognition of the individual in

person in a natural conversation.

Closed and Open Sets. We can further divide speaker-recognition into close-set and open sets. Close sets focus on identifying a speaker from a set of voices while an open set states whether the voice belongs to the set or not [48]. Close sets may also be known as N-ary classification, and open sets may be known as binary classification [21]. Open sets algorithms are sometimes utilised to infer whether a personal assistant is talking to its owner, such as the use of Siri [134] whereas close sets may utilise open sets to allow a return of unknown speaker and to improve the computational efficiency of the algorithm.

Binary and N-ary Recognition. Speaker-recognition can be classed as binary recognition and N-ary recognition [21]. Binary recognition is only able to recognise whether one person is speaking or not (returning true or false) and is often used by virtual assistances to identify whether they are talking to the device owner or not [134]. However, N-ary recognition is able to recongise multiple speakers but requires more computational power compared to binary recognition.

Local and Remote Inference. We can divide speaker-recognition further into local and remote inference and training. Due to the computational power that is required to run networks for inference, and the size of the dataset that is needed for training requirement of any new data, it may be unsuitable to have the training or inferring on the device. Here the voice sample would be sent to another device such as a remote server for inference with the reply sent back to the device stating the speaker's name.

On-device, on the other hand, does all the training and inference on the device, which means that it can work independently. However, these models may not perform as well as off-device training and may not be able to retrain due to the computational complexity that is required for retraining.

There may also be instances where training will occur off device such as on a remote computer, but inference occurs on the device. For example, feature extraction may occur on an external computer, but the actual inference from those features may occur on the smartphone. Feature extraction is a computationally expensive algorithm; however, the inference is less computationally expensive but trained around voices specific to that device.

5.2 Current Speaker-Recongition Algorithms

There are two types of speaker-recognition algorithms which we considered implementing for the application, open access or closed access. Each of these algorithms has its merits and detriments; however, we were unable to access close algorithms due to their closed nature.

5.2.1 Closed Speaker-Recognition Algorithms

We reviewed two closed source algorithms, Apple's Hey Siri and Google Voice Match, which are currently available in commercial products. While they are available in commercial products, they did not provide us with a way of accessing the algorithm and allow us to utilise them. However, the two algorithms that we reviewed were developed for mobile phones with a minimal detrimental effect on the performance of the device.

5.2.1.1 Apple "Hey Siri"

Apple's Hey Siri was developed to allow users to invoke Siri on a range of Apple products hands-free. Hey Siri utilised text-dependent binary recognition to identify whether it was the device owner attempted to invoke Siri or not, as a form of verification. While Apple utilises Hey Siri to identify the speaker, sensitive requests (such as texting someone) still require the device to be unlocked; this eliminates someone else triggering Siri other than the device owner, such as someone saying "Hey Siri" on the television.

Apple has self-published papers on the pipeline of Hey Siri; Hey Siri utilises a speech recognition algorithm to detect if the phrase "Hey Siri" has been said before triggering a deep neural network to produce a confidence score of whether it was the trained speaker or not. If the score is above a certain threshold, Siri is triggered; otherwise, Hey Siri will ignore the voice [134]. While Apple has published a paper on Hey Siri, during our research, we could not find any documentation to allow us to utilise Hey Siri.

5.2.1.2 Google Voice Match

Google Voice Match is an alternative to Hey Siri that is utilised by Google Assistant withing Google Home. Google Assist main difference from Hey Siri is that it utilises N-ary speaker-recognition, allowing up to 6 unique speakers to be recognised. Google Voice Match is trained in Googles servers when a new voice is appended or updated to Google Voice Match, while Apple trains Hey Siri on the device [52, 134]. Like Hey Siri, we could not find any API's or documentation that allowed us to utilise Google Voice Match

5.2.2 Open speaker-recognition Algorithms

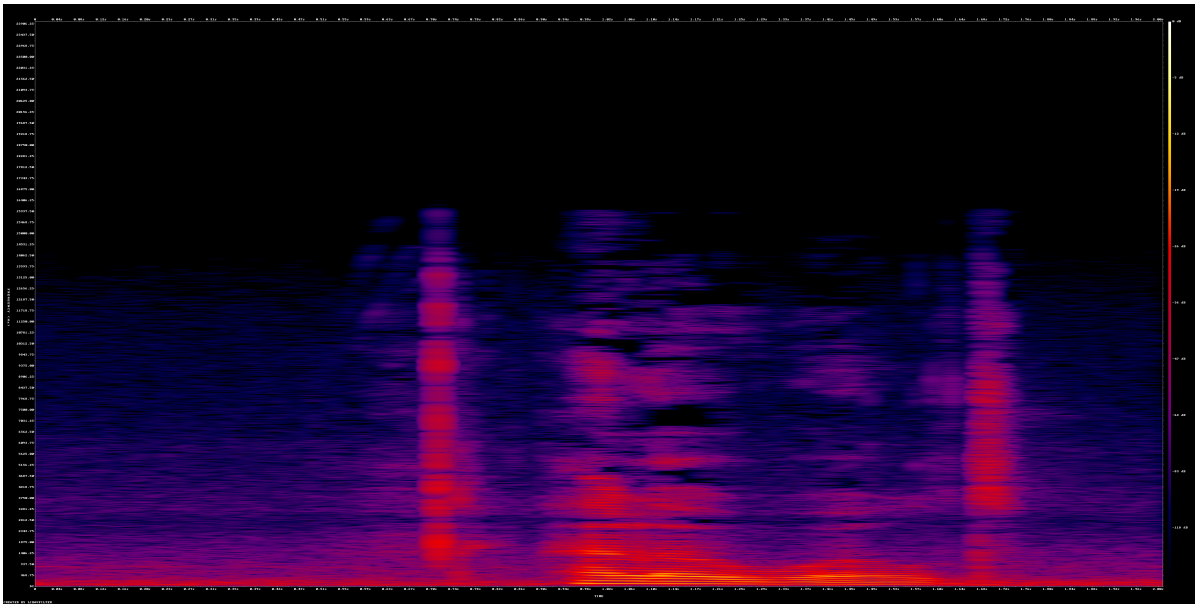


Figure 5.1: An example of a spectrogram that which is generated from an m4a file saying the phrase "Hello World" generated by FFMPEG

We considered three open speaker-recognition algorithms that each had their merits and detriments and we selected work by Christopher Gill as a foundation for our approach. Two of the three algorithms were open-source, with Microsoft solution being the only closed source approach.

5.2.2.1 Microsoft Azure Cognitive Services

We considered that the Microsoft Azure Cognitive Services (MS Speech) [104] system was a promising option; however, it is still in preview and not technically available in Europe with some technical limitations. MS Speech is an off-device speaker-recognition text-independent speaker-recognition system that is part of the Microsoft Azure, Microsoft's cloud platform. Microsoft allowed up to 1,000 voices on MS Speech and

returned a confidence level [104]. However, MS Speech is also still in preview and was not a finished product meaning access could change during development. Further MS Speech is a paid platform which could become costly if we were to go above the request thresholds. MS Speech is also not technically available in Europe, which we believe may be a result of GDPR.

5.2.2.2 Alizé

We also reviewed the Alizé project [71] and, while promising, we found it had poor documentation and was unreliable during testing. The Alizé project is an open-source platform for speaker-recognition that contains tools for speaker verification and speaker diarization along with an array of tools for manipulating audio formats. Alizé also provided an Android demo application which we tested on a virtual phone in Android Studio. While Alizé seemed encouraging, we quickly ran into difficulty with the demo application. The demo application would frequently loose voices and hang for no reason. Attempts to fix the application were frustrated bu Alizé logs which were cryptic, confusing to read and difficult to understand. Further, the lack of documentation in the Alizé code base and the demo application made it very difficult to use. Although we attempted to work with Alizé; we abandoned it once it became apparent of its shortcomings and lack of documentation [71].

5.2.2.3 Christopher Gill Approach

We reviewed a technical article by Christopher Gill in Towards Data Science [47], which gave a discussion on Gill's implementation along with the source code. We found this promising and selected this approach and modified it to increase performance and to optimise it for a mobile device. Gills approach was to use Spectrograms (example in figure 5.1), a visual representation of sound and feeding the voices through a Convolutional Neural Network (CNN) which trained on the CIFAR-10 architecture; a CNN designed explicitly for processing images. Computer vision is a subsection of machine learning with a broad set of tools that can be deployed to recognise objects and differences in images. Gill then removed the last layer of the CNN (which is used by the CNN as a classification layer) and fed the results into a Support Vector Machine (SVM), a supervised classifier. SVM performance is better at the large volumes of data that the CNN produced, which was 400,000 dimensions deep [47]. By using the SVM, we could utilise transfer learning (otherwise known as one-shot learning), which resulted in us not required in storing the training data, which is beneficial due to its large size. Gills pipeline is presented in figure 5.2

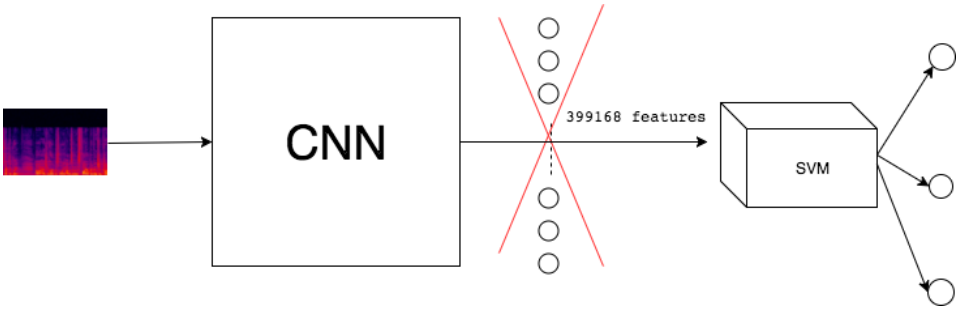


Figure 5.2: Above shows the implementation used by Gill [47]. Here the voice is already produced into a spectrogram which parsed into an SVM that can detect three voices and returns a result 0-2 relating to the speaker. An array was returned, highlighting who was speaking. Image source [47]

Gills code was also publicly available and well documented along with a discussion on Towards Data Science[1], a blogging website dedicated to articles on data science and machine learning, on Medium, a publishing plat-

[1] towardsdatascience.com

form[2], unlike Alizé project where documentation was difficult to analyse.

While Gills [47] work seemed promising, the performance was between 40% and 95% accuracy and was susceptible to the speakers gender and the size of the group that it had to recognise based off its training data. We believed that we could increase the performance of this approach using different algorithms within the pipeline and using a larger data set.

5.3 Gathering training and testing data

To allow us to train our machine learning algorithm, we had to collect training data. Training data is data that our algorithm can review and learn differences in to allow it to distinguish between further new samples. We used testing data to be able to evaluate our algorithm. We can derive testing data from the training data; however, testing data cannot be used to train the algorithm as we want to evaluate how the algorithm works on new data which it has not seen [21]. However, we also cannot train our algorithm on all the data the algorithm will interpret in its lifetime as this data does not exist and may be too large for us to train it.

To try to ensure we had a well performing algorithm, we had to select high-quality data for training. To ensure that we had high-quality training data, we produced the following requirements for each of our audio training set:

- The audio must be guaranteed only to contain a single speaker. Otherwise, the machine learning algorithm will not be able to infer the difference between speakers.
- The dataset should contain labels of who the speaker is if not all samples are unique. Otherwise, the machine learning algorithm may get taught that two samples by the same person are a unique person.
- The audio must be of high quality with minimal background noise. By having high-quality audio, we remove any chance of the machine learning algorithm trained on non-voice data.
- The utterance, the act of speaking, must be longer than 30 seconds. By having 30 seconds of utterance, it will use the machine learning algorithm to have enough data to learn differences in voices.
- The data should be publicly available and not covered under copyright. Although we do not plan to publish our dataset, downloading copyright data could cause our ISP (JANET) to flag our connection and possibly block our internet connection.

We reviewed four potential data sources and selected Librivox, a public domain audiobook repository, as it best fits our criteria. However, below, we will discuss the four data sources we reviewed.

5.3.1 Producing Data Ourselves

We initially investigated gathering the data ourselves; however, we quickly learned that this was unsuitable due to the size of the dataset needed. Further studies that do not target a specific demographic within the university commonly collects data on a single demographic - young university undergraduates are the largest demograpic we have access to. Targeting a specific demographic requires ethical approval by the College of Science. Previous experience within the research group has found that recruiting participants is difficult, and for the large dataset required, it would be impossible.

Nonetheless, during the investigation, it was decided that participants would read from a script contain the Harvard sentences, a standardised list of sentences developed by the IEEE [70]. These sentences were:

- *The birch canoe slid on the smooth planks.*
- *Glue the sheet to the dark blue background.*

[2] medium.com

- *It's easy to tell the depth of a well.*
- *These days a chicken leg is a rare dish.*
- *Rice is often served in round bowls.*
- *The juice of lemons makes fine punch.*
- *The box was thrown beside the parked truck.*
- *The hogs were fed chopped corn and garbage.*
- *Four hours of steady work faced us.*
- *Large size in stockings is hard to sell.*

5.3.2 Youtube

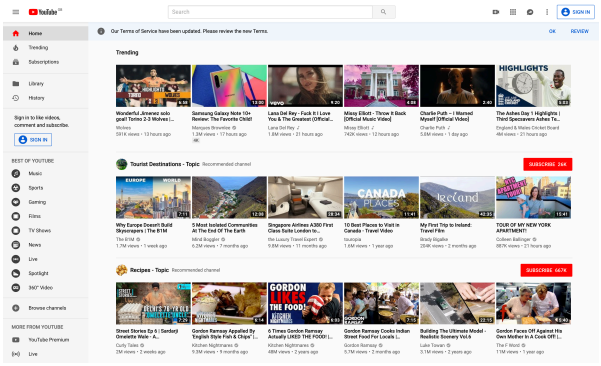


Figure 5.3: An screenshot of youtube.com

We considered using Youtube, an online repository of videos, as a source of retrieving audio for training data. Youtube has 300 hours of video uploaded every minute and although many of this is low-quality [102], we could utilise this for training data. However, we quickly realised that it was unsuitable for numerous reasons such as background music, licensing and guest actors. We also found it scrap audio from a single actor difficult as it required creating a playlist manually.

We did attempt to download audio from Youtube using Youtube-DL[3] but we found that it would require manual editing to remove introduction music, alternative actors and this was very time-consuming. Further Youtube actors do not use such a high-quality microphone resulting in poor audio quality, as many Youtube actors prioritise visual over audio. While subsections of actors such as ASMR actors[4], these actors may talk differently compared to how they talk when not on video. Video as a medium is significantly larger than audio files due to the need to carry visual data such as frames, resulting in the download size being larger, reducing the amount of data we could store. Further, copyright laws cover many Youtube videos.

5.3.3 Mozilla Common Voice

We also considered using the Common Voice project, a project by Mozilla that is designed to collect voices to teach computers to speak [107], as training and testing data. However we realised that found we could not guarantee how many samples within the dataset originated from the same person. The Common Voice dataset consists of an audio file, primary demographic data of the speaker along with votes stating the quality of the entity. The general public creates common voice entries and votes on the quality of the samples [106].

[3] https://ytdl-org.github.io/youtube-dl/index.html
[4] ASMR are a category of videos where an actor creates a physiological event through sound

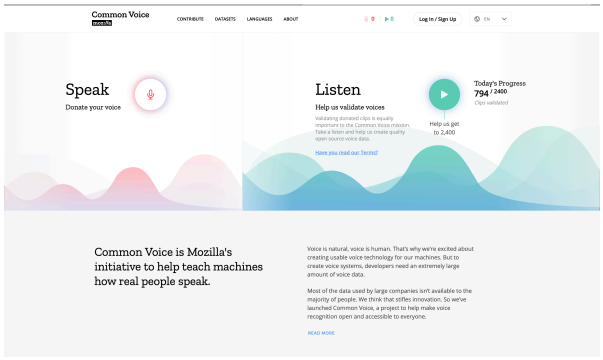


Figure 5.4: An screenshot of voice.mozilla.org, homepage of the Common Voice project

While reviewing the data we found numerous issues within the dataset. The same group of people creates many of the data entries; however, there no label that states whom the speaker was, apart from the basic demographic data, resulting in the fact that we were unable to guaranty the number of unique speakers within the dataset. There were also concerns about the utterances, the length of time someone speaks, being too short. Many of the phrases used were less than 5 seconds worth of text, making it difficult for us to train our machine learning algorithm to be accurate on identify of the speaker. For these reasons, it made the dataset unsuitable for the training of our neural network.

5.3.4 Librivox

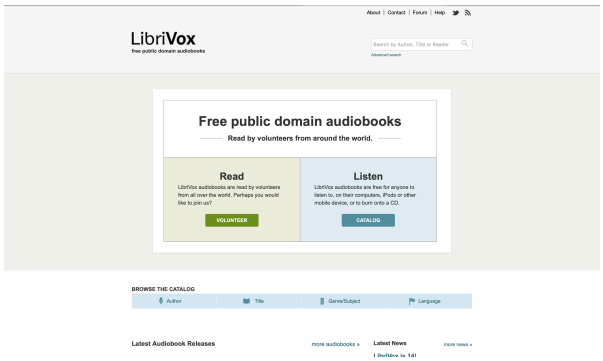


Figure 5.5: An screenshot of libirvox.org

We selected Librivox, a public domain repository of audiobooks [90], as a source of training and testing data as Librivox contained labelled data for a single speaker and could combine the same speaker for multiple books together. We could also guarantee that the audio was of high quality as Librivox encourages the use of dedicated recording equipment. Librivox recordings were released in the public domain, allowing us to download it legally.

Anyone can upload audiobooks to Librivox that were published before 1923 as books would have entered the public domain. Librivox also does not restrict downloading audiobooks; however, we did run into difficulty trying to scrape Librivox. Librivox supports solo work, and every book has labelled data of who spoke in the audiobook [91], allowing us to guaranty uniqueness within speakers. Further, all audiobooks had a large amount of utterance. Audio files were usually of high quality and scripts to scrape Librivox were available by Gill [47].

However we found that Librivox would return us an error code 403, which is a standard HTTP forbid-

den error, on downloading the file 100. Librivox would not return error code 403 on any other samples (even in multiples of 100). Error codes can often be thrown when websites detect when scraping occurs. While scraping websites can be illegal, Librivox is in the public domain which permits scraping. We cannot guarantee that Librivox was throwing the error as we scrapped Librivox through the University Ethernet. Swansea University. is an ISP through the JANET platform, a high-speed internet ISP designed for education and research [5]. Swansea University or JANET could have been throwing the 403 errors as a method to restrict scraping. To combat this, we modified Gill's scraper to encompass a 1-second pause between each request along with adding try-catches at each statement, ensuring we would not lose volatile data if our script crashed resulting in the deletion of hundreds of audiobooks.

We were able to download 303 unique voices from Librivox from several hundred audiobooks and combine each audiobook using FFMPEG[6] and SOX [7] into a single WAV file. We then removed any silences longer than 1 second in length and re-sampled all audio to be at a standard 16khz. We then exported each file was exported to a WAV file to remove any compression artefacts that may exist in mp3. The largest sample we generated was 471MBs in size being 4 hours and 17 minutes long, with the smallest sample being 99.5mb and only 52 minutes in length. We also downloaded multiple languages as our algorithm was not affected in the text but voices. In comparison,speech-to-text algorithms required labelled text to each sample; however, is are agnostic to languages.

5.4 Our Pipeline

Our speaker-recognition algorithm consisted of feature extraction to increase accuracy and then a secondary lighter weight algorithm to identify the speaker. In this section, we will discuss how we utilised each of the algorithms. Figure 5.6 below depicts the data flow within the application.

We pass the feature extraction algorithm the raw audio data, which returns a series of floats in an array in an array ([[Floats]] or a 3D Array). We then pass the array of floats into the identification algorithm, which will return a single integer (int). This integer is then used to lookup within the database to retrieve the speaker object to return to the watch.

Retraining is required for identification to allow the nodes to understand the new voice. Retraining is extremely computationally expensive, resulting in CoreML and TensorFlow Lite, the two most significant machine learning frameworks for machine learning on mobile devices, not supporting the retraining of algorithms. We can use simple machine learning algorithms that do not utilise the machine above learning frameworks. However, performance rapidly declines on more complex data sets. Each second of audio at 16khz is the equivalent of 16,000 samples.

To combat this we divided our algorithm in two. The first part is feature extraction, which is extremely computationally difficult to train, which in our case required a computer dedicated to Machine Learning. However, once produced and pruned, it does not require retraining and is efficient to run. The second algorithm is a much smaller algorithm that is not required to utilise the machine learning frameworks and trained on the results from the feature extraction. Once the user wants to add a new voice to the system, this smaller neural network does require retraining. Retraining is significantly computationally cheaper than retraining our feature extractor, and training can commence on the phones own processor.

[5] JANET homepage: https://www.jisc.ac.uk/janet
[6] FFMPEG hompage: https://ffmpeg.org
[7] SOX homepage: http://sox.sourceforge.net

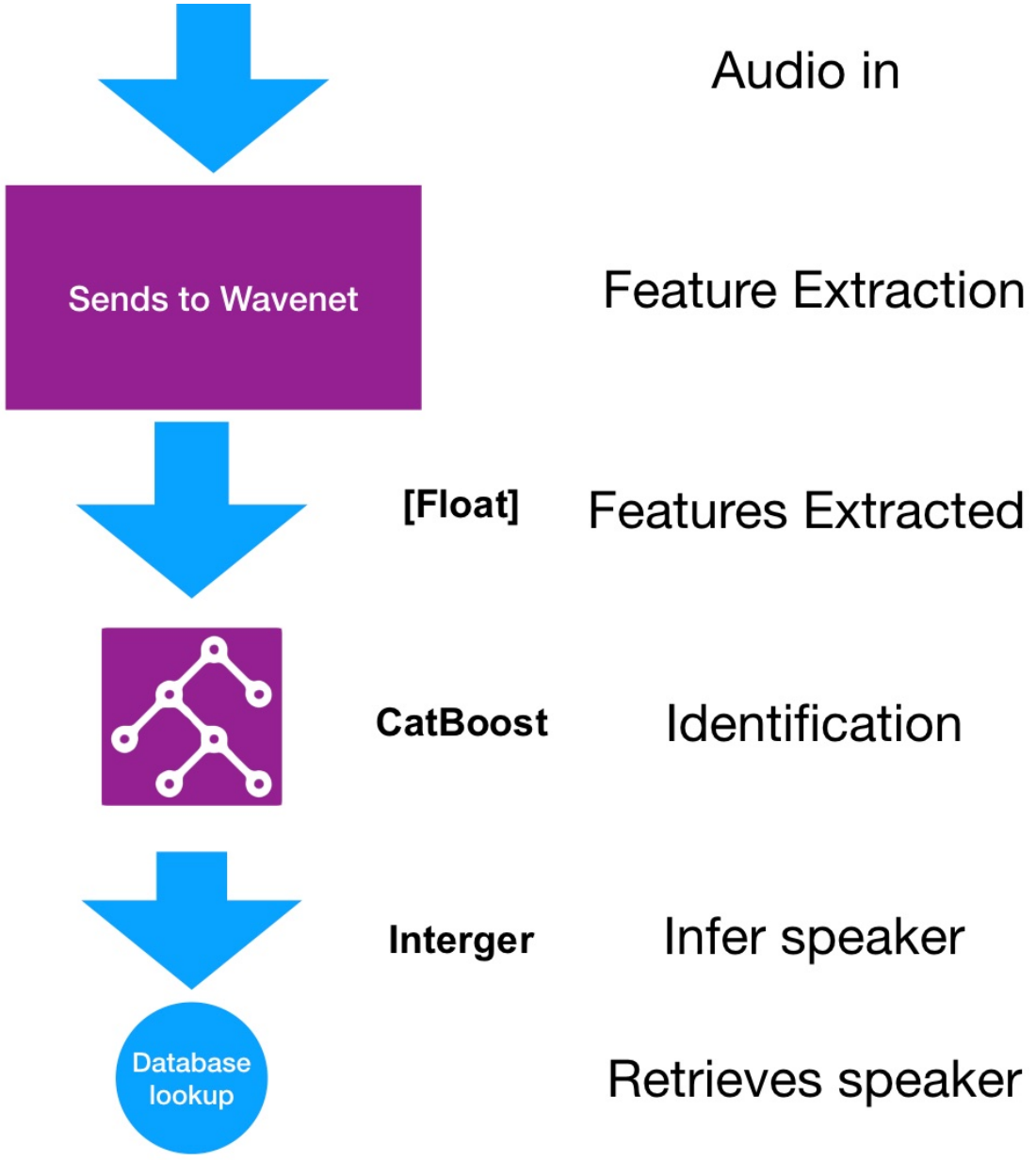


Figure 5.6: The above diagram depicts the information flow that we had produced for our speaker-recognition algorithm. When the phone receives audio data, the audio data is saved temporarily in the phone's storage (not shown) before being passed through a Wavenet, which is feature extraction. The feature extraction results are passed to a secondary algorithm, in this case, a CatBoost before an integer is returned that is processed as a database lookup.

5.4.1 Feature Extraction Through A Wavenet

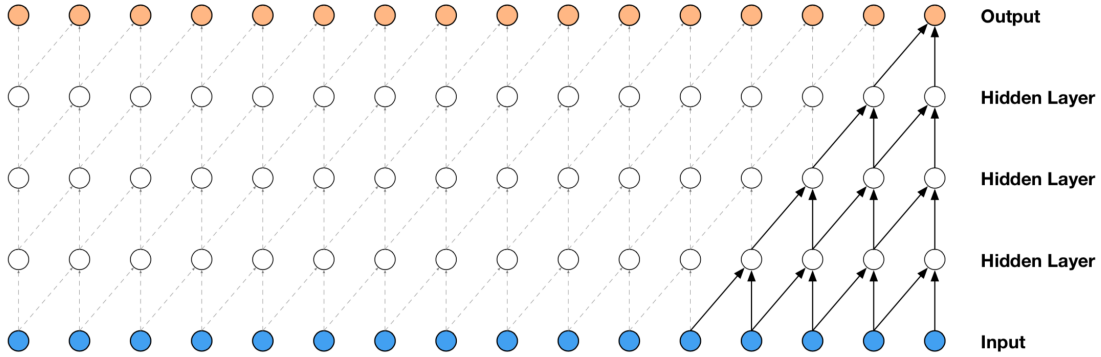


Figure 5.7: An visual reprentation of WaveNet. Credit: [146]

Our feature extraction algorithm that we selected consisted of DeepMinds WaveNet, a convolutional neural network which is designed to produce raw audio for high-quality text-to-speech algorithms for virtual assistants such as Google Assistant [146], which was modified to extract unique features within voices before passing the output to an identification machine learning model. By running feature extraction, we significantly decrease the amount of data that is passed to the identification machine learning algorithm.

WaveNet is designed to work with the properties of sound data, where each sample is relative to another audio sample. WaveNet is designed to be computationally cheap by utilising dilation between layers within the network. Dilation is the process of reviewing N inputs (N being the number of inputs) and returning {1,...,N-1} outputs. In the case of WaveNet, an input of 2 will result in an output of 1 per node resulting in every layer reduces in size by a factor of 2. In other words, an Input of 16 samples would result in layer 1 of 8 nodes, layer 2 of 4 nodes, layer 3 of 2 nodes, and layer 4 of one node which is returned by the function. A visual representation of WaveNet can be seen in figure 5.7.

By utilising dilation, we significantly reduce the computational cost of extracting features, allowing us to run feature extraction on a mobile device. Dilation also allows us to step over new samples with very little computation work as we can also cache dilation results temporary.

Dr Joss Whittle produced the WaveNet within TensorFlow (TF) and when tested it, achieving 99% accuracy. Whittle Further pruned, the process removing unnecessary layers of the model to further to increase efficiency. Once pruned the model was converted to a TFLite model. A TFLite model further increases the performance of the model by removing unnecessary functions and layers while optimising the architecture for an ARM processor; ARM processors are standard processors architectures that are used within phones for reduced architectural size compared to X86, the standard processors in laptops and desktops, while still maintaining accuracy.

During the conversion process to TFLite, we found that one of the functions within the model, the Leaky Relu quantisation function, incompatible with TFLite, resulting in Whittle transferring the model to TF2, which resulted in changing several functions to make the model compatible with TFLite 2 which contained the Leaky Relu quantisation function.

5.4.2 Identification

Here we will describe how we handle the data from feature extraction and infer the speaker. We required a smaller algorithm compared to the feature extraction as this would have to be retrained on the device once a new voice is added to the system. In this section, we reviewed three algorithms, random forests, CatBoost

and K-means clustering along with a novel approach to increase speed to identify frequency speakers through deploying algorithms in series.

5.4.2.1 Random Forests

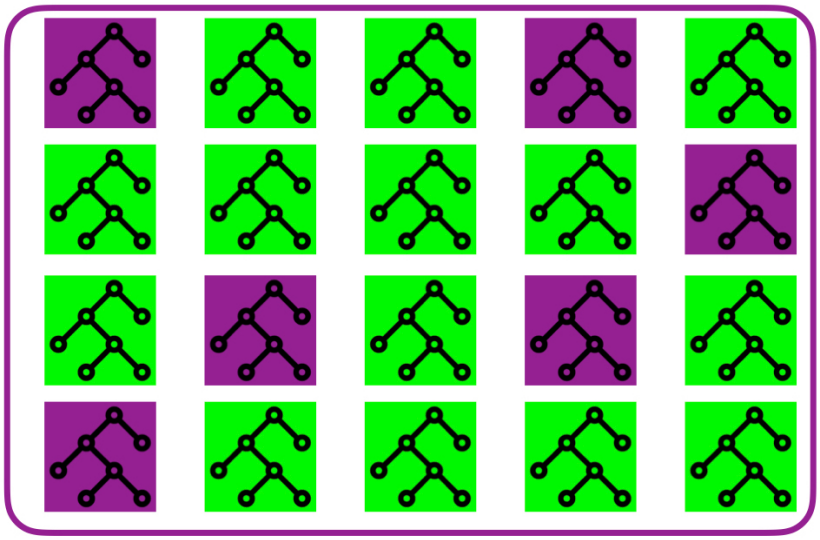


Figure 5.8: An visual reprentation of random forests. Green represents correct trees while purple represnets incorrect trees

A random forest is a classification algorithm that consists of several random decision trees which each tree receives the same data. Each tree votes on which speaker it believes the speaker is and the votes are averaged using the median function (otherwise the most). The result of the median function is then the decision of the Forrest. A visual representation of random forests can be seen in figure 5.8.

A tree consists of three main properties. These are a root node, which we enter the tree on and child nodes. A child node can be its recursive root node or a return node. Each node will contain a function which will decide which child node the algorithm will transverse down. At the training stage, trees receive an unique segment of data that no other tree will receive as their training data, resulting in each tree coming to a different conclusion.

We use multiple trees to generate forests because trees are sensitive to the data that they receive during training leading to substantially different results among the trees. By having multiple trees, we smooth out how sensitive trees are likely to be to results [158].

Random forests are computationally light to generate and run for inference. However, we will require a significant amount of trees that could make the process significantly slower on mobile. However trees are very sensitive to training data and if the testing data is difference is significant enough from training, trees will struggle to infer correctly. While many trees can compensate for this, this increases computational power, meaning random forests may not be suitable for large datasets on mobile phone proccessors [158].

5.4.2.2 K-Means Clustering

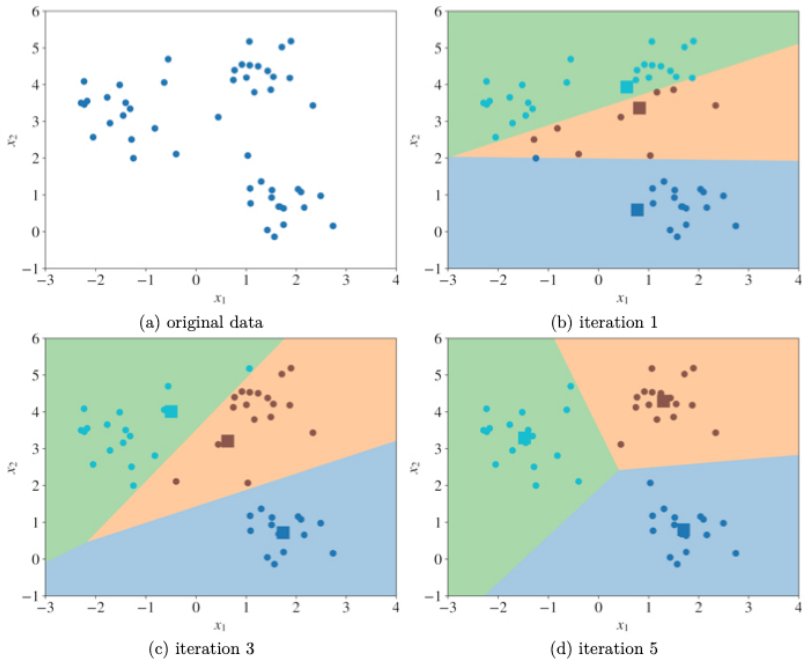


Figure 5.9: An visual representation of k-means clustering. The top left image represents the original data. In the top right image the centroids are placed into the image randomly. By iteration 3 in bottom left, clusters start to form and in iteration 5 clusters have sufficiently formed. Further iterations may result in a decrease in accuracy. Image credits: [21]

K-nearest neighbour is a classification algorithm that works by placing centroids in the training data and during each iteration of training moving centroids closer to their category until they are positioned centrally within their category. All data can only belong to one category. K-nearest neighbour contains K categories, K being in our case the number of voices that the system stores. This processes is presented in figure 5.9.

K-nearest neighbour performance deteriorates quickly after a certain amount of categories, which depends on the data. The performance deterioration problem will lead to poor user experience to the user, leading to frustration and confusion. As performance deterioration depends solely on the data that the algorithm receives, it means it is difficult to place a hard limit on the number of voices.

5.4.2.3 CatBoost

A CatBoost, short for Categorical Boosting, is a type of gradient boosting algorithm that operates on decision trees. Gradient boosting approaches are suitable for noisy data as they can perform gradient descent in feature space [125]. CatBoost is decision tree-based, combines features to produce a new feature, and considers this data greedily, a combination of the intra-tree feature - by combining fields - generation and inter-tree and inter-tree generation by combining previous tree features. CatBoost first splits, it does not consider the new features [33]. The other beauty of CatBoost is that it detects overfitting. [156].

Gradient boosting is traditionally one of the more efficient methods to build models and combining gradient boost with trees offers superior results compared to other algorithms. Traditionally iterative training of multiple trees usually results in overfitting, which is why random forests trees are trained on random data. Gradient Boosting improves this by computing the gradients from the loss functions and teaching the decision trees from the predicts gradients of the loss functions [36].

CatBoost incorporates ordered boosting along with a superior algorithm for processing categorical features. Ordered boosting performs random permutations of the training models and maintains these differences in other models. Usually, this will result in added complexity; however, as CatBoost utilises one tree structure that is shared by all the tree models, it reduces the added complexity [120].

5.4.2.4 Deploying identification algorithms in series

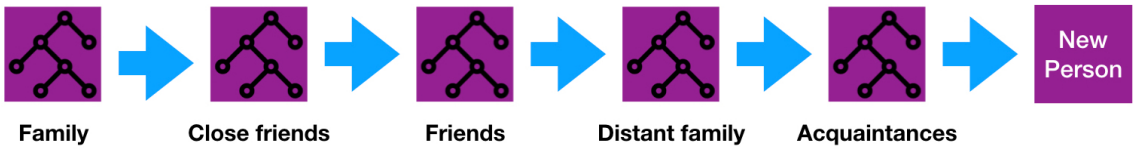


Figure 5.10: A visual representation of using random forests in series as a method of speeding up the recognition of the speaker.

We also considered using a combination of algorithms to increase speed and reduce the computation cost of these algorithms. As an alternative to testing all possible voices at once, here we consider the ability to place several algorithms in series and order voices in the frequency of recognition. An example can be seen in figure 5.10 However, this approach has numerous drawbacks that could affect the user experience of this approach.

To increase performance and to decrease the number of trees/categories that each algorithm would contain, we considered using smaller algorithms but used several of these based on the frequency of speakers. For example, if the user, Emily, were to see her partner, Tom the most, Tom would be in the first K-nearest neighbour algorithm. However, Ella, someone that Emily sees only at Christmas would be placed within one of the last few K-nearest neighbour algorithms.

By placing Tom at the first stage, it will quickly increase the speed that Tom is recognised resulting in better user experience. However, Emily may be able to recognise Tom regardless, depending on the extent of Emily's difficulty to recognise faces.

However, this method may significantly increase the time required for the algorithm to recognise Ella. As the algorithm will have to cascade through several algorithms, it may result in two scenarios, where an algorithm before the one Ella is positioned stating it is someone in that algorithm resulting in the incorrect name or Ella name taking so long to process that Emily has already had to asked it. Length of time in this situation may be less problematic due to social constructs; however, in this case, the technology has resulted in poor user experience for Emily.

5.5 Inability to place onto mobile devices

We were unable to convert the algorithm from TF to TFLite due to incompatibility issues between the Frameworks. We utilised LeakyReLU[8], which is a specific type of Rectified Linear Unit (RLU) in TF. An RLU is a common type of activation function which returns zero for negative inputs but a positive output for positive inputs which helps the model to smooth for the interaction effects. The interaction effects are where a variable in predicted different on different variable [15] or an unseen bias.

We used a Leaky RU in the activation layer of our WaveNet, but we found while converting the models to TFLite that this layer was not supported. While there are multiple open requests on TFLite [9]. As a

[8]https://www.tensorflow.org/api_docs/python/tf/Keras/layers/LeakyReLU
[9]Issue 1), https://github.com/TensorFlow/TensorFlow/issues/27996
Issue 2) https://github.com/TensorFlow/TensorFlow/issues/26755

result, we are unable currently to place our WaveNet on mobile devices without re-engineering our Machine Learning algorithm.

While we could change the layers to one that TFLite supported at this stage, we wanted to evaluate how our Machine learning algorithm would perform on voices and then alter the algorithm based from feedback from our evaluation studies.

5.6 Changes to the application based on Machine learning

For our machine learning, we took into account bystander privacy in designing the system. For identification from machine learning, we will only need to store temporal data. However, if we were to change any of the machine learning algorithms, we would need to train the new algorithms, and as a result, we would have to keep audio data on the device. However, this changes how we respected privacy as previously we made it clear that we would not be storing audio data, but here we would need to store audio data.

From this, we decided that we would store audio data, however, this data would not be exposed to the user or other services on the device. Audio data can be encrypted and stored in the secure Keychain, meaning that the extraction of the data is challenging. Any application that we deploy to users should make it clear that we are storing audio data and the reasoning behind it and also why the users are unable to access the audio data.

6 Evaluation

In this chapter, we will explore how delays will affect the quality of the conversation. From our literature review, we demonstrated an accuracy of 70% to 84% was required for algorithms to become acceptable, we did not find any literature surrounding how delays would effect acceptance. Machine learning requires some time for the algorithms on mobile devices to process, causing latency between listening and replying. We can increase the speed of our application by transferring audio data and machine learning tasks to an external computer, however this does trade-off privacy. We wanted to explore whether the delay was acceptable.

We further developed a speaker-recognition algorithm to evaluate its performance with watch data and Librivox data and achieved a low accuracy algorithm.

6.1 Evaluation of Application During Conversations

Figure 6.1: Experimental setup recreated with researchers: a participant (left) in discussion with an actor (right) with a researcher (centre) triggering the notification

As discussed in chapter 2.6 There is a tipping point for when technology performance becomes accept-

able for people to use in daily lives. While our literature review focused on the tipping point for where speech-to-text became accepted by users, we wanted to understand what would be the tipping point for delays in speaker recognition algorithms to support people with social interaction.

While Machine learning can be fast, it still requires some time for the algorithms on mobile devices to process, which is not instant and causes latency between listening and replying. We estimated that processing on the phone would take 8 seconds based of running the algorithm on a graphics card and comparing the performance difference to that of an iPhone with an A11 Bionic.

We hypothesise that offloading the processing to the cloud under 5G to Amazon Web Servers would lead to the virtually immediate inference of voices (within 2 seconds). However, the approach of sending data to the cloud weakens the users control of data, which was a pivotal decision we made to design the algorithm on a mobile device. While partial off device machine learning can improve speed and retain some control of the data; however, this approach still required data to be processed on the phone and internet latency (an estimated 5 seconds).

Further delays do occur between sending data between the watch and the phone. Within the Watch Communication Framework that allows communications via the Apple Watch and iPhone[1], delays do occur, for example when the iPhone application requires waking from memory. When the iPhone application is loaded, the watch then transfers data through Bluetooth, which in our experience adds an average two-second delay in the transfer process. As research by Mandal et al., [96] has demonstrated, using wearable devices to run inference will result in a worse user experience, as Mandal found that using Google Glass for facial recognition resulted in a significant usage of battery performance along with increased processing times.

In this study, we want to understand how the delays in our system will result in the acceptance of the application by general users with no known underlying condition. We want to work out the tipping point where our application becomes acceptable. From these numbers, we can infer whether the trade-off in speed for improvedment in privacy would impact the acceptance of our application.

To understand how delays would affect how people use the application, we developed a wizard of oz prototype of the application. We recruited 7 participants to evaluate our application by using the application in conversation on the Apple Watch. Each participant would meet an actor that they had not met before and asked to carry out a conversation with them. They would not know the actors name or the conversation topic until it displayed on the watch, where the participant was alerted by a tap and a chime as seen in figure 6.2c. The researcher controlled the time taken from the actor speaking until the name displayed on the watch.

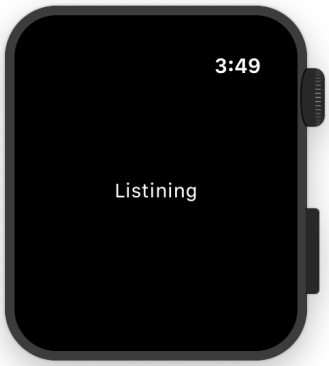
6.1.1 Methodology

To understand how delays would affect how people use the application, we developed a wizard of oz prototype as previously discussed of the application to ensure that we were testing the delay and not the performance of the algorithm. However, participants were not aware that the Wizard of Oz was taking place and that the application was listening and a machine learning algorithm was generating the results.

At the beginning of the session, we explained to the participants that the watch was attempting to infer the speaker and we were evaluating how delays between different modes of inference would affect the usefulness of the application. Participants were told they would have have five trial discussions with the researcher performing different personas before meeting five actors. Participants were made aware of the topics that would come up but were not aware of the names of the people they would meet. However, to ensure that they did not know any actors, participants were given a list of potential actors and asked if they knew any of them. If they did, we either used another actor or if this was not possible, we used a pseudonym on actors. When we used a pseudonym, we highlighted the use in data and noted it to the actor, and we removed the data from the dataset if it was an outlier. Topics were tailor made to the participant and the

[1] https://developer.apple.com/documentation/watchconnectivity

(a) Tap to listen button



(b) The watch passively listening

(c) Speaker has been detected

Figure 6.2: Above are the displays that the participants interacted with on the Apple Watch. Display 6.2a was the screen that participants tapped before going into conversations. We used a trigger was utilised to preserve battery life. Display 6.2b was displayed until the watch displayed a speaker in 6.2c. When display 6.2c was presented, the watch would sound a chime and vibrate.

actor. Actors were aware of the topic beforehand.

Delay was randomised to be either 2, 4 or 8 seconds by the researcher and the delay was triggered as soon as the actor started to speak. After the delay finished, the watch would vibrate, displaying the name of the actor and topic as can be seen in figure 6.2c. Participants would then try to work in the name of the actor and the topic without disrupting the conversation.

Once the conversation had finished, we asked participants to rate on a Likert scale the quality of the conversation, whether the prompt was useful and whether the delay that was triggered was too long or before they needed it. We compiled these results together, removing any identifiers of participants and actors.

We ensured that each participant apart from two had experienced each of the possible delays. We also varied delays with each actor to remove biases of the ability of the actor to converse. Two participants had delays set at a fixed time, one at two seconds and another at eight seconds to understand how outliers in the data would affect the results. These participants did not experience any other types of delays.

After meeting each of the actors, participants then took part in a short qualitative study to discuss their perceptions of delays, and how useful they found the application. Once participants had taken part in the interview, the researchers made participants aware that they were not using an algorithm and that the researcher controlled the name displaying along with the delay and that no personal information was collected.

We analysed each participants exit interview for key themes brought up by the participant and then compared to other participants themes. We did not share previous themes that came up in the sessions with participants to ensure that they gave their own perception of the app.

We recruited 7 participants who took part in 10 conversations each, resulting in a sample size of 57 conversations as we discounted the first discussion to consider learning biases. We discounted one conversation due to error by the researcher in triggering the notification (1x incorrect topic which confused the actor: 1x Actors unavailable: 1x participant recognised one actor and did not use the watch).

Ethics for this study was approved through a short form with the College Of Science Ethics Committee at Swansea Univeristy.

6.1.2 Limitations of this study

We have identified the following limitations of our study configuration:

Variance between actor speaking and researcher triggering

As the researcher triggers the alert, the researcher may introduce slight variance in trigger time. We instructed all researchers to trigger the alert as soon as the actor had spoken, however, this would still lead to a slight delay. We estimate that this might add up to a one-second delay on notifications which will affect the perception of delay more on the two seconds delay compared to the perception of delay on eight seconds delay.

Signal loss between Apple Watch and iPhone

We identified that in certain conditions, the Apple Watch might lose connection to the iPhone and fail to reconnect. These conditions that we noted were the participant and researcher climbing or descending stairs or use of a lift. To ensure that the Apple Watch had reconnected to the phone successfully, we instructed the researcher to test the application before meeting actors. If the application did not reconnect, the researcher was instructed to force the application to close.

Variance in Watch Communications.

The communication framework that Watch Connectivity offers is not instant and is delayed for function SendMessage(). Apple documentation states that the reply is instant for SendMessage() compared to transferUserInfo() and transferFile() which take place in the background [10]. While SendMessage() happens instantly, the iOS application itself requires to be woken up to reply [7].

In our experience, we have found that it takes on average one to two seconds for a response with no delay. We have not found any method with shorter delays. These delays add a two-second variance to our application which we must consider with results. Nevertheless, these delays would still occur without the speaker recognition algorithm, and we are testing the delay that the algorithm causes, not the delay on how long it takes the watch to communicate.

Error by actors

We briefed actors before the session when we explained the study and the topic. However, actors sometimes stated the topic beforehand to the participant or had lead to confusion where the participant stated another topic than they said. To mitigate the situation above from effecting our data, we attempted to meet another actor, or if this was not possible, we discarded the data of that conversation.

6.1.3 Results

Delay Time	Conversation Quality	Usefulness of prompts	Usefulness of watch
2 Seconds	8.29	8.38	7.86
4 Seconds	8.06	8.00	6.88
8 Seconds	7.60	6.65	5.80

Table 6.1: This table displays the mean averages from how participants rated the conversation on a scale 0 to 10.

We found that participants rated the quality of the conversations decreased in association with increasing delay between 2 and 4 seconds. Participants rating on how useful the watch was fell by 0.98. However, participants rated the usefulness of the watch decreased between 2 and 4 seconds by 1 on the Likerk scale. A further decrease in response time between 4 and 8 seconds resulted in the quality of the conversation and the usefulness of the prompt falling again. Participants found the usefulness of the watch declined as well demonstrating that an eight-second delay is not acceptable. These results can be seen in table 6.1.

Observations by the researchers suggested that the main factor affecting any given conversation was the ability of the actor and the participant to make small talk. Some participants were not able to make small talk at all and could not engage with any conversation until a prompt appeared on the watch. However, some participants (notably P1 and P4) were able to converse naturally with actors and generate small talk while waiting for the prompt to display. However that individual participants were unable to make any small talk and found any delays in the application difficult (specifically P5), highlighting that the ability to small talk was a part in the quality of the application.

These results suggest that the length of the delay impacts the user experience of the application and that users found a delay longer than four seconds was not acceptable. These results also demonstrate that while other contributing factors may play a significant part in the conversation, especially the ability to make small talk, which could warrant further exploration - users consider speed to be a vital factor in the application. This raises questions about how a wearable system might achieve these speeds that we discuss later.

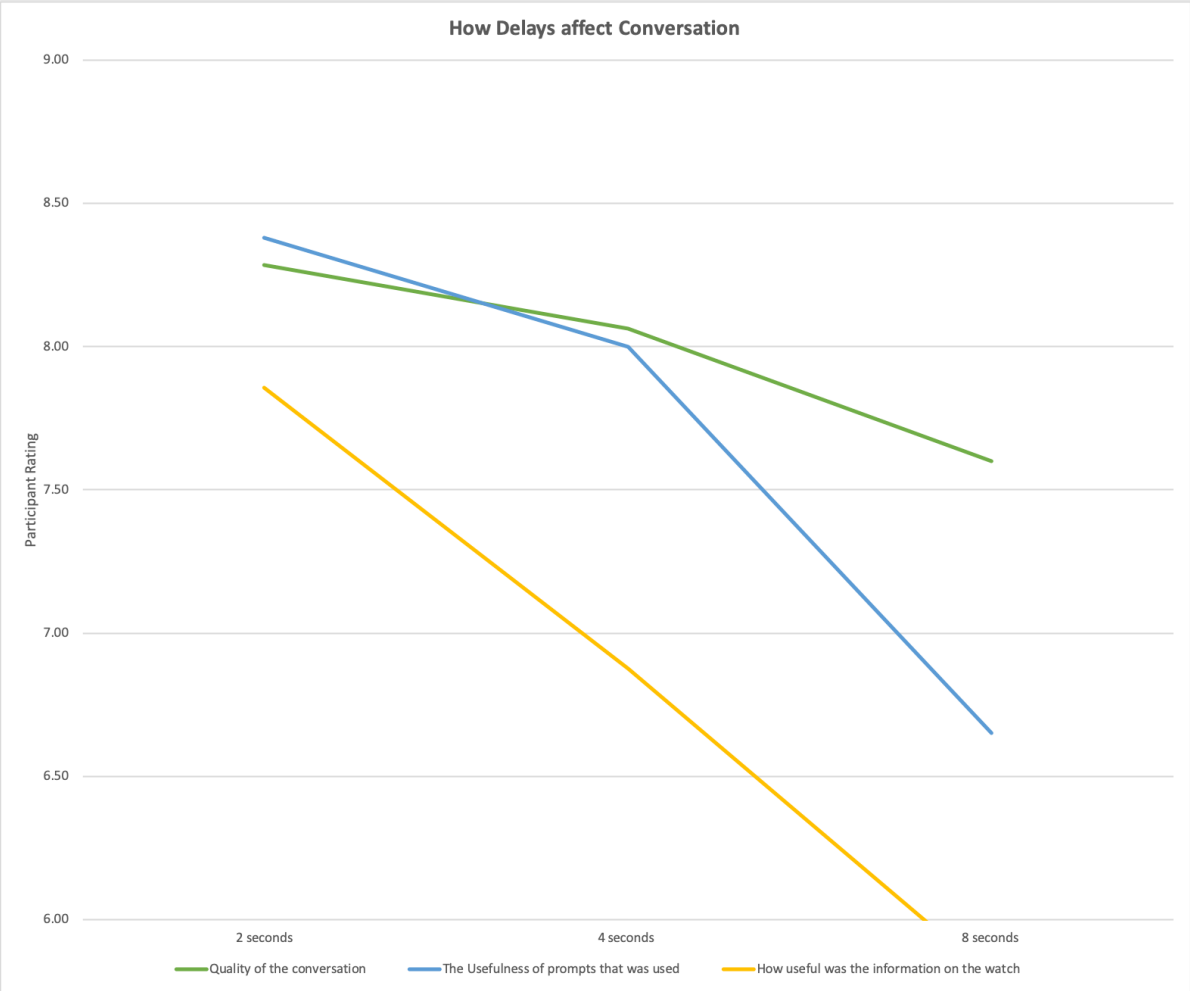


Figure 6.3: Above demonstrates the trend between delays and rating of the application

6.2 Evaulation on audio from an Apple Watch And Libivox

A crucial part of our application is the capturing of audio data from a smartwatch, or in this case, the Apple Watch. Several variables create the characteristics of the microphones, which can affect the quality of the microphone. We cannot control these variables, but we can test to see how these samples will affect the quality of the audio for machine learning purposes [37, 35].

Frequency response is the first characteristic that affects the quality of the microphone. Frequency response is how the microphone responds to the different frequency. Some applications require a low-frequency response such as a concert where more bass is needed, while a higher frequency will result in the more treble [37, 35].

Sensitivity is the characteristic that controls the output voltage to the input pressure, or how loud the microphone perceives noise relative to the actual sound. Sensitivity is task-dependent, which usually relies on how far the microphone is from the source. For example, a mobile phone would require low sensitivity as being too sensitive would result in too much distortion. However, a microphone which is far from the speaker, for example, microphone on a smart assistant such as Amazon Alexa Dot, would be more sensitive to noise [35].

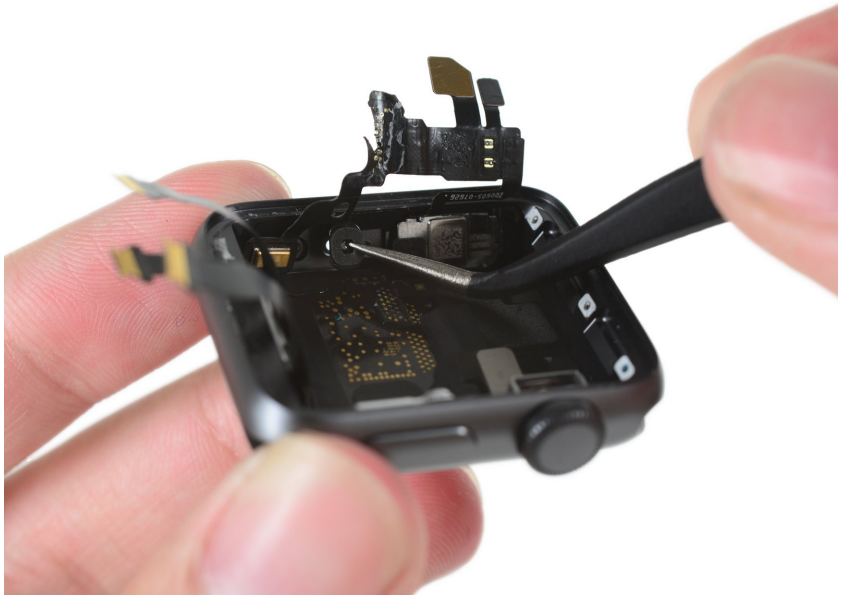


Figure 6.4: An image of the microphone assembly inside of the Apple Watch Series 2 during a teardown carried out by iFixIt [64]. Here the tweezers are removing the ingress protection for the microphone showing an O-Ring gasket.

Microphones also produce noise as a byproduct of recording from the small amount of current running through the speaker. As the signal is amplified by the factor of a thousand to make it audible, this can result in electrical noise produced by the microphone, resulting in the smallest amount of noise can become prevalent. Reducing the prevalence of sound is part of the design of the microphone and can be challenging to mitigate after the microphones have been manufactured [35].

We based the study on the work of Titze et al. [142], which explored how different microphone types work on extracting voice perturbation. Work by Titze et al. used three microphone placements (4cm, 30 cm and 1 meter from the source) and three angles of microphone placement from the source (0°, 45° and 90°). Titze used a loudspeaker as the source ,as loudspeakers offered a wide range of control and are more consistent than a human speaker [142].

We used a loudspeaker to control utterance lengths and to ensure that the audio was clear. As discussed in chapter 5.3.1, it would be difficult for us to collect a wide range of demographics.

6.2.1 Methodology

In this evaluation study, we evaluated the audio that the apple watch recorded through the app "Smart Record" and then fed this into the speaker. "Voice Recorder , voice memo"[2] (Voice Recorder) was utilised as it did not contain a limit on recording time and did audio transfers in the background, which is uncapped, unlike WatchCommunications SendData() method. However, background syncing has a delay while SendData() is immediate. We utilised Voice Recorder as it allowed us to develop this study rapidly.

We captured the audio with the following background noise:

- No background noise
- Beach

[2]https://apps.apple.com/gb/app/voice-recorder-voice-memo/id609030412

- Pub
- Night Club
- Coffee Shop
- Train
- Aeroplane

These background sounds were played through a Bluetooth speaker where an alternative Bluetooth speaker was playing audio of a speaker of librvox.

To control this study, we used the same quiet room to control background noise. We also used the same Bluetooth speakers, each doing the same task to ensure a consistent output. We also used the same audio loop. We also used the same Apple Watch. We used an Apple Watch Series 2 (42mm Silver Aluminium).

Data capture

We selected five male, and five female voices with clear utterance from the training data, which we then removed from the training data set. In the folder hierarchy, each voice had a labelled folder derived from the dataset with each sample labelled with the scenario that it captured.

Audio data was captured directly onto the Apple Watch using Voice Recorder. Each background noise was saved in a separate file to remove any drift from occurring. For the first ten seconds of audio recording, the user kept the phone's microphone a set distance from the speakers. For the final ten seconds, the user would then walk to a set location with the microphone facing the speaker to measure the sensitivity of the microphone and how it would impact training.

We configured background audio to play at 50Db (the equivalent of a quiet home), and we played the speaker audio at 70Db (the equivalent of a conversation). We measured this before the study took place using Decibel X [3].

6.2.2 Limitations

We identified the following limitations of our experiment configuration:

Bluetooth speaker output quality

Firstly output from the Bluetooth speakers may contribute to the output of the results. These variations in outputs is a result of their characteristics which could not be overcome. However, using Librivox, we could not get the narrators back to speak with a natural voice. Further, we did not visit these background locations. We gathered audio from Youtube, which may contain compression artefacts. We could not capture from these environments due to bystander privacy concerns.

Compression and alteration to audio in Voice Recorder

Secondly, the application used, Voice Recorder, may have also applied compression to their recording. While the output of the files was in 48khz which we downsampled to 16khz in FFMPEG, a compression algorithm may have been applied during the transfer. We think this is unlikely as the WatchCommunication framework handles large files already and any compression may add complexity.

Manufacturing tolerances and wear in the Apple Watch

[3] https://apps.apple.com/GB/app/decibel-x-DB-dba-noise-meter/id448155923

The final consideration is manufacturing tolerance in the Apple Watch. While consumers regarded that Apple quality is high, Apple themselves would have accepted tolerances. Apple does not publicise these tolerances. We could utilise several Apple Watches to test for the quality of its audio. However, we did not have access to these resources during this MRes.

Further, the Apple Watch that was used for data collection was used previously in watersports and cycling. Saltwater from watersports and dirt from road cycling may have developed inside the microphone cavity and decreased the quality of the microphone. While we did wash the watch 24 hours before any sample collection, this would not combat degradation or damage done to the microphone. This watch was three years old, and we believe that future work to investigate the comparison of a new Apple Watch and an older apple watch would be necessary to understand microphone degradation. While this specific Apple Watch is waterproof, it is protected by an O-ring which may not completely protect the microphone from the water, which can be seen in Figure 6.4. While the Apple Watch can pump water out [64], in our experience, this does not always remove all the water, resulting in evaporation being the best method to remove all the water. To mitigate this, we ensure that the Apple Watch did not get wet for at least 24 hours before being used for recording to allow any water to evaporate.

6.2.3 Results

We failed to achieve a high accuracy algorithm in our machine learning algorithm from the watch audio data. After the variance in noise and a large number of trees, we were only able to achieve an accuracy rating of 14.8%, which currently is not suitable for being used to support people with difficulty recognising faces. We did achieve higher accuracy predicting testing data of 90%. However, using audio data from Librivox data along with a more extensive training size, we were able to increase our accuracy to 42.8%, which is nowhere near the 75% accuracy we found as a tipping point for acceptance of the technology which we discussed previously. With further training and a more optimised algorithm, this number would further improve. When Random Forests were utilised instead of CatBoost on Apple Watch data, we achieved similar results (13% on testing data) which make both algorithms unsuitable.

We hypothesise the reason for the poor accuracy is down to the mismatch between Librivox data and Apple Watch data. While both algorithms trained on 16khz mono audio, there may be significant enough difference between the Apple watch audio and the Librivox audio to cause a mismatch within the feature extractor. To compensate for this, we can use audio data from the Apple Watch to train our feature extractor. Training the feature extractor on audio from the Apple watch could significantly increase accuracy to the same level we were having with Libriox data.

We further hypothesise the reasoning behind the poor accuracy from the smartwatch is the poor quality that the smartwatch microphones capture. When we listened back to the audio, we noted the low quality and how difficult it was to distinguish the speaker from the surrounding audio. Noise-cancelling can further improve this. However, the Apple Watch does not currently support noise cancelling through its microphone.

Further tuning to the WaveNet may improve the speaker and possibly training on raw audio from the Apple Watch may help further. A filter could also be placed before the feature extractor to be a feature extractor of noisy audio into clean audio. We can also improve the trees by training the trees on more variance and more data. We can also consider using another algorithm as a form of identification, such as utilising a K-means cluster as discussed in section 5 however, it is vital to remember that these algorithms should run on mobile devices and with as little variance as possible.

As discussed in Section 6.1 participants ratings of the watch as a communication aid decreased dramatically with longer delays demonstrating that delays have a significant impact on the user experience of a device. While dedicated hardware would be faster than a mobile phone neural network processor such as a remote server, it will still take a server time to compute. As a result, we want to reduce the computational load as much as viable. Increasing tree and leaf counts increase our computational load, and these trees should be as small as viable without a loss in performance.

At current timings, our approach to our algorithm is not suitable to support people with recognising people and did not archive our target of 80%. However, with further work, it should be possible to increase our accuracy.

6.3 Changes required to the specification for future workings

From our user evaluation, delays should be as short as possible. As a result, any further systems that we design should contain delay as short as possible, which may result in the reduction of privacy. There needs to be a balance between privacy and speed, where on device processing is more private as no data leaves the device however compared to off device processing is slower, and currently, we feel that speed should be a priority. However, we must ensure that privacy is respected as users will not use the application if they feel that is abusing bystander privacy.

The algorithm that we developed does not result in sufficient accuracy and significant work must be required in this area to ensure that the accuracy of the algorithm is for use within the app.

6.4 Conculusion

In this chapter, we have found that a small delay in inference is acceptable within the application. Speed should be the priority of the application, and long delays within the system are unacceptable and lead to a dramatic decline in user experience. Delays in the application influenced the quality of the conversation, which demonstrated that participants did use the watch as a communication aid.

We also found that the machine learning approach was unacceptable and resulted in a low accuracy with Apple Watch Audio. However, using Librivox data, we were able to achieve higher accuracy algorithm. Higher accuracy from Librivox demonstrated that with further pruning, we might be able to achieve a more accurate machine learning model. Our WaveNet may require further training on Apple Watch data to compensate for biases between Librivox audio and Apple Watch audio.

7 Conclusion

In this chapter, we will discuss our research and our findings from this MRes project. We will further develop our Conclusion and then will layout potential avenues for future work.

7.1 Discussion

In this chapter, we will develop the discussion of the MRes and highlight potential areas for future work. In this research, we have demonstrated that there is potential to use speaker-recognition to support people to recognise faces. However, there are limitations to this work which will restrict its adoption.

7.1.1 Privacy

Privacy was core to our development of Pwy, with us wanting to take an approach using on the device inference to ensure that users were in control of their data. It would also allow participants to use the application in situations where the internet was inaccessible. No group showed any concerns for self-privacy. However, worries about bystander privacy varied significantly between groups.

Expert designers felt the system was a useful communication aid for people and could be widely accepted though it would benefit from social media to share voice data. People with difficulty socialising wanted a passive listening application to help them with confidence with who they are in discussion with and had minimal regard for bystanders privacy. People with TBI were concerned with bystander privacy but felt that people would be accepted once made aware that it was an "extension of their brain" as they would happily explain what it was doing, they did not want it to become a topic of conversation.

During our evaluation of the delays of the watch, no participants showed any concern of privacy with many of them stating that they would use the application if it became available. However, in that specific application design, participants had very direct control of listening of the application, which adds an extra level of privacy.

In this research, we did not make a distinction between pausable passive listening, where the user could pause listening, and constant passive listening, where the user could not pause listening. With data that we collected in all activities, we are unable to infer how this would affect the watch, and further work in the area would be required to understand limitations.

We believe that capturing audio from the smartwatch has benefits compared to smart glasses which previous work [96, 148, 154]. While research does show that if participants disclosed their conditions that people are generally more accepting [124], we feel that with smartwatches people will be less willing to question the uses of AT, shifting the focus to conversations that do not surround specific conditions.

Finally, in this research, we did not make a distinction between pause-able passive listening, where the user could pause listening, and constant passive listening, where the user could not pause listening. With data that we collected in all activities, we are unable to infer how this would affect the watch, and further work in the area would be required to understand limitations.

7.1.2 Stigma Free Accessibility Tools

We wanted to develop an Accessibility Tool (AT) that was stigma-free. We wanted people to use application and the smartwatch without drawing attention to the technology. As discussed in our presentation "Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury" we outlined that we wanted to develop AT that looked commonplace and to utilise off the shelf products before having to produce anything custom. For example, we did look in quickly using a raspberry pi[1] as a method for running the neural networks. However, this would have required another device for the participant to have worn.

None of the participants stated that they felt that the watch was an AT or being used as one. Participants felt it was a ubiquitous piece of technology, with at least four participants already using a wearable piece of technology and a further two participants looking at purchasing one. Many participants stated that the would use the application today if it were further polished with one participant in our evaluation study stating "this would be very useful as people in my family have short term memory issues."

7.1.3 Limitations of Use

From evaluations of our machine learning algorithm and when listining from the watch, we achieved low accuracy in our machine learning algorithm. We discovered that microphones on the Apple Watch are low quality and do not work in all situations. While the audio from a beach, train and coffee shop had achieved higher accuracy, in social situations such as a bar or a night club, the accuracy deteriorated. These are common situations that users may find themselves in and find that they are unable to use the system.

The situations can be stressful enough, especially for people with difficulty socialising, where a lot of new people will be present. If the application keeps failing, this is likely to lead to users becoming further stressed. A filter for voices could be applied to help with these situations. However, a filter may still not be enough if the quality of the audio that is coming into the watch is still not sufficient. Spatial awareness from audio may help with the watch to identify where people are to allow audio to focus on.

However, without further work, such as further training the machine learning algorithm on audio data from the watch, we are unable to archive a higher accuracy algorithm. As an outcome from this project, we are unable to have an application that can support people to infer the speaker. However, further works should focus on increasing the accuracy of speaker-recognition models to an accuracy level that is suitable to support people.

7.1.4 Effect of Delays on the system

Delays in the system do impact the user experience (UX) reducing the perceived usefulness of the prompts and the quality of conversation a user can have. A significant decrease in UX was linked to a decline of between 2 and 4 seconds, with a further decrease in UX when increasing the delay from 4 to 8 seconds. This study demonstrates that for the application to support people, the application must be able to work as quickly as possible. Current hardware introduces a 2 second delay that is difficult if not impossible to work around but each delay beyond this further reduces the quality of the work done. This suggests that speed needs to be a top priority in the future.

[1] https://www.raspberrypi.org

A trade-Off of privacy for speed could help in these situations as the general population found that 8 second delays was not effective support. We did not run the study with people with TBI. However, we hypothesise that people with TBI would be more accepting of a delay.

The capabilities of machine learning on mobile phones has drastically improve over the few years [65], and in this future, it may be possible to run our algorithm in a suitable timeframe on the device. However, on current hardware with current limitations, a trade-off would be required between speed and privacy. A deeper understanding of how the effects with people with TBI would be needed to evaluate the trade off.

7.1.5 Approach of the application

Our participatory design workshops demonstrated that each group had a significant, technically different approach to the problems that arise in conversation which needed different technological approaches to address. While people with social difficulties wanted complete discreetness and constant listening, our design experts wanted social media to share voices and data. Technically, social media and discreetness are not easily reconciled. A social media would have to hand over data every time that an individual has had their voice inferred which is not very discreet.

Furthermore, social media may add complexity for people with TBI which is not suitable when trying to reduce cognitive load. Discreetness and the requirements of people living with TBI were incompatible because those with TBI did not want to deceive bystanders. Participants with TBI were willing to display that they were using the application more with using other modalities such as headphones.

These results demonstrate that one size fits all approach in this instants is not suitable and that each set of participants requires a tailor-made solution. Compromises in the system are possible. However, we do not understand how this would affect the UX of the application. Furthermore, without running an evaluation study with users using the application in their daily lives, it is also difficult to understand the impact of compromises. However, three separate applications may need to be developed to support each of the groups.

7.2 Conclusion

In this research, we explored how speaker-recognition can support people to recognise other people through the development of the Pwy application. Firstly, we explored the current research based on supporting people with recognising others and discovered that researchers predominantly used facial recognition through smart-glasses. We also became aware that there are significant ethical and legal concerns surrounding the recognition of people. We also explored the tipping point of acceptance of technology and found it was within the 75% to 85% range.

We explored the design requirements and specifications that stakeholders have for a speaker-recognition application. We explored these requirements through participatory design workshops with expert designers, people who find socialising difficult and patients with Traumatic Brain Injury. We found that the requirements of each of the groups was significantly different, and as a result, a one size fits all approach was not suitable. Further significant bystander privacy considerations between groups highlighted that a single technical solution might not be achievable.

For our requirements, we explored speaker-recognition from a machine learning perspective and built a speaker-recognition algorithm using WaveNets and CatBoost. We then evaluated the algorithm and reached a mid-accuracy algorithm with audiobook data; however, we were only able to achieve accuracy of less than 15% with smartwatch audio data. The reasoning behind these low accuracy is attributed to poor audio from the smartwatch or a mismatch between training data and watch audio data.

We ran a further evaluation study to understand how delays affect the usability of the watch as a communication aid and discovered that the longer the delay, the decrease in its usability to the participant. Delays of between two and four seconds were acceptable but delays of 8 seconds were not. These results demonstrate that off device inference may be required at present, which trades privacy for speed.

7.3 Future Work

As part of our research, we have identified future works for this research project to explore. Below is a compiled list of all identified future works:

- Develop our machine learning algorithm, specifically our feature extractor on watch data to increase accuracy. We do not believe that our current accuracy is not suitable to support people, and a higher accuracy algorithm would then allow evaluation study of people using the app outside of the lab. We would suggest accuracy rating of 80% and above on the speaker-recognition algorithm would be optimal .
- Run evaluation studies on a larger pool of participants from a larger demographic pool. By running with an larger evaluation pool, we can further understand how delays can effect the acceptance of the application to other demographics. For example, currently we do not know how people with TBI find the lenght of delays acceptable.
- Run further design studies to explore the UI with a larger set of participants.By running a larger study this allows us to understand requirements of other groups such as those with prosopagnosia.
- Run a large scale questionnaire with the general population to understand society's current attitude for speaker recognition. We currently do not understand how bystanders would perceive the application and through running a large scale questionnaire it will allow us to further understand how society will react to speaker-recognition as an AT.
- Apply for NHS ethics to research with people with prosopagnosia. Currently we do not have ethical approval to approach people with prosopagnosia and we believe that they provide a valuable insight to our research.
- Run a system study with people who find faces difficult to recognise in the real-world to understand real-world performance. Currently our results are lab based and may not reflect the usage outside of the lab.
- Develop an application that worked on a cross-platform watch such as Samsung Galaxy Watch[2] to ensure the application is not limited to people to the iOS platform. This would allow a further range of participants who do not have an iPhone to take part in a system study.
- Investigate other modalities such as earphones to notify the user to the speaker's identity. We briefly touched upon the use of Bluetooth earphones with people with TBI, however we believe there is other modalities that can be used. For example vibration patterns. Further we do not currently understand non-TBI participants feel about the use of other modalities.
- Launch the application commercially or open source to support the people who had not been part of this research project. This will allow for the general public to use the application who we were unable to work with during our research such as those with Turning Syndrome and Alzheimer's disease.

[2]https://www.samsung.com/global/galaxy/galaxy-watch/

References Section

[1] Abner, L. Google speech recongition is now almost as accurate as humans. *9 to 5 Google* (June 2017).

[2] Ahmed, T., Hoyle, R., Connelly, K., Crandall, D., and Kapadia, A. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 3523–3532.

[3] Ahmed, T., Kapadia, A., Potluri, V., and Swaminathan, M. Up to a limit?: Privacy concerns of bystanders and their willingness to share additional information with visually impaired users of assistive technologies. In *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* (New York, New York, USA, 2018), vol. 2, ACM, pp. 89:1–89:27.

[4] Amazon Customer Service. About alexa voice profiles. https://www.amazon.com/gp/help/customer/display.html?nodeId=202199440, Retrived 28 April 2019.

[5] Android Open Source Project. Privacy: Mac randomization — android open source project. `https://source.android.com/devices/tech/connect/wifi-mac-randomization`.

[6] Andy, I., and Username AriX. How big can the payload be when sending data via watch-connectivity? `https://stackoverflow.com/questions/33025915/how-big-can-the-payload-be-when-sending-data-via-watchconnectivity`.

[7] Apple inc. Instance Method sendMessage(_:replyHandler:errorHandler:) [Apple Documentation website]. https://developer.apple.com/documentation/watchconnectivity/wcsession/1615687-sendmessage.

[8] Apple Inc. iOS Security iOS 12.1 November 2018. Tech. rep., Apple inc., November 2018.

[9] Apple inc. transferFile [Apple Documentation website]. `https://developer.apple.com/documentation/watchconnectivity/wcsession/1615667-transferfile?language=objc`, 2019.

[10] Apple inc. WCSession [Apple Documentation website]. `https://developer.apple.com/documentation/watchconnectivity/wcsession`, 2019.

[11] Arthur, C. Google glass: is it a threat to our privacy? *The Guardian* (Mar 2013).

[12] Atal, B. S. Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America 52*, 6B (1972), 1687–1697.

[13] Bâce, M., Sörös, G., Staal, S., and Corbellini, G. Handshakar: Wearable augmented reality system for effortless information sharing. In *Proceedings of the 8th Augmented Human International Conference* (New York, NY, USA, 2017), AH '17, ACM, pp. 34:1–34:5.

[14] Ball, K. Workplace surveillance: an overview. *Labor History 51*, 1 (2010), 87–106.

[15] Becker, D. Rectified linear units (relu) in deep learning. `https://www.kaggle.com/dansbecker/rectified-linear-units-relu-in-deep-learning`, 2018.

[16] Beigi, H. Speaker recognition. In *Fundamentals of Speaker Recognition* (2011), Springer.

[17] Berkun, S. How to run a design critique. `https://scottberkun.com/essays/23-how-to-run-a-design-critique/`.

[18] BEYN, E. S., and KNYAZEVA, G. R. The problem of prosopagnosia. *Journal of neurology, neurosurgery, and psychiatry 25*, 2 (may 1962), 154–158.

[19] Bowyer, K. W. Face recognition technology: security versus privacy. *IEEE Technology and Society Magazine 23*, 1 (Spring 2004), pp 9–19.

[20] Bryant, R. A. Disentangling mild traumatic brain injury and stress reactions. *New England Journal of Medicine 358*, 5 (2008), 525–527. PMID: 18234757.

[21] Burklov, A. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[22] Busigny, T., Mayer, E., and Rossion, B. Prosopagnosia. In *The Behavioral and Cognitive Neurology of Stroke*, O. Godefroy, Ed. Cambridge University Press, Cambridge, 2013, pp. 231–246.

[23] Cadwalladr, C., and Graham-Harrison, E. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The Guardian* (2018).

[24] Campbell, J. P. Speaker recognition: a tutorial. *Proceedings of the IEEE 85*, 9 (Sep. 1997), 1437–1462.

[25] Choi, J., and Seongcheol, K. Is the smartwatch an it product or a fashion product? a study on factors affecting the intention to use smartwatches. *Computers in Human Behavior 63* (10 2016), 777–786.

[26] Corrow, S. L., Dalrymple, K. A., and Barton, J. J. Prosopagnosia: current perspectives. *Eye and brain 8* (2016), 165–175.

[27] Dana, S. The revolt against google "glassholes"'. *New York Post* (July 2014).

[28] De Luca, A., Hang, A., von Zezschwitz, E., and Hussmann, H. I feel like i'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 1411–1414.

[29] DeGutis, J. M., Chiu, C., Grosso, M. E., and Cohan, S. Face processing improvements in prosopagnosia: successes and failures over the last 50 years. *Frontiers in Human Neuroscience 8* (2014), 561.

[30] Deibel, K. A convenient heuristic model for understanding assistive technology adoption. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2013), ASSETS '13, ACM, pp. 32:1–32:2.

[31] Denning, T., Dehlawi, Z., and Kohno, T. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, US, 2014), no. 10 in CHI '14, ACM, pp. 2377–2386.

[32] Dinka, D., and Lundberg, J. Identity and role—a qualitative case study of cooperative scenario building. *International Journal of Human-Computer Studies 64*, 10 (2006), 1049 – 1060.

[33] Dorogush, A. V., Ershov, V., and Gulin, A. Catboost: gradient boosting with categorical features support. *arXiv* (2018).

[34] Dyer, C. Law lords rule there is no right to privacy. *The Guardian* (Oct 2003).

[35] Elsea, P. Microphones How they work. http://artsites.ucsc.edu/EMS/Music/tech$_{\text{b}}$ackground/TE-20/teces$_2$0.html, 1996.

[36] Ershov, V. Catboost enables fast gradient boosting on decision trees using gpus. *Nividia Developer Blogs* (2018).

[37] ESR Eletronic Compnents Ltd. Microhpone characterisitcs. https://www.esr.co.uk/sound-light/Infomation/microphones.htm, Accessed 5 September 2019 2019.

[38] European Commission. Do we always have to delete personal data if a person asks. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/do-we-always-have-delete-personal-data-if-person-asks.

[39] Ferdous, M. S., Chowdhury, S., and Jose, J. M. Analysing privacy in visual lifelogging. *Pervasive and Mobile Computing 40* (2017), pp 430 – 449.

[40] Francis, R., Riddoch, M. J., and Humphreys, G. W. 'Who's that girl?' Prosopagnosia, person-based semantic disorder, and the reacquisition of face identification ability. *Neuropsychological Rehabilitation 12*, 1 (jan 2002), 1–26.

[41] Frantz, G., Reimer, J., and Wotiz, R. Julie: The application of dsp to a consumer product. 82–86.

[42] Frauenberger, C., Makhaeva, J., and Spiel, K. Blending methods: Developing participatory design sessions for autistic children. In *Proceedings of the 2017 Conference on Interaction Design and Children* (New York, NY, USA, 2017), IDC '17, ACM, pp. 39–49.

[43] Gagliardi, C., Frigerio, E., Burt, D., Cazzaniga, I., Perrett, D. I., and Borgatti, R. Facial expression recognition in williams syndrome. *Neuropsychologia 41*, 6 (2003), 733 – 738.

[44] Gainotti, G., and Marra, C. Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders. *Frontiers in Human Neuroscience 5* (2011), 55.

[45] García, E. L., Banegas, J. R., Pérez-Regadera, A. G., Cabrera, R. H., and Rodríguez-Artalejo, F. Social network and health-related quality of life in older adults: A population-based study in spain. *Quality of Life Research 14*, 2 (Mar 2005), 511–520.

[46] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report 93* (1993).

[47] Gill, C. Automatic speaker recognition using transfer learning. *Towards Data Science* (Dec 2017).

[48] Gish, H., and Schmidt, M. Text-independent speaker identification. *IEEE Signal Processing Magazine 11*, 4 (Oct 1994), 18–32.

[49] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[50] Goodin, D. Shielding MAC addresses from stalkers is hard and Android fails miserably at it. *arstechnica* (March 2018).

[51] Google Glass. Glass explorers do's and don'ts. https://sites.google.com/site/glasscomms/glass-explorers, Accessed 2019-04-29.

[52] Google Support. Link your voice to your google assistant device with voice match. https://www.support.google.com/assistant/answer/9071681?co=GENIE.Platform%3DAndroid&hl=en, 2019.

[53] Google Trends. Call dad - explore - google trends. https://trends.google.com/trends/explore?date=all&q=Call%20dad.

[54] Gotterbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M. S., Miller, K., Varansky, K., J, W. M., Anderson, E., Anderson, R., Bruckman, A., Carter, K., Davis, M., Duquenoy, P., Epstein, J., Kimppa, K., Kisselburgh, L., Kumar, S., McGettrick, A., Milic-Frayling, N., Oram, D., Rogerson, S., Shama, D., Sipior, J., Spafford, E., and Waguespack, L. ACM Code of Ethics and Professional Conduct. `https://www.acm.org/code-of-ethics`, 2018.

[55] Grüter, T., Grüter, M., and Carbon, C. C. Neural and genetic foundations of face recognition and prosopagnosia. *Journal of Neuropsychology 2*, 1 (2008), 79–97.

[56] Halperin, Y., Buchs, G., Madienbaum, S., Amenou, M., and Amedi, A. Social Sensing: a Wi-Fi based Social Sense for Perceiving the Surrounding People. In *Augmented Human Interaction Confernece* (Geneva, Switzerland, 2016), ACm New York, pp. 42–43.

[57] Halskov, K., and Hansen, N. B. The diversity of participatory design research practice at pdc 2002–2012. *International Journal of Human-Computer Studies 74* (2015), 81–92.

[58] Headway - the brain injury association. Types of brain injury. https://www.headway.org.uk/about-brain-injury/individuals/types-of-brain-injury/.

[59] Hébert, M. *Text-Dependent Speaker Recognition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 743–762.

[60] Hern, A. Apple halts practice of contractors listening in to users on siri. *The Guardian* (Aug 2019).

[61] HM Revenue & Customs. Voice id showcases latest digital development for hmrc customers. `https://www.gov.uk/government/news/voice-id-showcases-latest-digital-development-for-hmrc-customers`, Jan 2017.

[62] Hosokawa, S. The walkman effect. *Popular Music 4* (1984), 165–180.

[63] Howard, T. Journey mapping: A brief overview. *Communication Design Quarterly Review 2*, 3 (may 2014), 10–13.

[64] iFixit. Apple watch series 2 teardown. `https://www.ifixit.com/Teardown/Apple+Watch+Series+2+Teardown/67385`, Retrived 6 September 2019.

[65] Ignatov, A., Timofte, R., Chou, W., Wang, K., Wu, M., Hartley, T., and Van Gool, L. Ai benchmark: Running deep neural networks on android smartphones. In *The European Conference on Computer Vision (ECCV) Workshops* (September 2018).

[66] Ijaz, S., Davies, P., Williams, C. J., Kessler, D., Lewis, G., and Wiles, N. Psychological therapies for treatment-resistant depression in adults. *Cochrane database of systematic reviews*, 5 (2018).

[67] Information Commissioner's Office. Right to be informed. `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed/`.

[68] Information Commissioner's Office. What is personal data? `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/`.

[69] Institute Of Amateur Cinematographers. To film or not to film. https://www.theiac.org.uk/resourcesnew/filming-in-public/filming-in-public.html.

[70] Institute of Electrical and Electronics Engineers. Ieee recommended practice for speech quality measurements. *IEEE transactions on audio and electroacoustics 17*, 3 (1969), 225–246.

[71] J. -. Bonastre and F. Wils and S. Meignier. Alize, a free toolkit for speaker recognition. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* (2005), vol. 1, IEEE, pp. I/737–I/740 Vol. 1.

[72] Jain, A., Hong, L., and Pankanti, S. Biometric identification. *Commun. ACM 43*, 2 (Feb. 2000), 90–98.

[73] Jain, A. K., Nandakumar, K., and Ross, A. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters 79* (2016), 80 – 105.

[74] Jeske, D., and Santuzzi, A. M. Monitoring what and how: psychological implications of electronic performance monitoring. *New Technology, Work and Employment*, 1 (2015), 62–78.

[75] Jingu Heo, Kong, S. G., Abidi, B. R., and Abidi, M. A. Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. In *2004 Conference on Computer Vision and Pattern Recognition Workshop* (June 2004), pp. 122–122.

[76] Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience 17*, 11 (1997), 4302–4311.

[77] Kaplan, K. When and how to create customer journey maps. https://www.nngroup.com/articles/customer-journey-mapping/, Jul 2016.

[78] Kelion, L. Amazon alexa: Luxembourg watchdog in discussions about recordings. *BBC News* (Aug 2019).

[79] Kelly, H. Google glass users fight privacy fears. *CNN* (Dec 2013).

[80] Kensing, F., and Blomberg, J. Participatory design: Issues and concerns. *Computer Supported Cooperative Work (CSCW) 7*, 3 (Sep 1998), 167–185.

[81] Kleinberg, S. 5 ways voice assistance is shaping consumer behavior. *Think With Google* (2018).

[82] Klin, A., Sparrow, S. S., de Bildt, A., Cicchetti, D. V., Cohen, D. J., and Volkmar, F. R. A normed study of face recognition in autism and related disorders. *Journal of Autism and Developmental Disorders 29*, 6 (Dec 1999), 499–508.

[83] Koelle, M., Kranz, M., and Möller, A. Don't look at me that way!: Understanding user attitudes towards data glasses usage. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2015), MobileHCI '15, ACM, pp. 362–372.

[84] Kothari, J. Glass enterprise edition 2: faster and more helpful. https://www.blog.google/products/hardware/glass-enterprise-edition-2/, May 2019.

[85] Kotti, M., Moschou, V., and Kotropoulos, C. Speaker segmentation and clustering. *Signal processing 88*, 5 (2008), 1091–1124.

[86] Lakdawala, N., Fontanella, D., and Grant-Kels, J. M. Ethical considerations in dermatologic photography. *Clinics in Dermatology 30*, 5 (2012), 486 – 491. Ethics in Dermatology: Part II.

[87] Lawrence, K., Kuntsi, J., Coleman, M., Campbell, R., and Skuse, D. Face and emotion recognition deficits in turner syndrome: a possible role for x-linked genes in amygdala development. *Neuropsychology 17*, 1 (203), 39.

[88] legislation.gov.uk. The telecommunications (lawful business practice) (interception of communications) regulations 2000) no 2699 regulation 3. http://www.legislation.gov.uk/uksi/2000/2699/regulation/3.

[89] Lei, Y., Scheffer, N., Ferrer, L., and McLaren, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2014), pp. 1695–1699.

[90] LibriVox. Librivox: Free public domain audiobooks. https://librivox.org, Retrived 22 August 2019.

[91] LibriVox. Volunteer for librivox. https://librivox.org/pages/volunteer-for-librivox/, Retrived 22 August 2019.

[92] Lindsay, S. Basic Sketching Pratical Session, Feb 2019.

[93] Lindsay, S., Brittain, K., Jackson, D., Ladha, C., Ladha, K., and Olivier, P. Empathy, participatory design and people with dementia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 521–530.

[94] Lishka, K. Google-Debatte: Datenschützer kritisieren W-Lan-Kartografie[German: privacy advocates criticize W-Lan cartography [German: Google debate privacy advocates criticize W-FI cartography], April 2010.

[95] Luckerson, V. Google will stop selling glass next week. https://time.com/3669927/google-glass-explorer-program-ends/, January 5 2015.

[96] Mandal, B., Chia, S.-C., Li, L., Chandrasekhar, V., Tan, C., and Lim, J.-H. A wearable face recognition system on google glass for assisting social interactions. In *Computer Vision - ACCV 2014 Workshops* (Cham, 2015), C. V. Jawahar and S. Shan, Eds., Springer International Publishing, pp. 419–433.

[97] Markel, J., and Davis, S. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Transactions on Acoustics, Speech, and Signal Processing 27*, 1 (February 1979), 74–82.

[98] Marshall, G. Google glass: say goodbye to your privacy. *TechRadar* (Mar 2013).

[99] Marx, G. T. Ethics for the new surveillance. *The Information Society*, 3 (1998), 171–185.

[100] Maslow, A. H. A theory of human motivation. *Psychological review 50*, 4 (1943), 370.

[101] McCarthy, G. M., Rodriguez Ramírez, E. R., and Robinson, B. J. Participatory design to address stigma with adolescents with type 1 diabetes. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (New York, NY, USA, 2017), DIS '17, ACM, pp. 83–94.

[102] McConnell, F. Youtube is 10 years old: the evolution of online video. *The Guardian* (February 2015).

[103] McNaney, R., Vines, J., Roggen, D., Balaam, M., Zhang, P., Poliakov, I., and Olivier, P. Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 2551–2554.

[104] Microsoft Azure. Speaker recognition preview. https://azure.microsoft.com/en-gb/services/cognitive-services/speaker-recognition/, Retrived 22 August 2019.

[105] Moraiti, A., Vanden Abeele, V., Vanroye, E., and Geurts, L. Empowering occupational therapists with a diy-toolkit for smart soft objects. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction* (New York, NY, USA, 2015), TEI '15, ACM, pp. 387–394.

[106] Mozila. Common voice dataset. `https://www.kaggle.com/mozillaorg/common-voice`, 2017.

[107] Mozila. Mozila common voice. https://voice.mozilla.org/en, 2017.

[108] Munteanu, C., Baecker, R., Penn, G., Toms, E., and James, D. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2006), CHI '06, ACM, ACM, pp. 493–502.

[109] National Health Service. Prosopagnosia (face blindness). https://www.nhs.uk/conditions/face-blindness/, May 2016.

[110] National Health Service. Severe head injury, November 2018.

[111] National Health Service. Fake and harmful autism 'treatments'. https://www.nhs.uk/conditions/autism/autism-and-everyday-life/fake-and-harmful-treatments/, 2019.

[112] National Health Service. Overview osteoarthritis. https://www.nhs.uk/conditions/osteoarthritis/, August 2019.

[113] National Health Service. Overview turner syndrome. https://www.nhs.uk/conditions/turner-syndrome/, 2019.

[114] Nebeker, C., Linares-Orozco, R., and Crist, K. A multi-case study of research using mobile imaging, sensing and tracking technologies to objectively measure behavior: Ethical issues and insights to guide responsible research practice. *Journal of Research Administration 46*, 1 (2015), 118–137.

[115] Newell, A., Carmichael, A., Morgan, M., and Dickinson, A. The use of theatre in requirements gathering and usability studies. *Interacting with Computers 18*, 5 (08 2006), 996–1011.

[116] Newzoo. Top countriesmarkets by smartphone penetration & users. https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/.

[117] Norman, D., and Tognazzini, B. How Apple Is Giving Design A Bad Name. *FastCompany* (October 2015).

[118] Parette, P., and Scherer, M. Assistive technology use and stigma. *Education and Training in Developmental Disabilities 39*, 3 (2004), 217–226.

[119] Pascolini, D., and Mariotti, S. P. Global estimates of visual impairment: 2010. *British Journal of Ophthalmology 96*, 5 (2012), pp 614–618.

[120] Peretz, T. Mastering the new generation of gradient boosting. *Towards Data Science* (October 2018).

[121] Perez, S. Siri usage and engagement dropped since last year, as alexa and cortana grew. *Tech Crunch* (2017).

[122] Pollack, I., Pickett, J. M., and Sumby, W. H. On the identification of speakers by voice. *the Journal of the Acoustical Society of America 26*, 3 (1954), 403–406.

[123] Polster, M. R., and Rapcsak, S. Z. Representations in learning new faces: Evidence from prosopagnosia. *Journal of the International Neuropsychological Society 2*, 03 (may 1996), 240.

[124] Profita, H., Albaghli, R., Findlater, L., Jaeger, P., and Kane, S. K. The at effect: How disability affects the perceived social acceptability of head-mounted display use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), no. 12 in CHI '16, ACM, pp. 4884–4895.

[125] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 6638–6648.

[126] Prospecro, M. Vuzix Blade Review: These $1,000 AR Glasses Are Fun But Frustrating. https://www.tomsguide.com/us/vuzix-blade,review-6065.html, Febuary 2019.

[127] Ratha, N. K., Connell, J. H., and Bolle, R. M. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal 40*, 3 (2001), 614–634.

[128] Reynolds, D. A. An overview of automatic speaker recognition technology. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (May 2002), vol. 4, pp. IV–4072–IV–4075.

[129] Richardson, F., Reynolds, D., and Dehak, N. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters 22*, 10 (Oct 2015), 1671–1675.

[130] Roudier, M., Marcie, P., Grancher, A.-S., Tzortzis, C., Starkstein, S., and Boller, F. Discrimination of facial identity and of emotions in alzheimer's disease. *Journal of the Neurological Sciences 154*, 2 (1998), 151 – 158.

[131] Samsung Electronics America. Use a mobile hotspot on your galaxy phone. https://www.samsung.com/us/support/answer/ANS00079036/, August 2019.

[132] Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., and Pollick, F. E. Vision in autism spectrum disorders. *Vision Research 49*, 22 (2009), 2705 – 2739.

[133] Singhal, S., Neustaedter, C., Schiphorst, T., Tang, A., Patra, A., and Pan, R. Are Being Watched: Bystanders' Perspective on the Use of Camera Devices in Public Spaces. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16* (New York, NY, US, 2016), no. 7 in CHI EA '16, ACM, pp. 3197–3203.

[134] Siri Team. Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant - Apple. *Apple Machine Learning Journal* (2017).

[135] Siri Team. Personalized Hey Siri. *Apple Machine Learning Journal* (2018).

[136] Spam Laws. War Driving Attack. https://www.spamlaws.com/war-driving-attack.html.

[137] Stals, S., Smyth, M., and Mival, O. Exploring People's Emotional Bond with Places in the City. In *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems - DIS '17 Companion* (2017).

[138] Stark, L., Whittaker, S., and Hirschberg, J. Asr satisficing: the effects of asr accuracy on speech retrieval. In *Sixth International Conference on Spoken Language Processing* (2000), vol. 3, pp. pp 1069–1072.

[139] The Hamburg Commissioner for Data Protection and Freedom of Information. PRESS RELEASE Speech assistance systems put to the test - Data protection authority opens administrative proceedings against Google. https://datenschutz-hamburg.de/assets/pdf/2019-08-01_press-release-Google_Assistant.pdf, August 2019.

[140] The National Institute for Health and Care Excellence. Search Results for "prosopagnosia". `https://www.nice.org.uk/Search?q={%}22face+recognition{%}22`, 2018.

[141] Tigwell, G. W., Menzies, R., and Flatla, D. R. Designing for Situational Visual Impairments. In *DIS 2018* (Hong Kong, 2018), Association for Computing Machinery, pp. 387–399.

[142] Titze, I., and S. Winholtz, W. The effect of microphone type and placement on voice perturbation measurements. *Journal of speech and hearing research 36* (12 1993), 1177–1190.

[143] Trewin, S. Automating Accessibility: The Dynamic Keyboard. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, New York, USA, 2004), no. 8 in Assets '04, ACM, pp. pp 71–78.

[144] United States District Court Northern District Of California San Jose Division. FUMIKO LOPEZ, and FUMIKO LOPEZ, as guardian of A.L., a minor, individually and on behalf of all others similarly situated V APPLE INC., a Delaware corporation. https://www.classaction.org/media/lopez-et-al-v-apple-inc.pdf, August 2019. Accessed on https://www.classaction.org/media/lopez-et-al-v-apple-inc.pdf.

[145] U.S Food and Drug Administration. Danger: Don't drink miracle mineral solution or similar products. https://www.fda.gov/consumers/consumer-updates/danger-dont-drink-miracle-mineral-solution-or-similar-products, August 2019.

[146] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio, 2016.

[147] Vines, J., Clarke, R., Wright, P., McCarthy, J., and Olivier, P. Configuring participation: on how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI '13, ACM, ACM, pp. 429–438.

[148] Wang, H., Bao, X., Roy Choudhury, R., and Nelakuditi, S. Visually Fingerprinting Humans without Face Recognition. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '15* (New York, NY, USA, 2015), no. 345–358 in MobiSys '15, ACM, ACM.

[149] Warman, M. Google glass: We'll all need etiquette lessons. *The Telegraph* (Apr 2013).

[150] Watson, J., and Nesdale, D. Rejection sensitivity, social withdrawal, and loneliness in young adults. *Journal of Applied Social Psychology 42*, 8 (2012), 1984–2005.

[151] Williams, E. Creepy wireless stalking made easy. https://hackaday.com/2016/12/04/creepy-wireless-stalking-made-easy/more-232845, December 2016.

[152] Wobbrock, J. O. The Future of Mobile Device Research in HCI. In *2006 workshop proceedings: what is the next generation of human-computer interaction, 131-134.* (2006), CHI 2006 workshop proceedings: what is the next generation of human-computer interaction, pp. pp. 131–134.

[153] www.legislation.gov.uk. Human Rights Act 1998 Article 8. https://www.legislation.gov.uk/ukpga/1998/42/schedule/1/part/I/chapter/7, 1998.

[154] Xi Wang, Xi Zhao, Prakash, V., Weidong Shi, and Gnawali, O. Computerized-eyewear based face recognition system for improving social lives of prosopagnosics. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops* (ICST, Brussels, Belgium, Belgium, 2013), no. 4 in PervasiveHealth '13, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 77–80.

[155] Yale Vision Group. Yale face database. http://vision.ucsd.edu/content/yale-face-database, 1997.

[156] Yandex. Using the overfitting detector. https://catboost.ai/docs/features/overfitting-detector-desc.html.

[157] Yardley, L., McDermott, L., Pisarski, S., Duchaine, B., and Nakayama, K. Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of Psychosomatic Research 65*, 5 (nov 2008), 445–451.

[158] Yie, T. Understanding random forest. *Towards Data Science* (June 2019).

[159] Zunger, J. Computer science faces an ethics crisis. the cambridge analytica scandal proves it. *Boston Globe 22* (2018).

Appendices

A Paper submission made to the Workhop CHI 2019 Workshop on Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction.

Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury

Osian Smith
869024@swansea.ac.uk
Swansea University
Swansea, United Kingdom

Stephen Lindsay
s.c.lindsay@swansea.ac.uk
Swansea University
Swansea, United Kingdom

ABSTRACT

In this document, we discuss our work into a speaker recognition to support people with prosopagnosia and the limitations of alerting the user of whom they are in discussion with. We will discuss how current research into Situationally Induced Impairments Disabilities (SIIDs) can assist people with disabilities and vice versa and how our work can support people who may find themselves in a situation where they are impaired with facial recognition.

CCS CONCEPTS

• **Human-centered computing** → *Accessibility technologies.*

Glasgow '19, May 04–09, 2018, Glasgow, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Situationally-induced imparements, cognitive brain injury, prosopagnosia

ACM Reference Format:
Osian Smith and Stephen Lindsay. 2019. Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury. In *Glasgow '19: ACM Computer Human Interaction, May 04–9, 2019, Glasgow, UK*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

INTRODUCTION

Situationally-induced impairments and disabilities (SIIDs) are an expanding issue in ubiquitous computing. While our research area is supporting patients with Cognitive Brain Injury (CBI), the work of SIIDs supports our research interests by allowing existing devices to be more accessible for people with impairment. Current estimates of the abandonment rate Adaptive Tools (AT) is as high as 75%, with many of people using AT finding that they give themselves stigma of being unwell [13]. By using standard technologies, we can remove the stigma of the patient using a medical device, which in turn can increase the life quality of the patient.
We chosen to focus on supporting patients with prosopagnosia by speaker recognition on a smartwatch as we believe that by using a smartwatch, the patient should not feel any stigma as smartwatches are commonplace. We also feel that speaker recongition can support people with SIIDs.

PROSOPAGNOSIA

Prosopagnosia, commonly known as facial blindness, is the condition where patients cannot recognise faces, people or voices [11]. Prosopagnosia is estimated to affect around 2.5% of the population, approximately the equivalent number of people who have dyslexia [8]. It can be cognitive at birth or acquired through CBI and although treatment is available, it is widely accepted not all will be successful with many patients with no suitable treatment [4, 6, 14].

Prosopagnosia is usually present in an individual at birth although can be acquired. There is evidence that suggests that prosopagnosia can be a trait of an individual with Autism Spectrum Disorder, although this link cannot be considered established due to little evidence supporting them. There is also debate whether prosopagnosia should be kept as its condition when ASD cannot be used to explain face blindness [11, 22]. People who suffered brain trauma can also develop prosopagnosia due to damage to the segment of the brain that processes facial data. Prosopagnosia can also be genetic with evidence showing that the ability to recognise faces is an evolutionary trait as research shows that the Macaque Monkey has neurons specifically tuned to facial recognition [10, 11, 18].
There is no treatment for prosopagnosia that has demonstrated an improvement to all patients with many patients demonstrating no improvement to to treatment[2, 14] . Treatments for prosopagnosia

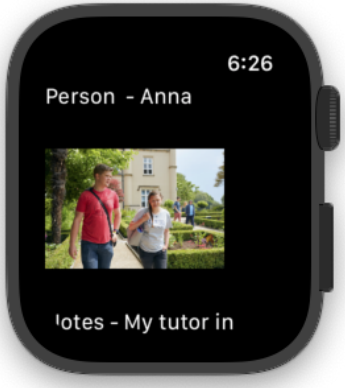


Figure 1: A screenshot of our prototype application displaying information on a speaker after identifying an individual. On this screen we display the person's name, a photo and notes on them

focuses on training of processing facial features [1, 3, 9, 17]. This adds a cognitive load to the patient which may be unstable for people with ABI due to their decrease cognitive ability [2, 6]

OUR CURRENT WORK

Our current research is investigating supporting patients with prosopagnosia to identify the person that the patient is in conversation. We are currently developing on a system that uses a microphone on a smartwatch to then listen to speakers, before sending the audio to the smartphone for further processing. The smartwatch then uses a one-shot learning neural network to identify the person in conversation with on their smartphone before alerting them on the smartwatch.
We are currently running participatory design workshops with patients with prosopagnosia to understand how they would want such a device to work. Users did not want to have a device that showed that they had a condition, such as a microphone on their neck but would prefer a smartwatch like the Apple Watch, a device that many people already wear. However, many of the participants do have impairments that may make it difficult to use the smartwatch.
As a result, we are using these participatory design workshops to understand the impairments of using a device. For example, we are looking into how the user can interact with their smartwatch to activate the microphone without making it evident to the person who they're in converse. This research may also benefit the SIIDs community, by understanding how to interact with a smartwatch without people knowing - such as sending a message to a family member while the user is in a meeting.

We are also interested in alerting the user that we have identified the speaker. Our current working prototype as seen in Figure 1 uses visual feedback of showing the user who the speaker is from a photo and a quick description. However, we are investigating audio feedback, where the user has an earphone in their ear which an automatic voice speaks to them stating who it is, or do we have the smartwatch give off a unique chime for every user. Could we also use haptic feedback, where we use the watch's vibration motor to give unique vibration to every person in the speaker?
There are limitations to each of these modalities. The chime and haptic feedback add a cognitive load onto the user, which may not be suitable for patients with CBI due to their diminished cognitive ability [2, 6]. However visual feedback may be rude because the user has to look at their smartwatch which may give the impression that the user is ignorant. Our headphone implementation could also be considered being rude as it may seem that the user is listening to music when in conversation.
Many of these concerns are also valid to the general population with SIIDs. How can users interact with their devices without being rude [7]? Is the user in a suitable situation where they can use voice assistance? Do these technologies cause concern for user and bystander privacy? Google Glass was infamous for the lack of bystander privacy with many of its users being called "glassholes" [5].

Furthermore, research by McNaney et al. found that during his study with Google Glass, many patients felt uncomfortable wearing this technology due to privacy concerns [12].

Situationally Induced Impairments Disabilities

Many patients who require AT can also benefit from the research into SIIDs. For example, patients with Parkinson's disease may struggle to interact with a smartphone device due to their tremors [15]. Many users who don't have Parkinson's may find themselves in a situation that they cannot use a device for similar reasons of patients with Parkinson's, such as being in a moving vehicle. As a result, could an interface designed for someone moving in a car also be transferred to someone with Parkinson's?

Research by Trewin [20] has previously researched dynamic keyboards and how people with motor disabilities and SIIDS interact with their phone. Trewin noted that there were technologies such as sticky keys to help people use keyboards, however Trewin noted that their were difficulties with adjusting the delays which many users may not want to constantly change [20]. While people with motor disabilities would likely configure the settings correctly, there may need to be more of an adaptive approach to people with SIIDS. We do have to consider that Trewin works was published in 2004 before touchscreen smartphones became commonplace.

Research by Nicolau et al. [15] found that people with motor disabilities while can tap on keyboards, although their error rate was higher than participants with no impairments [15]. While technologies such as predictive text and dictation may help, we did not find any literature that investigate this area.

Wobbrock [21] believes that once SIIDs are better understood, research should focus on comparing SIIDs to medical conditions such as comparing people with poor eyesight to people with poor vision. From this we can gather relationships between conditions and have better design decisions that support a wider range of people, which would also reduce stimga in these technologies.

Research into the design process of applications by Tigwell et al. [19] found that many clients were resistant to accessibility as they felt that the designers were overbearing. Tigwell et al. stated that while they can use well-known guidelines such as they Apple Human Interface Guidelines [1] and Google Material Design [2] , they may become problematic due to the amount of content available. There are also inconsistencies with these guidelines with Apple's low contrasting fonts being challenging for people with typical vision to read. [16] [19]

Tigwell et al. [19] made the argument that this approach to accessibility could be made to support people with SIIDs, although we believe that this argument of supporting SIIDS could also support users with impairments.

Our research on supporting patients with prosopagnosia can also be deployed to support patients with SIIDs. Many people without prosopagnosia find remembering names challenging, especially

[1]https://developer.apple.com/design/human-interface-guidelines/

[2]https://material.io/design/

in certain situations, for example, a lecture at a university. A lecture may be responsible for 300+ students enrolled on a module and may need to discuss an important topic with one student; however, they cannot remember every student on the course. By using the speaker recognition system that we developed, could the lecture put notes on that student and be alerted when they are talking to the student to discuss the important topic?

CONCLUSION

In this document, we have discussed the high abandonment rate of Assistive Technologies with research pointing this is a result of stigma. We have also outlined our work that supports patients with prosopagnosia by listening to the person they are in conversation with and stating who they are talking to. We also highlighted how research into SIIDs can help people who require AT and vice versa and that our work could also support people who do not suffer from prosopagnosia.

REFERENCES

[1] E S BEYN and G R KNYAZEVA. 1962. The problem of prosopagnosia. *Journal of neurology, neurosurgery, and psychiatry* 25, 2 (may 1962), 154–158. https://doi.org/10.1136/JNNP.25.2.154

[2] Richard A. Bryant. 2008. Disentangling Mild Traumatic Brain Injury and Stress Reactions. *New England Journal of Medicine* 358, 5 (2008), 525–527. https://doi.org/10.1056/NEJMe078235 PMID: 18234757.

[3] Thomas Busigny, Eugene Mayer, and Bruno Rossion. 2013. Prosopagnosia. In *The Behavioral and Cognitive Neurology of Stroke*, Olivier Godefroy (Ed.). Cambridge University Press, Cambridge, 231–246. https://doi.org/10.1017/CBO9781139058988.020

[4] Sherryse L Corrow, Kirsten A Dalrymple, and Jason Js Barton. 2016. Prosopagnosia: current perspectives. *Eye and brain* 8 (2016), 165–175. https://doi.org/10.2147/EB.S92838

[5] Schuster Dana. 2014. The revolt against Google "Glassholes", url = https://nypost.com/2014/07/14/is-google-glass-cool-or-just-plain-creepy/,. *New York Post* (July 2014).

[6] Joseph M DeGutis, Christopher Chiu, Mallory E Grosso, and Sarah Cohan. 2014. Face processing improvements in prosopagnosia: successes and failures over the last 50 years. *Frontiers in Human Neuroscience* (2014). https://doi.org/10.3389/fnhum.2014.00561

[7] Thomas Eddie, Juan Ye, and Graeme Stevenson. 2015. Are Our Mobile Phones Driving Us Apart? Divert Attention from Mobile Phones Back to Physical Conversation! (2015), 1082–1087. https://doi.org/10.1145/2786567.2794331

[8] A Feola, V Marino, A Masullo, M Trabucco Aurilio, and LT Marsella. 2015. The protection of individuals affected with Specific Learning Disorders in the Italian Legislation. *La Clinica terapeutica* 166, 3 (2015). https://doi.org/10.7417/ct.2015.1851

[9] R Francis, M Jane Riddoch, and Glyn W Humphreys. 2002. 'Who's that girl?' Prosopagnosia, person-based semantic disorder, and the reacquisition of face identification ability. *Neuropsychological Rehabilitation* 12, 1 (jan 2002), 1–26. https://doi.org/10.1080/09602010143000158

[10] Guido Gainotti and Camillo Marra. 2011. Differential Contribution of Right and Left Temporo-Occipital and Anterior Temporal Lesions to Face Recognition Disorders. (2011), 55. https://doi.org/10.3389/fnhum.2011.00055

[11] Thomas Grüter, Martina Grüter, and Claus Christian Carbon. 2008. Neural and genetic foundations of face recognition and prosopagnosia. https://doi.org/10.1348/174866407X231001

[12] Roisin McNaney, John Vines, Daniel Roggen, Madeline Balaam, Pengfei Zhang, Ivan Poliakov, and Patrick Olivier. 2014. Exploring the Acceptability of Google Glass As an Everyday Assistive Device for People with Parkinson's. (2014), 2551–2554. https://doi.org/10.1145/2556288.2557092

[13] Argyro Moraiti, Vero Vanden Abeele, Erwin Vanroye, and Luc Geurts. 2015. Empowering Occupational Therapists with a DIY-toolkit for Smart Soft Objects. (2015). https://doi.org/10.1145/2677199.2680598

[14] NHS. 2016. Prosopagnosia (face blindness) - NHS. https://www.nhs.uk/conditions/face-blindness/

[15] Hugo Nicolau and Joaquim Jorge. 2012. *Elderly Text-entry Performance on Touchscreens*. ACM, New York, NY, USA. 127–134 pages. https://doi.org/10.1145/2384916.2384939

[16] Don Norman and Bruce Tognazzini. 2015. How Apple Is Giving Design A Bad Name. https://www.fastcompany.com/3053406/how-apple-is-giving-design-a-bad-name

[17] Michael R Polster and Steven Z Rapcsak. 1996. Representations in learning new faces: Evidence from prosopagnosia. *Journal of the International Neuropsychological Society* 2, 03 (may 1996), 240. https://doi.org/10.1017/S1355617700001181

[18] David R Rubinow and Robert M Post. 1992. Impaired recognition of affect in facial expression in depressed patients. *Biological Psychiatry* 31, 9 (may 1992), 947–953. https://doi.org/10.1016/0006-3223(92)90120-O

[19] Garreth W. Tigwell, Rachel Menzies, and David R. Flatla. 2018. Designing for Situational Visual Impairments. In *DIS 2018*. Association for Computing Machinery, Hong Kong, 387–399. https://doi.org/10.1145/3196709.3196760

[20] Shari Trewin. 2004. *Automating Accessibility: The Dynamic Keyboard*. Technical Report. http://www.research.ibm.com/KeyboardOptimizer

[21] Jacob O Wobbrock. 2006. *The Future of Mobile Device Research in HCI*. Technical Report. 131–134 pages. http://www.fossil.com

[22] Lucy Yardley, Lisa McDermott, Stephanie Pisarski, Brad Duchaine, and Ken Nakayama. 2008. Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of Psychosomatic Research* 65, 5 (nov 2008), 445–451. https://doi.org/10.1016/J.JPSYCHORES.2008.03.013

B Presintation made to the Workhop CHI 2019 Workshop on Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction.

Looking At Situationally-Induced Impairments And Disabilities (SIIDs) With People With Cognitive Brain Injury

Osian Smith
Swansea University

Swansea University
Prifysgol Abertawe

About me

- Currently:
  MRES in Computing and Future Interaction Technology

- Starting a PhD in October - Making with Meaning - researching patients with Acquired Brain Injury (ABI) to co-create Assistive Technology (AT) in Swansea University

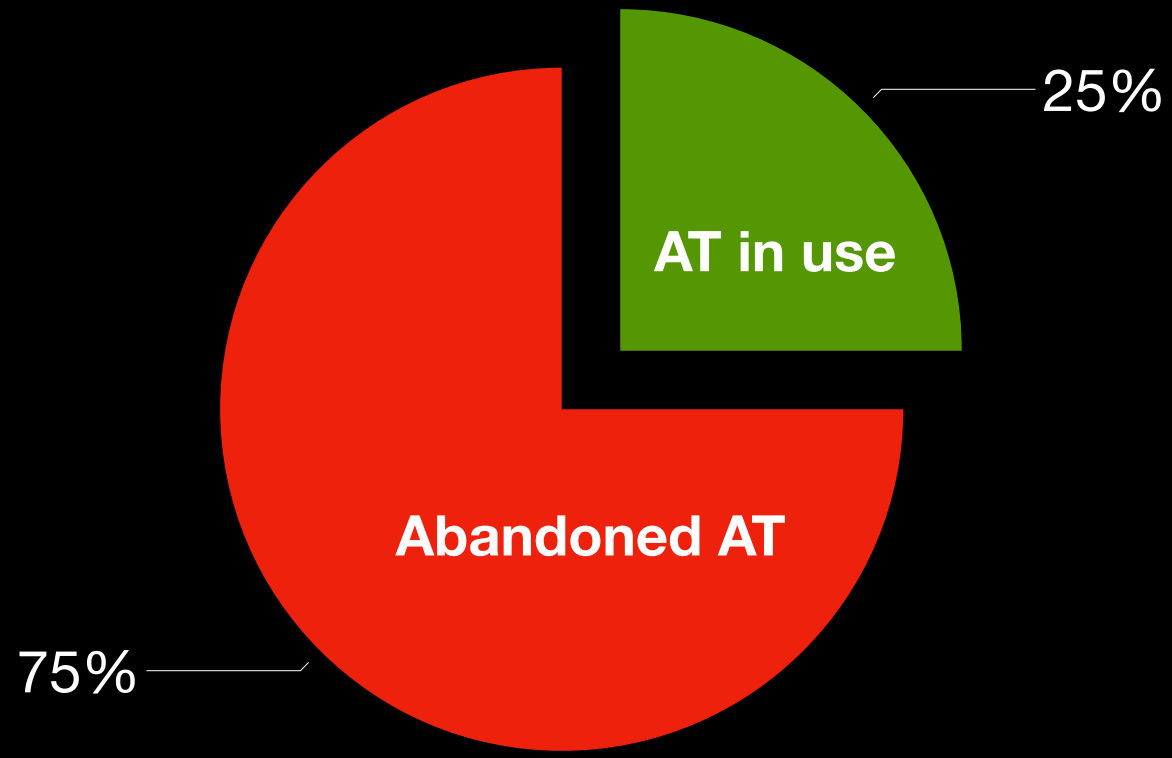
Current Research

- How can we use speaker recognition to support patients with prosopagnosia
- Smartwatch chosen due to be common place
- Currently running participatory design workshops with patients

Disability + SIIDs

- AT should introduce as little stigma as possible
- Can making AT common place also reduce stigma?
- Can we use AT to support SIIDS?

75% of all AT are abandoned



Argyro Moraiti, Vero Vanden Abeele, Erwin Vanroye, and Luc Geurts. 2015. Empowering Occupational Therapists with a DIY-toolkit for Smart Soft Objects. (2015). https://doi.org/10.1145/2677199.2680598

SOPUR


9:67
Friday, 09 December
MTN-SA

How can AT help people without disabilities?

- Through SIIDs
- However there are challenges:
  - Continuous
  - Reliability
- However, we can adapt AT for SIIDS and vise versa


$= \sum_{n=0}^{\infty} \int_0^t \frac{(-1)^n x^{2n}}{n!} dx = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{1}{(2n+1)} x^{2n+1} \Big|_0^t$ (10+)
$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(2n+1)} t^{2n+1}$ numerisch berechenbar!
Es gilt: $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$ (Laplace 1782)

Thank you

C Paper submission made to CHI 2020 = Pwy?: Designing a Discrete speaker-recognition App for Conversational Support on Smartwatches

This paper is currently in review at time of print

Pwy?: Designing a Discrete speaker-recognition App for Conversational Support on Smartwatches

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

We investigate conversational speaker-recognition systems, inferring identity from any spoken phrase, to support people who find recalling names in conversation difficult by discretely providing them with speakers' names and other relevant personal information via a smartwatch. We ran participatory design sessions with expert designers, people who self-identify as finding socialising difficult and people diagnosed with Traumatic Brain Injury. Sessions addressed social attitudes, privacy and adding new people to the system for future recognition. We discuss significant differences the process uncovers between groups. We train a speaker-recognition algorithm based on spectrogram feature extraction and classification. However, the implementation had delays of two to eight seconds between the start of conversation and recognition of speakers. Consequently, we ran studies to understand how delays in alerting users to speaker identity impacted on the perceived usefulness of the application.

Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; this section is required.

Mobile Devices: Phones/Tablets ; Accessibility ; Health - Wellbeing ; Individuals with Disabilities & Assistive Technologies

CCS Concepts

•**Human-centered computing** → **Accessibility technologies; Human computer interaction (HCI);** *Haptic devices;* User studies; Please use the 2012 Classifiers and see this link to embed them in the text: https://dl.acm.org/ccs/ccs_flat.cfm

INTRODUCTION

The ability to recognise faces and link them to names is fundamental to functioning within society with evidence suggesting this trait is evolutionary [20]. The ability to link faces and names allows us to determine our relationship to whoever we are in discussion with, know where a conversational partner is looking or infer a stranger's gender, age, health, and mood [16, 54]. However, several health conditions inhibit the ability to recognise faces including brain injury, Alzheimer's disease, Autism Spectrum Disorders (ASD) [8] and Prosopagnosia which can prevent facial recognition entirely. The current work of the CHI community to offer support for people living with these conditions has focused on supporting people living with Prosopagnosia or similar conditions using facial recognition through video capture [34, 51]. However, video-based recognition relies on specific circumstances so can be inhibited by, for example, poor lighting, a long or short distance to the face or partial or full obfuscation of the face. Furthermore, the use of cameras introduce serious, complex privacy concerns for both users and those being observed [31, 40]. These constraints limit the practical support video-based recognition can offer in the real world.

In this work we explore the design of a smartwatch wearable to discretely tell you who you are talking to in order to support the wearer in conversation. The smartwatch records and analyses voices in the environment and presents best guesses as to the identity of the person you are in conversation with based off of machine learning ran entirely on a companion smartphone. This work is timely as today's smartphones are starting to include dedicated neural networking processors that allow us to run machine learning frameworks such as TensorFlow Lite[1] [TFLite] and CoreML [2], which allow us to run new machine learning algorithms locally [26]. Video-based facial recognition is a computationally cheap process which explains its use to date in prototypes that recognise people in social situations [51]. However, using machine-learning frameworks such as TFLite and a smartphone neural network processor, we can execute novel machine learning algorithms such as speaker-recognition to identify the person that one is in conversation with in real time. Speaker-recognition has several advantages over facial recognition, such as improved perceptions of privacy intrusion and practical reductions in the data being stored for recognition's security risk coupled with a reduction in social stigma.

In this paper, we discuss our research by using speaker-recognition to aid social interaction. We developed Pwy[3] to

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

[1] https://www.tensorflow.org/lite

[2] https://developer.apple.com/documentation/coreml

[3] Welsh for "Who" as in "Pwy ydych chi?" or "Who are you?"

use in participatory design workshops and to run evaluation studies with. We analyse audio data captured from the Apple Watch to understand the quality of the audio from the watch. Our work also shows how HCI research, through participatory design workshops, can support choices made by the machine learning community and drive decisions on trade-offs between speed, features and privacy.

PREVIOUS WORK

Understanding speaker-recognition

Speaker-recognition is the sub-element of the wider area of voice recognition that answers the question of "who is speaking" [7]. This is different to the more common approach of speech recognition that focuses on turning what a person has said into machine-readable text or "what is said". Speaker-recognition is broken up into several subcategories, each with their features which separates them from other categories. While specific applications do utilise both speech recognition and speaker-recognition, such as voice assistants on smartphones like Siri, Google Now and Amazon Alexa [1, 19, 43, 44], these require two discreet algorithms in order to function.

Text-Dependent and Text-Independent speaker-recognition

We can categorise speaker-recognition as text-dependent speaker recognition or text-independent speaker-recognition. Text-dependent requires that the phrase used in training a model for voice recognition be used within inference [23]. Text-Dependant speaker-recognition is currently primarily used within the financial sector, with organisations such as the UK Government HMRC using speaker-recognition with the phrase "My Voice is my Password." [24]. Text-independent recognition allows for training and inferring phrases to be different with no effect on the accuracy of the system [18], allowing the use of this model in real-world conversations.

Closed-sets and Open Sets

We can further divide speaker-recognition into close-sets and open sets. Closed sets focus on identifying a speaker from a set of voices while an open set states whether the voice belongs to the set or not [18]. Closed sets may also be known as N-ary classification, and open sets may be known as binary classification [5]. Open sets algorithms are sometimes utilised to infer whether a personal assistant is being addressed by its owner, such as the use of "Hey Siri" [43] whereas close sets may utilise open sets to return an unknown speaker and to improve the computational efficiency of the algorithm.

Binary and N-ary Recognition

Speaker-recognition can be classed as binary recognition and N-ary recognition [5]. Binary recognition is only able to recognise whether one person is speaking or not (returning true or false) and is often used by virtual assistants to identify whether they are talking to the device owner or not [43]. N-ary recognition is able to recognise multiple speakers but requires more computational power compared to binary recognition.

Local and Remote Inference

We can divide speaker-recognition further into local and remote inference and training. Due to the computational power that is required to run networks for inference, and the size of the data-set that is needed for training requirement of any new data, it may be unsuitable to have the training or inferring on the device. Here the voice sample would be sent to another device such as a remote server for inference with the reply sent back to the device stating the speaker' name.
On-device, on the other hand, does all the training and inference on the device, which means that it can work independently. However, these models may not perform as well as off-device training and may not be able to retrain due to the computational complexity that is required for retraining.
There may also be instances where training will occur off device, but inference occurs on the device. For example, feature extraction may occur on an external computer, but the actual inference from those features may occur on the smartphone. Feature extraction is a computationally expensive algorithm; however, the inference is less computationally expensive but trained around voices specific to that device.

Applications of Recognition for Social Support

The work of the HCI community has focused on two ways to support people in social situations through recognising people around them: biometrics and alternative identifiers. Biometrics is the identification of individuals on their anatomical and behaviour characteristics such as fingerprints, facial patterns or characteristics of their speech [34, 51]. Alternative identifiers work by identifying things associated with a person such as digital devices they carry or items they wear. Examples of successful alternative identifies include clothing identification [49] and mobile phone signal logging [21]. Biometrics is the "automated recognition of individuals based on their anatomical and behavioural characteristics such as fingerprint, face, iris, and voice" [28]. Biometrics is made up of biometric identifiers that are the measurable characteristics of an individual, such as the distance between the eyes, ears and nose [27]. Biometrics recognition is the most commonly used approach within the HCI community for recognising people [34, 51]. Biometrics are also widely used within the security field as a form of authentication [42], with many phones containing a facial recognition system or fingerprint sensor [10]. This is because biometrics include the most accurate methods to infer who an individual is as they do not rely on a variable external factor like carrying a mobile phone. However, they are also the most intimate type of recognition requiring processing of personal data such as facial data or voice data [28].

Biometrics have been widely employed in smart-glasses to support people with facial blindness or similar conditions. Research by Wang et al. [51] explored using smart glasses with an augmented reality (AR) screen to highlight faces and overlay the name of the individual. A camera captured images continuously and used facial recognition to detect faces, that were analysed to see whether they were known. If known, the name and relation to the person was displayed in AR in the sight-line of the wearer. Using facial recognition, Wang et al. were able to achieve 99% percent accuracy for detection of faces and a 95% rate of recognition of known faces on the Yale data-set which consisted of 165 gray-scale images of 15 individuals [52] along with a proprietary database which we

could not access [51]. Mandal et al. [34] extended the work of Wang et al. by using Google Glass (GG) a headset available to consumers to run facial recognition and understand the limitations of on device inference, where inference occurs on GG, and off device inference, where inference occurred on a mobile phone. Running inference on GG, was only able to achieve less than half on device recognition when compared to off device inference highlighting that processing on wearables are not currently comparable to the performance of a smartphone as wearable technology generally contains, weaker processing and battery capability [34] .

Work by Wang et al. [49] develops walking signatures to infer an individual's identity by producing "temporary fingerprints." Temporary fingerprints allowed the system to recognise an individual, but the individuals could tell the application not to track them which would tell the server to ignore their fingerprint. Wang et al. used clothing signatures to support the recognition of an individual. The system infers a person by analysing footage for walking features such as whether the person is standing or walking, their step-duration and phase of step as the step along with the direction that they were walking [49].

However, smart glasses have been unpopular with members of the general public and participants. When Google launched GG in 2013, privacy quickly became a concern for the public with many news outlets covering the issue [2, 29, 35, 50]. There was no external way to tell if someone was being recorded which made many people worry about their privacy when a glass wearer was in the vicinity and, as a result of the privacy concerns, GG wearers were given diminutive nicknames such as "Glassholes" [9]. McNaney et al. [36] highlight that patients with Parkinson's disease who wore GG for a period of 5-days were concerned about their privacy. The users of the GG system also worried about the data that they were collecting and the potential for it to be turned against them as some participants thought that relatives might abuse the video linking features as a way of monitoring what they were doing.

Beyond biometrics, research by Halperin et al. has focused on using WiFi signals for identifying people. The system would listen to the WiFi address that phones transmitted as part of the WiFi hot-spotting standard 802.11. Halperin et al. generated an auditory mapping of the distance between the phones and gave feedback based on the location of the phone to the individual. They then taught participants to identify a phone from the tones that their device had emitted [21]. However, the act of tracking someone through wireless signals is considered to be part of the wider practice of wardriving, which, while not explicitly illegal, still might be challenged in several countries in the EU [33].

Wearable technology can also be used to identify someone through their device without having to wardrive. Bâce et al. [3] has looked at using gestures for sharing an identity with "HandshakeAR." Here both parties would wear a wearable such as a smartwatch and would detect when they do a predefined movement, such as a handshake or a high-five. When one wearer does the predefined gesture, it scans for the other device to see if they both made the same gesture and they both share and display a business card [3].

Not every individual wants to disclose their disability as they feel that it may give them a "social weight", where patients feel stigmatised for their conditions, which in turns drastically impacts the adoption and the use of devices [11, 40]. Many people who require Accessibility Tools [AT] may also not use it when needed, for example, some individuals who require a white cane due to vision impairment abandon it to avoid drawing attention [39, 40].

Research by Munteanu et al. evaluated the acceptance of accuracy on text to speech algorithms on transcripts generated from internet broadcast videos. Munteanu et al. found that transcripts of 75% Word Accuracy Rate (WAR) were acceptable, but a WAR of less than 55% was unacceptable [38]. Work by Stalk et al. further demonstrated that people who had access to high-quality (above 84% WAR) transcripts were less likely to abandon them and people with moderate to low transcripts (less than 70% WAR) are more likely to abandon transcripts [45]. These results demonstrate that there is a tipping point where people are willing to use technology that relies on around 75% and 84% WAR. We consider this to be the benchmark of any solution to support people to recognise people. We developed a speaker-recognition algorithm to understand how such an algorithm would work with audio data that came from the smartwatch. We used two machine learning algorithms, a feature extractor and an identifying algorithm to understand the audio output of such an algorithm. Each second of audio data contains 16,000 samples with sound parameters requiring an understanding of surrounding data to create a wave and to discriminate the differences between speakers sign waves, resulting in it being too complicated for humans to program with the low accuracy required.

THE PARTICIPATORY DESIGN OF PWY

The Pwy system is envisioned as being a primarily smartwatch based one. When in conversation with someone, the wearer will be alerted by a vibration on their wrist when the Pwy system has established who they are talking to and their name displayed on the screen as a discrete form of support. Pwy uses the smartwatch microphone to capture audio because it is relatively free to do so when compared to a phone in ones pocket, then relays this to the companion phone where machine learning established the identity of the person you are in conversation with and pushes a notification to this effect back to the smartwatch. This relatively simple use case gives rise to several questions though such as "how does the system get told the name associated with a voice?" and "what are peoples feelings about using this sort of device in conversation?" or "do they feel that it is rude to use?". To address these, we conduct a range of Participatory Design activities with several stakeholder groups.

Participatory Design is a group of design and research

practices that foreground the needs of users by actively including them in the design process to improve the design and to ultimately smooth the integration of novel technology into their daily life [22]. Participatory Design emerged in Scandinavian countries in the 1970's as computers were being introduced into the workplace and concerns arose surrounding the effects that these systems would have on workers. [30]. Participatory Design has become a popular method in the HCI community for developing technologies with users, allowing for sharing control of the development process of technology by treating the end-user as an expert with tacit knowledge of the lived experience of their health condition [48]. Participatory Design Workshops require different approaches depending on the participants, and understanding that not all activities are suitable for each set of participants.

We ran three Participatory Design workshops with independent design experts, people who self-identified as having difficulty socialising and people with Traumatic Brain Injury [TBI] diagnosis (see Table 1). During each session participants viewed three design theaters: 1) system working as intended 2) meeting a new person and 3) the system failing where the user called the bystander by the wrong name. Once this had taken place participants used scenario cards to understand how the application could be used by them. We also ran design crits with the expert designers and with participants who had difficulty socialising, however were unable to run it with participants with TBI due to the richness of data gathered in the other work leading into time constraints.

Session	Gender
Expert designers	3 males aged 20 to 35 currently working towards a PhD in the design field
People with difficulty socialising	4 males and 2 female aged 18 to 25 who are currently students and not receiving any formal support
People with TBI	4 males and 3 females who are outpatients with TBI however were attending a TBI support session

Table 1. Above states a basic outline demographic data of participants in each of our participatory design sessions

Findings of our Participatory Design Workshop

We developed the interface of Pwy to consist of a watch application, with the task of listening and notification and a phone application which was designed as voice management. The phone application consisted of three principle screens - a list of known voices by name, a list of new voices and a detailed screen which allowed the user to view, modify or add people to the application. On the Watch application, users can trigger a listing by pressing the "Tap to listen" button, view a recording of the application take place and then an alert of whom the speaker is and an option to listen again. These screens were labelled and given to participants to design critic.

In a series of participatory design workshops, participants consistently raised several themes such as failure and privacy. All three groups accepted that failure would occur, though, participants with TBI were, to our surprise, the more accepting of failure, saying that any assistance was better than no assistance. We understood that this was a result of people with TBI being more accustomed to having the support and struggling with faces already, unlike people with difficulty socialising who don't have access to support. This mean that expert designers and participants who lacked social confidence wanted a tree-like structure that could hone in on an identity. For example, the app could ask whether a person mentions their dog and if so then that must be Tom, else the app would look for another identifying detail like talking about their new house to ensure that they were using the correct name. However, participants with TBI were happy for a simpler solution just seeing a confidence score such as 85%percent chance this is John' to allow them to make their judgments.

Participants needs varied depending on their condition. Participants with TBI were more accepting of using the device compared to those with difficulty socialising. We hypothesise this is a result of people with TBI being more willing to receive help because they have already had assistance through rehabilitation, while people with social anxiety have not received any support and the act of receiving that support could, in their view, stigmatise them. Participants with TBI wanted to look at the watch as little as possible and did not want the distraction of the watch when not required. Participants with TBI did not want to seen as rude by constantly looking at the watch, as they felt that if they did have to explain

People who lack social confidence worried about the system adding to anxiety. Participants envisioned seeing a name displayed on the watch, but being unsure whether this was the correct name so would then avoid using the name. Participants with TBI felt that this could be overcome by using confidence scores stating how confident the system was in it's prediction. Privacy was a concern that all three sets of participants highlighted with specific concerns about bystanders privacy. Bystander privacy is the concern of privacy for people who are captured by the application, even though they may have no desire or be unwilling to interact with the user [15]. Concerns for bystander privacy were present in all sets of groups, and each group had a different approach to this privacy problem. The expert designers felt that the best approach was a social media of voices, where users could request access to peoples voices. This seems unlikely as the general public may be willing to share data for something seen as a relatively niche service, especially since the Facebook and Cambridge Analytica scandal [6]. Participants with TBI focused more on explaining that they were recording and explained the reasoning behind it; however, they felt that most people would be accommodating and that participants would rather talk about other topics. Unlike the findings of McNanney, none of the groups we worked with were concerned about their own privacy [34] The lack of self-privacy concerns was likely a result of being aware that a recording was taking place and that they could

stop if they wanted privacy, such as if the user was discussing a sensitive topic.

Furthermore, each set of participants had a different approach to discreetness and how they felt that they would accept the application. Participants with social difficulty showed less concern for bystander privacy. Participants wanted all audio recorded and processed without the knowledge of bystanders to be as discreet as possible. Nonetheless, when discussion turned to the use of the system in public, the participants were worried about the possibility that bystanders could catch them using the application and they quickly rejected it, finding the thought of being caught creepy. We hypothesise this is a result of participants not wanting to be seen using an accessibility device as well as feeling their own behaviour was rude. In contrast, participants with TBI explained their relative lack of concern stating that they were happy to use the watch and if someone saw them using it, they would explain what the watch was doing and say 'I have had a brain injury, this is my communication aid, I'd rather not talk about it, let's talk about puppies.' Participants with TBI were also interested in the use of other modalities such as audio feedback through Bluetooth earphones such as AirPods as seen in Figure 1. Finally, our expert designers group and participants with social difficulty felt that the interface should be more discreet and should not be displayed to bystanders to stop them believing that the user is stalking.

The ability to listen to conversations to generate notes proposed by participants with social anxiety proposed was a novel concept which could support people with short term memory issues. Participants wanted to know what they commonly talk about with the person such as whether the person was a client or that they had a new kitten based on previous conversations. Participants with TBI extended this concept wanting notes to create calendar events or work with other third-party apps and support people during a conversation.

From these participatory design workshops, we found that using a one size fit all approach would not be possible due to differences between each of the group's requirements. Each set of participants had their unique approaches to privacy and discreteness that were conflicting with one another. Further, some of the ideas and features that participants required were incompatible with one another. For example, the expert design group wanted a social media of voices. However, participants with difficulty socialising wanted the system to be as discreet as possible, with no voices sent away from the device to notify others of the speaker. These differences present new challenges to producing an application to support people with inferring faces.

BUILDING A SPEAKER-RECOGNITION ALGORITHM

We developed a speaker-recognition algorithm to evaluate how well the approach could work with audio data that came from a smartwatch. We used two machine learning algorithms, a feature extractor and an identifying algorithm to understand the audio output of such an algorithm. Each second of audio

Figure 1. An example of a user using wireless Bluetooth headphones.

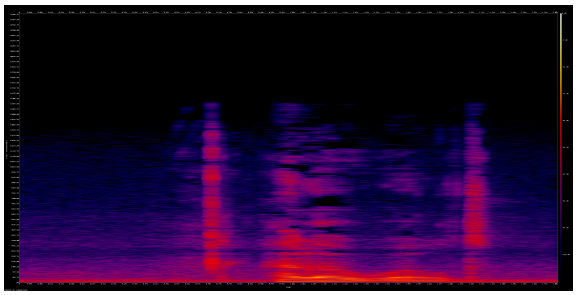


Figure 2. An example of a spectrogram that was produced from the phrase "Hello World".

data contains 16,000 samples with sound parameters requiring an understanding of surrounding data to create a wave and to discriminate the differences between speakers sign waves, resulting in it being too complicated for humans to program with high accuracy required.

Development and Training

The training of machine learning approaches is computationally expensive, and can require specialist computers to execute them. As a result, we reviewed several approaches, some pre-built and trained while others required training from scratch. Multiple speaker-recognition algorithms exist and each had their merits and drawbacks. We selected a modified approach by Christopher Gill [17] because Gill's code was available along with a discussion on implementation, unlike many algorithms We considered several alternatives before settling on Gill's approach. The Microsoft Azure Cognitive Services (MS Speech) [37] system seemed promising. MS Speech is an off-device speaker-recognition text-independent speaker-recognition system that is part of the Microsoft Azure, Microsoft's cloud platform. However, it is still in preview and not technically available in Europe with some technical limitations. Microsoft allowed up to 1,000 voices on MS Speech and returned a confidence level [37]. We also considered the Alizè project [4], an open-source platform for speaker-recognition that contains tools for speaker verification. Although promising, Alizè had poor documentation and proved unreliable during testing. While Aliè seemed encouraging, we quickly ran into difficulty with the demo application.

Gill [17] gave a discussion on implementation along with source code. Gills work was also interesting as it utilised algorithms that should perform well on Mobile devices. We selected this approach and modified it to increase performance and to optimise it for the mobile environment. Gills approach was to use Spectrograms (example in figure 2), a visual representation of sound feeding the voices through a Convectional Neural Network (CNN) which trained on the CIFAR-10 architecture; a CNN designed specifically for processing images. Gill then removed the last layer of the CNN and fed the results into a Support Vector Machine (SVM) acting as a supervised classifier. By using the SVM, we could utilise transfer learning (otherwise known as one-shot learning), which resulted in us not required in storing the training data, which is beneficial due to its large size. To allow us to train our machine learning algorithm, we collected training data. We then used the audio data that we captured as testing data to be able to evaluate our algorithm. We derived testing data from the training data. [5].

We selected high-quality data for training requiring the audio data of a single speaker per sample which was labeled, of high recording quality with minimal background noise. We also required the utterance, the act of speaking, to be longer than 30 seconds.We selected Librivox, a public domain repository of audio books [32], as a source of training and testing data as Librivox contained labelled data for a single speaker and could combine the same speaker for multiple books together. We could also guarantee that the audio was of high quality as Librivox encourages the use of dedicated recording equipment. Librivox recordings were released in the public domain, allowing us to download it legally. We were able to download 303 unique voices from Librivox from several hundred audio books and combine each audio book using FFMPEG[4] and SOX [5] into a single WAV file. We then removed any silences longer than 1 second in length and re-sampled all audio to be at a standard 16khz. We exported each file to a WAV file to remove any compression artefacts that may exist in mp3.

From this data, we produced a pipeline consisting of a feature identification algorithm, along with an identification algorithm. Retraining is required for identification to allow the nodes to understand a new voice. Retraining is extremely computationally expensive resulting in the main mobile machine learning algorithms not supporting algorithm retraining. To combat this, we divided our algorithm in two. The first part is feature extraction, which is computationally expensive to train, however, once produced and pruned, does not require retraining and is efficient to execute. The second algorithm, CatBoost, is a much smaller algorithm that is trained on the data that comes from feature extraction. When the user wants to add a new voice to the system, this smaller neural network requires retraining but retraining CatBoost compared to our WaveNet is significantly computationally cheaper.

[4]FFMPEG hompage: https://ffmpeg.org
[5]SOX homepage: http://sox.sourceforge.net

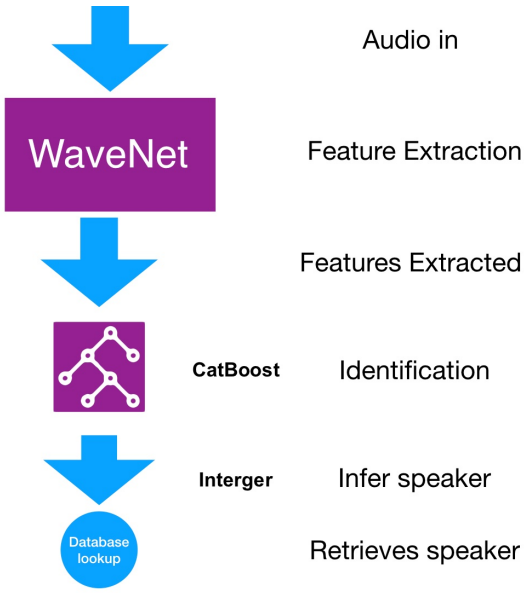


Figure 3. Information flow in our speaker-recognition algorithm. When the phone receives audio data, it is saved temporarily before being passed through a Wavenet for feature extraction. The results are passed to a CatBoost algorithm for classification used in a database lookup.

Our feature extraction algorithm that we selected consisted of DeepMinds WaveNet, a convolution neural network which is designed to produce raw audio for high-quality text-to-speech algorithms for virtual assistants such as Google Assistant [47], which was modified to extract unique features within voices before passing the output to an identification machine learning model. By running feature extraction, we significantly decrease the amount of data that is passed to the identification machine learning algorithm. A CatBoost, short for Categorical Boosting, is a type of gradient boosting algorithm that operates on decision trees. Gradient boosting approaches are suitable for noisy data as they can perform gradient descent in feature space [41]. CatBoost is decision tree-based, combines features to produce a new feature, and considers this data greedily, a combination of the intra-tree feature - by combining fields - generation and inter-tree and inter-tree generation by combining previous tree features. CatBoost first splits, it does not consider the new features [12]. The other beauty of CatBoost is it detects overfitting. [53].

EVALUATION OF AUDIO FROM SMARTWATCH

A crucial part of our application is the capturing of audio data from a smartwatch, in this case, the Apple Watch. Several variables create the characteristics of the microphones, which affect the quality of the recording. Frequency response is how the microphone responds to the different frequency. Some applications require a low-frequency response such as a concert where more bass is needed, while a higher frequency will result in the more treble [13, 14]. Sensitivity is the characteristic that controls the output voltage to the input pressure, or how loud the microphone perceives noise relative to the actual sound.Microphones also produce noise as a byprod-

uct of recording from the small amount of current running through the speaker. As the signal is amplified by the factor of a thousand to make it audible, this can result in electrical noise produced by the microphone, resulting in the smallest amount of noise becoming prevalent. Reducing the prevalence of sound is part of the design of the microphone and can be challenging to mitigate after the microphones have been manufactured [13].We cannot control these variables, but we can test to see how these samples impact on the quality of the audio for machine learning purposes [13, 14].

Methodology

We evaluated the audio that the apple watch recorded through the app "Voice Recorder , voice memo"[6] (Voice Recorder) and then fed this into the speaker.. We based the methodology on the work of Titze et al. [46], which explored how different microphone types work on extracting voice perturbation. Titze used a loudspeaker as the source as they offered a wide range of control and were more consistent than a human speaker [46]. Playback also allowed us to control utterance lengths and to ensure that the audio was clear. We captured the audio with the following background noises:

No background noise, beach, pub, night club, coffee shop, train, aeroplane

To run the study, we used the same quiet room to control background noise. We used an Apple Watch Series 2 (42mm Silver Aluminium). While this specific Apple Watch is waterproof, it is protected by an O-ring which may not completely protect the microphone from the water. If the user was to shower with the Apple Watch, it may impair the success of the recording. While the Apple Watch can pump water out [25], in our experience, this does not always remove all the water, resulting in evaporation being the best method to remove all the water. To mitigate this, we ensure that the Apple Watch did not get wet for at least 24 hours before being used for recording to allow any water to evaporate.

Data capture

We selected five male, and five female voices with clear utterance from the training data, which we then removed from the training data set. Audio data was captured directly onto the Apple Watch using Voice Recorder. Each background noise was saved in a separate file to remove any drift from occurring. For the first ten seconds of audio recording, the user kept the phone's microphone a set distance from the speakers. For the final ten seconds, the user would then walk to a set location with the microphone facing the speaker to measure the sensitivity of the microphone and how it would impact training. We configured background audio to play at 50Db (the equivalent of a quiet home), and we played the speaker audio at 70Db (the equivalent of a conversation). We measured this before the study took place using Decibel X [7].

[6]https://apps.apple.com/gb/app/voice-recorder-voice-memo/id609030412

[7]https://apps.apple.com/GB/app/decibel-x-DB-dba-noise-meter/id448155923

Results

We failed to achieve a high accuracy algorithm in our machine learning algorithm from the watch audio data. After the variance in noise and a large number of trees, we were only able to achieve an accuracy rating of 14.8%, which currently is not suitable for being used to support people with difficulty with faces. We did achieve higher accuracy predicting testing data of 90%. However, using audio data from Librivox data along with a more extensive training size, we were able to increase our accuracy to 42.8%, which is nowhere near the 75% accuracy we found as a tipping point for acceptance of the technology which we discussed previously. With further training and a more optimised algorithm, this number would further improve. We hypothesise the reasoning behind the poor accuracy from the smartwatch is the poor quality that the smartwatch microphones capture. When we listened back to the audio, we noted the low quality and how difficult it was to distinguish the speaker from the surrounding audio. Noise-cancelling can further improve this. However, the Apple Watch does not currently support noise cancelling through its microphone.

Limitations

Variation in output from the Bluetooth speakers may contribute to variability in the results. These variations are a result of characteristics which could be overcome, however, using Librivox, we may encounter compression. Another consideration is manufacturing tolerance in the Apple Watch. While consumers regarded that Apple quality is high, Apple themselves do have accepted tolerances. Apple does not publicise these tolerances. We could utilise several Apple Watches to test for the quality of its audio, however we did not have the resources during our research.

EVALUATION OF DELAY TIME

Machine learning requires some time for the algorithms on mobile devices to process, causing latency between listening and replying. In this study, we want to understand how the delays in our system will result in the acceptance of the application by general users with no known underlying condition. We want to work out the tipping point where our application becomes acceptable. From these numbers, we can infer whether the trade-off in speed for improved in privacy would impact the acceptance and user acceptance of our application. To understand how delays would affect how people use the application, we developed a Wizard of Oz prototype and recruited 7 participants to evaluate our application. Each participant would meet an actor that they had not met before and were asked to carry out a conversation with them. They would not know the name or the conversation topic until it came upon the watch, where the participant was alerted by a vibration.

To determine delay times, We estimated that processing on the phone would take up to 8 seconds based on analysis of the algorithm running on a graphics card. Alternatively, we hypothesised that offloading the processing to the cloud with 5G to Web Servers would lead to the virtually immediate inference of voices but the approach weakens the user's

Figure 4. Experimental setup recreated with researchers: a participant (left) in discussion with an actor (right) with a researcher (centre) triggering the notification

control of data, which was a pivotal part of the decision we made to design the algorithm to run on a mobile device. Further delays occur sending data between the watch and the phone [8]. As research by Mandal et al., [34] has demonstrated, using wearable devices to run inference will result in a worse user experience Mandal found that using Google Glass for facial recognition resulted in a significant increase in processing times. This led us to testing a range of 2, 4 and 8 seconds with our participants.

Methodology

At the beginning of the session, we explained to participants that the watch they wore was attempting to infer who they were talking with and we wanted to evaluate how delays between different modes of inference would affect the usefulness of the application. Participants would then have five trial discussions with the researcher performing different personas before meeting five actors. The actors here were research staff. Participants were made aware of the topics that could come up but were not aware of the names of the people they would meet or who was linked to what prompt. This allowed for some level of recognition of a topic they were familiar with even without knowing the actor they were talking to. Delay was randomised between 2 seconds, 4 seconds and 8 seconds by the researcher and the delay was triggered as soon as the actor started to speak. After the delay finished, the watch would vibrate, displaying the name of the actor and topic to discuss. Participants would then try to work in the name of the actor and the topic without disrupting the conversation. We ensured that each participant apart from two had experienced each of the possible delays. We randomised delays across actors. A example of a conversation can be seen in Figure 4

Once the conversation had finished, we asked participants to rate the quality of the conversation, whether the prompt was useful and whether the delay that was triggered was too long or before they needed it. We compiled these results together, removing any identifiers of participants and actors. After meeting each of the actors, participants then took part in a short qualitative study to discuss their perceptions of delays, and how useful they found the application. Once participants had taken part in the interview, the researchers made participants aware that they were not using an algorithm and that the researcher controlled the name displaying along with the delay and that no personal information was collected. We recruited 7 participants, generating a total of 67 conversations. We discounted one conversation due to error by the researcher in triggering the notification.

[8] https://developer.apple.com/documentation/watchconnectivity

Delay Time	Conversation Quality	Usefulness of prompts	Usefulness of watch
2 Seconds	8.29	8.38	7.86
4 Seconds	8.06	8.00	6.88
8 Seconds	7.60	6.65	5.80

Table 2. This table displays the mean averages from how participants rated the conversation on a scale 0 to 10.

Results

We found that that quality of the conversations decreased in association with increasing delay between 2 and 4 seconds. Participants rating on how useful the watch was fell by 0.98. However, the usefulness of the watch decreased between 2 and 4 seconds by 1. A further decrease in response time between 4 and 8 seconds resulted in the quality of the conversation and the usefulness of the prompt falling again. Participants found the usefulness of the watch declined as well demonstrating that an eight-second delay is not acceptable. These results can be seen in table 2.

Observations by the researchers suggested that the main factor affecting any given conversation was the ability of the actor and the participant to make small talk. Some participants were not able to make small talk at all and could not engage with any conversation until a prompt appeared on the watch. However, some participants (notably P1 and P4) were able to converse naturally with actors and generate small talk while waiting for the prompt to display. However, individual participants were unable to make any small talk and found any delays in the application difficult (specifically P5), highlighting that the ability to small talk was a part in the quality of the application.

These results do suggest that the length of the delay impacts the user experience of the application and that users found a delay longer than four seconds was not acceptable. These results also demonstrate that while other contributing factors may play a significant part in the conversation especially the ability to make small talk, which could warrant further exploration - users consider speed to be a vital factor in the application. This raises questions about how a wearable system might achieve these speeds that we discuss later.

Limitations

Variance between actor speaking and researcher triggering.
As the researcher triggers the alert, the researcher may introduce slight variance int rigger time. We instructed all researchers to trigger the alert as soon as the actor had spoken, however, this would still lead to a slight delay. We estimate that this might add up to a one-second delay on notifications which will affect the perception of delay more on the two seconds delay compared to the perception of delay on eight seconds delay.

Variance in Watch Communications.
In our experience, we have found that it takes on average one to two seconds for a response with no delay and have not found any method to reduce this delay. These delays add two-seconds of variance to our application which we must consider in the results. These delays would still occur without the speaker-recognition algorithm, and we are testing the delay that the algorithm causes, not the delay that the watch takes to communicates.

Actor Confusion
We briefed actors before the session when we explained the study and the topic. However, actors sometimes stated the topic beforehand to the participant or had lead to confusion where the participant stated another topic than they said. To mitigate the situation, we attempted to meet another actor, or if this was not possible, we discarded the data of that conversation.

DISCUSSION

In this research, we have demonstrated that there is potential to use speaker-recognition to support people to recognise faces. However, there are limitations to this work which will restrict its adoptions.

Privacy

Privacy was core to our development of Pwy, with us wanting to take an approach of using on the device inference to ensure that users were in control of their data. A side effect of this is that it would also allow participants to use the application in situations where the internet was inaccessible. We believe that capturing audio from the smartwatch has privacy benefits compared to video from smart glasses which previous work used [34, 49, 51]. While research does show that if participants disclosed their conditions that people are generally more accepting [40], we feel that with that if a user is wearing a smartwatch as AT, bystanders will be less likely to think that the user is wearing an AT as smartwatches are more prevalent than smart glasses.

No group showed any concerns for self-privacy. However, bystander privacy between groups varied significantly. Expert designers felt the system was a useful communication aid for people and could be widely accepted. People with difficulty socialising wanted a passive listening application to help them with confidence about who they are in discussion with and had minimal concern for bystanders privacy. People with TBI were concerned with bystander privacy but felt that people would be accepted once made aware that it was an extension of their brain and that they would happily explain what it was doing, they did not want it to become a topic of conversation. During our evaluation of the delays of the watch, no participants showed any concern of privacy with many of them stating that they would use the application if it became available. However, from the application design, participants could pause listening of the application, which adds an extra level of privacy.

Figure 5. An person wearing a smartwatch

Finally, in this research, we did not make a distinction between pause-able passive listening, where the user could pause listening, and constant passive listening, where the user could not pause listening. With data that we collected in all activities, we are unable to infer how this would affect the watch, and further work in the area would be required to understand limitations.

Building Stigma Free Accessibility Tools

We wanted to develop an Accessibility Tool that was stigma-free and We wanted people to use the application without drawing attention to the technology. As discussed, we outlined that we wanted to develop AT that looked commonplace and to utilise off the shelf products before having to produce anything bespoke. For example, we did investigate possibly using a Raspberry Pi[9] as a method for running the neural networks. However, this would have required another device for the participant to have worn. None of our participants stated that they felt that the watch was an Accessibility Tool or that it was being used as one in qualitative discussion. Participants felt it was a ubiquitous piece of technology, with at least four participants wearing a wearable piece of technology and a further two participants looking at purchasing one. Many participants stated that the would use the application today if it were further polished with one participant in our evaluation study stating ‘this would be very useful as people in my family have short term memory issues.‘

[9]https://www.raspberrypi.org

Limitations of use

From evaluations of our machine learning algorithm and through passing data through a watch, we discovered that microphones on the Apple Watch are low quality and were not sufficient for speaker recognition, which presents a severe barrier to accurate speaker-recognition going forward. A filter for voices might increase the accuracy of the audio. However, a filter may still not be sufficient if the quality of the audio that is coming into the watch is not sufficient. Currently, an external microphone would be required if this application was to be used by users, which is not ideal. In contrast, if higher quality hardware became available, spatial awareness from audio may help with the watch to identify where people are.

Effect of Delays on the System

Delays in the system do impact the user experience (UX) reducing the perceived usefulness of the prompts and the quality of conversation a user can have. A significant decrease in UX was linked to a decline of between 2 and 4 seconds, with a further decrease in UX when increasing the delay from 4 and 8 seconds. This study demonstrates that for the application to support people, the application must be able to work as quickly as possible. Current hardware introduces a 2 second delay that is difficult if not impossible to work around but each delay beyond this further reduces the quality of the work done. This suggests that speed needs to be a top priority in the future. The capabilities of machine learning on mobile phones have drastically increased in the last few years [26], and in the future, it might become possible to run our algorithm in a suitable time frame on a smartphone or even a wearable device. However, on current hardware with current limitations, a trade-off seems to be required between speed and privacy. A deeper understanding of how the effects with people with TBI would be needed to evaluate the trade off.

Approach of the application

Our participatory design workshops demonstrated that each group had a significant, technically different approach to the problems that arise in conversation which needed different technological approaches to address. While people with social difficulties wanted complete discreetness and constant listening, our design experts wanted social media to share voices and data. Technically, social media and discreetness are not easily reconciled. A social media would have to hand over data every time that an individual has had their voice inferred which is not very discreet. Furthermore, social media may add complexity to people with TBI, which is not suitable when trying to remove cognitive loads. Discreetness and the requirements of people living with TBI were incompatible because TBI did not want to deceive bystanders. Participants with TBI were willing to display that they were using the application more with using other modalities such as headphones. These results demonstrate that one size fit all approaches in this instance are not suitable and that each set of participants requires a tailor-made solution. Compromises in the system are possible. However, we do not understand how this would affect the UX of the application. Further, without running an evaluation study with users using the application in their daily lives, it is also difficult to understand the impact of compromises.

CONCLUSION

In this paper, we researched how speaker-recognition can support people recognising their conversational partners. We found during participatory design workshops that different user groups have different concerns about bystanders privacy and design requirements that make a one size fits all approach unsuitable. A trade-off between privacy and speed is required for an application to support people as small delays in the system significantly limit the ability of the application to support people. Furthermore, limitations in the quality of smartwatch microphones lead to the application failing in noisy situations which will impact the usefulness of the application and may lead to further anxiety. However, speaker-recognition is a promising technology which has got the potential to support people with problems recognising faces.

ACKNOWLEDGEMENTS

Blank for review

REFERENCES

[1] Amazon Customer Service. Retrived 28 April 2019. About Alexa Voice Profiles. (Retrived 28 April 2019). https://www.amazon.com/gp/help/customer/display.html?\nodeId=202199440

[2] Charles Arthur. 2013. Google Glass: is it a threat to our privacy? https://www.theguardian.com/technology/2013/mar/06/google-glass-threat-to-our-privacy, *The Guardian* (Mar 2013).

[3] Mihai Bâce, Gábor Sörös, Sander Staal, and Giorgio Corbellini. 2017. HandshakAR: Wearable Augmented Reality System for Effortless Information Sharing. In *Proceedings of the 8th Augmented Human International Conference (AH '17)*. ACM, New York, NY, USA, Article 34, 5 pages. http://doi.acm.org/10.1145/3041164.3041203

[4] Bonastre, J-F and Wils, Frédéric and Meignier, Sylvain. 2005. ALIZE, a free toolkit for speaker recognition. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. IEEE, I–737. DOI: http://dx.doi.org/10.1109/ICASSP.2005.1415219

[5] Andriy Burklov. 2019. *The Hundred-Page Machine Learning Book*. Andriy Burkov.

[6] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, *The Guardian* (2018).

[7] J. P. Campbell. 1997. Speaker recognition: a tutorial. *Proc. IEEE* 85, 9 (Sep. 1997), 1437–1462. DOI: http://dx.doi.org/10.1109/5.628714

[8] Sherryse L Corrow, Kirsten A Dalrymple, and Jason Js Barton. 2016. Prosopagnosia: current perspectives. *Eye and brain* 8 (2016), 165–175. DOI: http://dx.doi.org/10.2147/EB.S92838

[9] Schuster Dana. 2014. The revolt against Google "Glassholes"'. *New York Post* (July 2014). https://nypost.com/2014/07/14/is-google-glass-cool-or-just-plain-creepy/

[10] Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. 2015. I Feel Like I'm Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1411–1414. DOI: http://dx.doi.org/10.1145/2702123.2702141

[11] Katherine Deibel. 2013. A Convenient Heuristic Model for Understanding Assistive Technology Adoption. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '13)*. ACM, New York, NY, USA, Article 32, 2 pages. DOI:http://dx.doi.org/10.1145/2513383.2513427

[12] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. (2018).

[13] Peter Elsea. 1996. Microphones How they work. (1996). http://artsites.ucsc.edu/EMS/Music/tech_background/TE-20/teces_20.html

[14] ESR Eletronic Compnents Ltd. 2019. Microhpone Characterisitcs. https://www.esr.co.uk/sound-light/Infomation/microphones.htm. (September 2019).

[15] Md Sadek Ferdous, Soumyadeb Chowdhury, and Joemon M. Jose. 2017. Analysing privacy in visual lifelogging. *Pervasive and Mobile Computing* 40 (2017), 430 – 449. DOI:http://dx.doi.org/https://doi.org/10.1016/j.pmcj.2017.03.003

[16] Guido Gainotti and Camillo Marra. 2011. Differential Contribution of Right and Left Temporo-Occipital and Anterior Temporal Lesions to Face Recognition Disorders. *Frontiers in Human Neuroscience* 5 (2011), 55. DOI:http://dx.doi.org/10.3389/fnhum.2011.00055

[17] Christopher Gill. 2017. Automatic Speaker Recognition using Transfer Learning. (2017). https://github.com/hamzag95/voice-classification

[18] H. Gish and M. Schmidt. 1994. Text-independent speaker identification. *IEEE Signal Processing Magazine* 11, 4 (Oct 1994), 18–32. DOI: http://dx.doi.org/10.1109/79.317924

[19] Google Support. 2019. Link your voice to your Google Assistant device with Voice Match. https://www.support.google.com/assistant/answer/9071681?co=GENIE.Platform%3DAndroid&hl=en. (2019).

[20] Thomas Grüter, Martina Grüter, and Claus Christian Carbon. 2008. Neural and genetic foundations of face recognition and prosopagnosia. *Journal of Neuropsychology* 2, 1 (2008), 79–97. DOI: http://dx.doi.org/10.1348/174866407X231001

[21] Yoni Halperin, Galit Buchs, Shachar Madienbaum, Mayer Amenou, and Amir Amedi. 2016. Social Sensing: a Wi-Fi based Social Sense for Perceiving the Surrounding People. In *Augmented Human Interaction Confernece*. ACm New York, Geneva, Switzerland, 42–43. DOI:http://dx.doi.org/10.1145/2875194.2875228

[22] Kim Halskov and Nicolai Brodersen Hansen. 2015. The diversity of participatory design research practice at PDC 2002–2012. *International Journal of Human-Computer Studies* 74 (2015), 81–92. DOI: http://dx.doi.org/https://doi.org/10.1016/j.ijhcs.2014.09.003"

[23] Matthieu Hébert. 2008. *Text-Dependent Speaker Recognition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 743–762. DOI: http://dx.doi.org/10.1007/978-3-540-49127-9_37

[24] HM Revenue & Customs. 2017. Voice ID showcases latest digital development for HMRC customers. https://www.gov.uk/government/news/voice-id-showcases-latest-digital-development-for-hmrc-customers. (Jan 2017). https://www.gov.uk/government/news/voice-id-showcases-latest-digital-development-for-hmrc-customers

[25] iFixit. Retrived 6 September 2019. Apple Watch Series 2 Teardown. (Retrived 6 September 2019). https://www.ifixit.com/Teardown/Apple+Watch+Series+2+Teardown/67385

[26] Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool. 2018. AI Benchmark: Running Deep Neural Networks on Android Smartphones. In *The European Conference on Computer Vision (ECCV) Workshops*.

[27] Anil Jain, Lin Hong, and Sharath Pankanti. 2000. Biometric Identification. *Commun. ACM* 43, 2 (Feb. 2000), 90–98. DOI: http://dx.doi.org/10.1145/328236.328110

[28] Anil K. Jain, Karthik Nandakumar, and Arun Ross. 2016. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters* 79 (2016), 80 – 105. DOI: http://dx.doi.org/https://doi.org/10.1016/j.patrec.2015.12.013

[29] Heather Kelly. 2013. Google Glass users fight privacy fears. https://edition.cnn.com/2013/12/10/tech/mobile/negative-google-glass-reactions/index.html, *CNN* (Dec 2013).

[30] Finn Kensing and Jeanette Blomberg. 1998. Participatory Design: Issues and Concerns. *Computer Supported Cooperative Work (CSCW)* 7, 3 (01 Sep 1998), 167–185. DOI: http://dx.doi.org/10.1023/A:1008689307411

[31] Marion Koelle, Matthias Kranz, and Andreas Möller. 2015. Don't Look at Me That Way!: Understanding User Attitudes Towards Data Glasses Usage. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 362–372. DOI: http://dx.doi.org/10.1145/2785830.2785842

[32] LibriVox. Retrived 22 August 2019. Librivox: Free public domain audiobooks. https://librivox.org. (Retrived 22 August 2019). https://librivox.org

[33] Konrad Lishka. 2010. Google-Debatte: Datenschützer kritisieren W-Lan-Kartografie[German: privacy advocates criticize W-Lan cartography [German: Google debate privacy advocates criticize W-FI cartography]. (April 2010). http://www.spiegel.de/netzwelt/web/google-debatte-datenschuetzer\-kritisieren-w-lan-kartografie-a-690600.html

[34] Bappaditya Mandal, Shue-Ching Chia, Liyuan Li, Vijay Chandrasekhar, Cheston Tan, and Joo-Hwee Lim. 2015. A Wearable Face Recognition System on Google Glass for Assisting Social Interactions. In *Computer Vision - ACCV 2014 Workshops*, C. V. Jawahar and Shiguang Shan (Eds.). Springer International Publishing, Cham, 419–433.

[35] Gary Marshall. 2013. Google Glass: say goodbye to your privacy. https://www.techradar.com/news/mobile-computing/google-glass-say-goodbye-to-your-privacy-1134796, *TechRadar* (Mar 2013).

[36] Roisin McNaney, John Vines, Daniel Roggen, Madeline Balaam, Pengfei Zhang, Ivan Poliakov, and Patrick Olivier. 2014. Exploring the Acceptability of Google Glass As an Everyday Assistive Device for People with Parkinson's. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2551–2554. DOI: http://dx.doi.org/10.1145/2556288.2557092

[37] Microsoft Azure. Retrived 22 August 2019. Speaker Recognition PREVIEW. https://azure.microsoft.com/en-gb/services/cognitive-services/speaker-recognition/. (Retrived 22 August 2019). https://azure.microsoft.com/en-gb/services/cognitive-services/speaker-recognition/

[38] Cosmin Munteanu, Ronald Baecker, Gerald Penn, Elaine Toms, and David James. 2006. The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, ACM, New York, NY, USA, 493–502. DOI: http://dx.doi.org/10.1145/1124772.1124848

[39] Phil Parette and Marcia Scherer. 2004. Assistive technology use and stigma. *Education and Training in Developmental Disabilities* 39, 3 (2004), 217–226.

[40] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4884–4895. DOI: http://dx.doi.org/10.1145/2858036.2858130

[41] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 6638–6648. http://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf

[42] N. K. Ratha, J. H. Connell, and R. M. Bolle. 2001. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal* 40, 3 (2001), 614–634.

[43] Siri Team. 2017. Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant - Apple. *Apple Machine Learning Journal* (2017). https://machinelearning.apple.com/2017/10/01/hey-siri.html

[44] Siri Team. 2018. Personalized Hey Siri. *Apple Machine Learning Journal* (2018). https://machinelearning.apple.com/2018/04/16/personalized-hey-siri.html

[45] Litza Stark, Steve Whittaker, and Julia Hirschberg. 2000. ASR satisficing: the effects of ASR accuracy on speech retrieval. In *Sixth International Conference on Spoken Language Processing*, Vol. 3.

[46] Ingo Titze and William S. Winholtz. 1993. The Effect of Microphone Type and Placement on Voice Perturbation Measurements. *Journal of speech and hearing research* 36 (12 1993), 1177–1190. DOI:http://dx.doi.org/https://doi.org/10.1044/jshr.3606.1177

[47] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. (2016).

[48] John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. 2013. Configuring participation: on how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, ACM, New York, NY, USA, 429–438. DOI: http://dx.doi.org/10.1145/2470654.2470716

[49] He Wang, Xuan Bao, Romit Roy Choudhury, and Srihari Nelakuditi. 2015. Visually Fingerprinting Humans without Face Recognition. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '15 (MobiSys '15)*. ACM, ACM, New York, NY, USA. DOI: http://dx.doi.org/10.1145/2742647.2742671

[50] Matt Warman. 2013. Google Glass: We'll all need etiquette lessons. https://www.telegraph.co.uk/technology/google/10015697/Google-Glass-well-all-need-etiquette-lessons.html, *The Telegraph* (Apr 2013).

[51] Xi Wang, Xi Zhao, V. Prakash, Weidong Shi, and O. Gnawali. 2013. Computerized-eyewear Based Face Recognition System for Improving Social Lives of Prosopagnosics. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops (PervasiveHealth '13)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 77–80. DOI:http://dx.doi.org/10.4108/icst.pervasivehealth.2013.252119

[52] Yale Vision Group. 1997. Yale Face Database. http://vision.ucsd.edu/content/yale-face-database. (1997).

[53] Yandex. Using the overfitting detector. (????). https://catboost.ai/docs/features/overfitting-detector-desc.html

[54] Lucy Yardley, Lisa McDermott, Stephanie Pisarski, Brad Duchaine, and Ken Nakayama. 2008. Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of Psychosomatic Research* 65, 5 (nov 2008), 445–451. DOI:http://dx.doi.org/10.1016/J.JPSYCHORES.2008.03.013

D Ethics Application for Participatory Design Workshop

Our participatory design workshops requrired us to submit a "Application For Ethical Approval Of Projects Involving Human Subjects" Form to Swansea University College Of Science Ethics Committee. This was approved.

For our evaluation, as we did not work with any vulnerable people (which TBI patients are classed as vulnerable) we were able to complete a short form.

APPLICATION FOR ETHICAL APPROVAL OF PROJECTS INVOLVING HUMAN SUBJECTS

RESEARCH CANNOT COMMENCE UNTIL ETHICAL APPROVAL HAS BEEN OBTAINED

Please note form is opened as read-only.

Reference Number: STU_CSCI_105526_301018173646_1

Status: Approved Proposal :College Ethics Committee DECISION Details

Submitted By: Stephen Lindsay

Submitted Date: 28 Nov 2018

1. TITLE OF PROJECT

Supporting patients with Prosopagnosia through speaker recognition

2. APPLICANT NAME(s)

Osian Smith, Stephen Lindsay, Joss Whittle

3. PROPOSED START DATE

December 2018

4. DURATION (months)

Nine months

5. OBJECTIVES

Briefly state what the project is designed to achieve.

In this project, we are planning to do a participatory design session with patients with brain trauma to understand how we can design a digital system that identifies people they are speaking with using speech recognition to alleviate embarrassment in social situations and improve wellbeing.

6. LOCATION OF STUDY

6.1. List the country and location(s) where the data will be collected

Computational Foundry, Headway Centres and South West wales Brian Injury Group (SWWBIG) facilities, Swansea, Wales.

6.2. Identify the person(s) who will be present to supervise the research at that location

Osian Smith, Stephen Lindsay, Joss Whittle.

7. STUDY DESIGN

7.1. Outline the study design (e.g. cross-section, longitudinal, intervention, RCT, questionnaire etc.)

Participatory Design Workshop Series (3 sessions, 2 hours approx.)

We are planning to run participatory design workshop with people living with brain injury to understand the problems that they face with identifying who they are talking to and how digital technologies like Smart Watches might support their social interactions. Activities will include:

Collecting demographic data via survey on all participants.
Show a video following a design theatre method where actors will use the application that we are currently developing to give a sense of the possibilities technology offers.
Have focus group discussion in response to the idea.
Use paper prototyping (post-it notes, large sheets to record ideas, sketches) to design potential responses to the problem areas outlined.

7.2. State the number and characteristics of study participants

We will aim to work with 10 participants living independently with brain trauma in groups of 3-4.

7.3. State the inclusion criteria for participants

Our inclusion criteria for participants are:
- They self-report a diagnosis of Prosopagnosia or they self-report issues with it
- They currently live independently

7.4. State the exclusion criteria for participants and identify any requirements for health screening

Our exclusion criteria for participants are:
- If they are reliant on daily home care
- If they have other medical issues that prevent them from giving informed consent

7.5. Will the study involve vulnerable populations (i.e. children, elderly, those with cognitive impairment or in unequal relationships, disabled, clinical, etc.) or people who are unable to give informed consent? **Yes/No – if Yes, please justify.**

(Please note that people with learning disabilities fall under The Mental Capacity Act 2005 and must be reviewed by the NHS; other vulnerable groups may not require NHS review but will typically require Disclosure & Barring Service (DBS) clearance (formerly CRB checks) - Evidence for this will be required.

This group is likely to include people with brain injuries that lead to memory problems which mean that they are classified as vulnerable.

We will work with Headway and SWWBIG staff to ensure that participants are able to give informed consent. In addition, we will take the following precautions:
Research team members have prior experience working with people with brain injury.
Members of Headway or SWWBIG will be present in all sessions to offer support
If research staff feel that participants do not understand the study, we will not work with them or we will not include their data if we feel it would upset or distress them to exclude them.
If their situations change during the study their data will be destroyed.

7.6. Will parental/coach/teacher consent be required? **Yes/No - If Yes, please specify which and how this will be obtained and recorded**

No

7.7. Are there any requirements/commitments expected of participants (e.g. time, exertion level)?

They are expected to take part in 2-hour studies.

8. PARTICIPANT RECRUITMENT

8.1. Briefly outline how and from where will participants be recruited

We will approach local charities and organisations such as Headway and the South West Wales Brain Injury Group (SWWBIG) to recruit participants. We will talk to staff first and then ask for introductions to participants to explain the research.

9. DATA COLLECTION

9.1. Briefly describe the type of data that will be collected

We will gather necessary information about the participants to understand their background
Age
Gender
Time since injury
Symptoms
Contact telephone number

We will focus on collecting qualitative data documenting their thoughts about the video prototype that we have demonstrated and paper copies of designs that we will work on in the sessions.

9.2. Briefly describe how the data will be collected

1.Demographic data will be collected at the beginning of the session by questionnaire.
2.Sessions will be audio-recorded for later analysis
3. Paper designs will be collected at the end of the sessions and stored

9.3. Will the collection of data be undertaken by Swansea University staff or students? **Yes/No - if No, please explain who is responsible for data collection and give permit details (permit number, date, issuing body) or explain why these are not needed**

Yes, this will be collected by Swansea University Staff and student

9.4. Briefly describe how you propose to ensure participant confidentiality and anonymity. If anonymity is not to be preserved explain why not

We will not record any names in the research data and each participant will receive a unique
ID number on any documents that file the information. Publications will not include any information that can identify the participant unless we are requested to include their name.

Past experience has shown that some participants are proud of their contributions and would like their name to appear in the published work.

9.5. Will the research involve respondents to the Internet or other visual / vocal methods where they may be identified (e.g. IP address)

No

9.6. Will participants be given information on the study and consent forms? **Yes/No - If No please justify**

Yes - see attached

9.7. Will the research involve the sharing of data of confidential information beyond initial consent? **Yes/No - If Yes please explain**

No

9.8. Will the information be collected from participants without their knowledge and consent at the time? (e.g. secondary use or re-use of social media content; covert observation/photos of people in non-public places, etc.). **Yes/No If Yes, please justify:**

No

9.9. Will any substance be administered to participants? **Yes/No - if Yes, please explain:**

(Please note that substances falling under the auspices of the Medicines for Human Use (Clinical Trials) Regulations 2004 and will require additional review by the NHS)

No

9.10. Will tissue samples (including blood) be obtained from participants? **Yes/No - if Yes, please explain:**

(Please note that collection of tissue samples would fall under the terms of the Human Tissue Act 2004 and will require additional review by the NHS)

No

9.11. Is a first aider needed? **Yes/No – If Yes, please identify**

No

9.12. Will the study discuss or collect sensitive information (e.g. terrorism; sexual activity; drug use, criminal activity) **Yes/No - if Yes, please explain:**

No

9.13. Will the research involve the collection of administrative or secure data that requires permission before use? **Yes/No - if Yes, please explain:**

No

10. STORAGE AND DISPOSAL OF DATA and SAMPLES

10.1. Briefly describe the procedures to be undertaken for the storage and disposal of data and samples

Any electronic data that captured will be shared security between researchers by the university OneDrive network. The data will be disposed of by Dr Stephen Lindsay after 5 years.

10.2. Who will have the responsibility for the storage and disposal of data and/or samples?

Osian Smith and Stephen Lindsay

10.3. For how long will the data and/or samples be retained after completion of the study? (normally 5 years, or end of award)

11. POTENTIAL RISKS AND DISCOMFORTS

11.1. Are there any potential physical risks or discomforts to the participants in the study? **Yes/No – if Yes, please explain**

No

11.2. Are there any potential physical risks or discomforts to the researcher(s) conducting the study? **Yes/No – if Yes, please explain**

No

12. OTHER ETHICAL ISSUES OF CONCERNS

If none, then please state 'none'

Participants living with traumatic brain injury frequently experience mild memory problems. These can lead to problems recalling details and purposes of a study which raises questions around informed consent. However, their personality does not change and memory can, typically, be 'jogged' by re-presentation of that information therefore, we do not anticipate anyone being upset that they are taking part in a study after being reminded of that fact.

Participants reasoning about the world is not typically impaired and they are able to live independent lives so we do not anticipate any issues around initial understanding of the project.

13. APPLICATION CHECK LIST

Tick as appropriate below.

	Yes	No	N/A
Have you included a **Participant Information Sheet** for participants in the study?	☑	☐	☐
Have you included a **Parental/Guardian Information Sheet** for parents/guardians?	☐	☑	☐
Have you included a **Participant Consent (or Assent) Form** for participants in the study?	☑	☐	☐
Have you included a **Parental/Guardian Consent Form** for parents/guardians?	☐	☑	☐
For collaborative projects carried by **outside organisations**, have you included details of ethics permits?	☐	☑	☐

14. DECLARATION

Please read the following declarations carefully and provide details below of any ways in which your project deviates from these. ☑

I certify the answers to the questions given above are true and accurate to the best of my ability

I have ensured that there will be no active deception of participants.

I have ensured that no data will be personally identifiable.

I have ensured that no participant should suffer any undue physical or psychological discomfort (unless specified and justified in methodology).

I certify that there will be no administration of potentially harmful drugs, medicines or foodstuffs.

I have (or will obtain) written permission from an appropriate authority before recruiting members of any outside institution as participants.

I certify that the participants will not experience any potentially unpleasant stimulation or deprivation.

I certify that the above statements are true with the following exception(s):

Conset form.pdf

Bill of rights.pdf

Amendment 1.docx

Consent form changes.pdf

Bill of rights.pdf

Amendment 2.docx

Consent form_Amendment 2.pdf

E Ethical Consent forms and Bill of Rights

Each participant for all studies in this research project recived the following consent forms

Swansea University Department of Computer Science

Bill of Rights

The Following Document Lists Your Rights While Participating In The Participatory Design Session For Supporting Patients With Prosopagnosia Through Speaker Recognition.

As A Research Participant You Have The Right:

To be treated with respect and dignity during every phrase of the research.

- To be fully and clearly informed of all aspects of the research prior becoming involved.
- To be fully and clearly informed of all aspects of the research prior to becoming involved in it.
- To enter into clear, informed, and written agreement with the researcher prior to becoming involved in the activity. You should sense **NO** pressure, explicit or otherwise, to sign this contract.
- To choose explicitly whether or not you will become involved in the research under the clearly stated provision that refusal to participate or the choice to withdraw during the activity can be made at any time without penalty to you.
- To be treated with honesty, integrity, openness, and straightforwardness in all phases of the research, including a guarantee that you will not unknowingly be deceived during the course of the research.
- To demand proof that an independent and competent ethical review of human rights and protections associated with the research has been successfully completed.
- To demand complete patient confidentiality and privacy in any reports of the research unless you have explicitly negotiated otherwise.
- To expect that your personal welfare is protected and promoted in all phases of the research, including knowing that no harm will come to you.
- To be informed of the results of the research study in a language you understand.
- To be offered a range of research studies or experiences from which to select, if the research is part of fulfilling your educational or employment goals.

The contents of this bill was prepared by the University of Calgary who examined all of the relevant Ethical Standards from the Canadian Psychological Association's Code of Ethics for Psychologists, 1991 and rewrote these to be of relevance to research participants. Descriptions of the CPA Ethical Code and the CPA Ethical Standards relevant to each of these rights are available at http://www.cpa.ca/ethics2000.html and http://www.psych.ucalgary.ca/Research/ethics/bill/billcode.html if you would like to examine them.

The complete CPA Ethical Code can be found in Canadian Psychological Association "Companion manual for the Canadian Code of Ethics for Psychologists" (1992).

Swansea University Department of Computer Science

Consent form

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information that you may receive.

Research Project Title

Supporting patients with Prosopagnosia through speaker recognition

Researchers

Mr O Smith, Dr S Lindsay, Dr J Whittle

Study Purpose

This participatory design session is to understand how patients with prosopagnosia (face blindness) can be supported through a mobile application.

Participant Recruitment and Selection

Morriston Hospital Brain Trauma Center will recruit participants, Swansea, Wales by being personally approached.

Procedure

You will be asked to attend a 2-hour session where the researchers will demonstrate a video design theatre where a participatory design session will then take place.

Data Collection

Your age and sex will be collected to allow us to understand the basic demographics of the group.

We will also capture relevant information that is discussed within the session that we feel will assist us in designing our future product for research.

We will also collect an audio recording of the session and may be processed by our speaker recognition algorithm. Any voice data will not be used to train or test our algorithm outside of the workshop. We may photo capture and record workings on the desk. We will not capture anything that may identify you in any manner.

Data archive

Data will be kept securely. Data will be stored electronically which will be encrypted and password protected. The password will be kept safe and separate from the data. The investigator will destroy study data after it is no longer of use. Usually, this will be at the end of the research project when results are fully reported and disseminated.

Likelihood of Discomfort

There is no likelihood of discomfort or risk associated with participation.

Confidentiality

Your name and your feedback may be stated in the final piece of work. Your telephone number and email will not be shared to anyone and will be destroyed along with policy stated in data archive.

Researcher

Mr O Smith is working on his master's in research in Computing and Future Interaction Technologies at Swansea University. This study will contribute to his research into supporting patients with prosopagnosia through speaker recognition. His supervisor is Dr S Lindsay.

Mr O Smith can be contacted by email via o.l.smith@swansea.ac.uk

Dr S Lindsay can be contacted by email via s.c.lindsay@swansea.ac.uk or by calling 01792 606958

Finding out about Results

Participants can find out the results of the study by contacting the researcher after September 30, 2019.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and your agreement to take part as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free not to answer specific items or questions in interviews or on questionnaires. You are free to withdraw from the study at any time without penalty. Your continued participation should be as informed as your initial consent, and you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact the researcher.

_____ _____

Participant Date

_____ _____

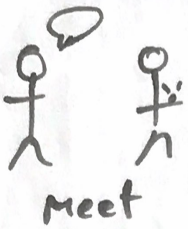
Investigator/Witness Date

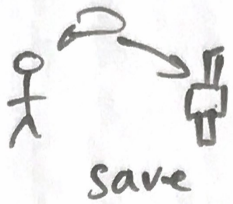
A copy of this consent form has been given to you to keep for your records and reference.

F Mateirnal Generated From Participatory Design Workshop

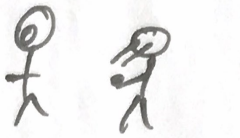
F.0.1 Expert Design Study

ADD

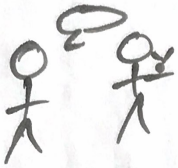


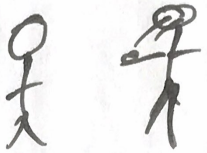


Retrive

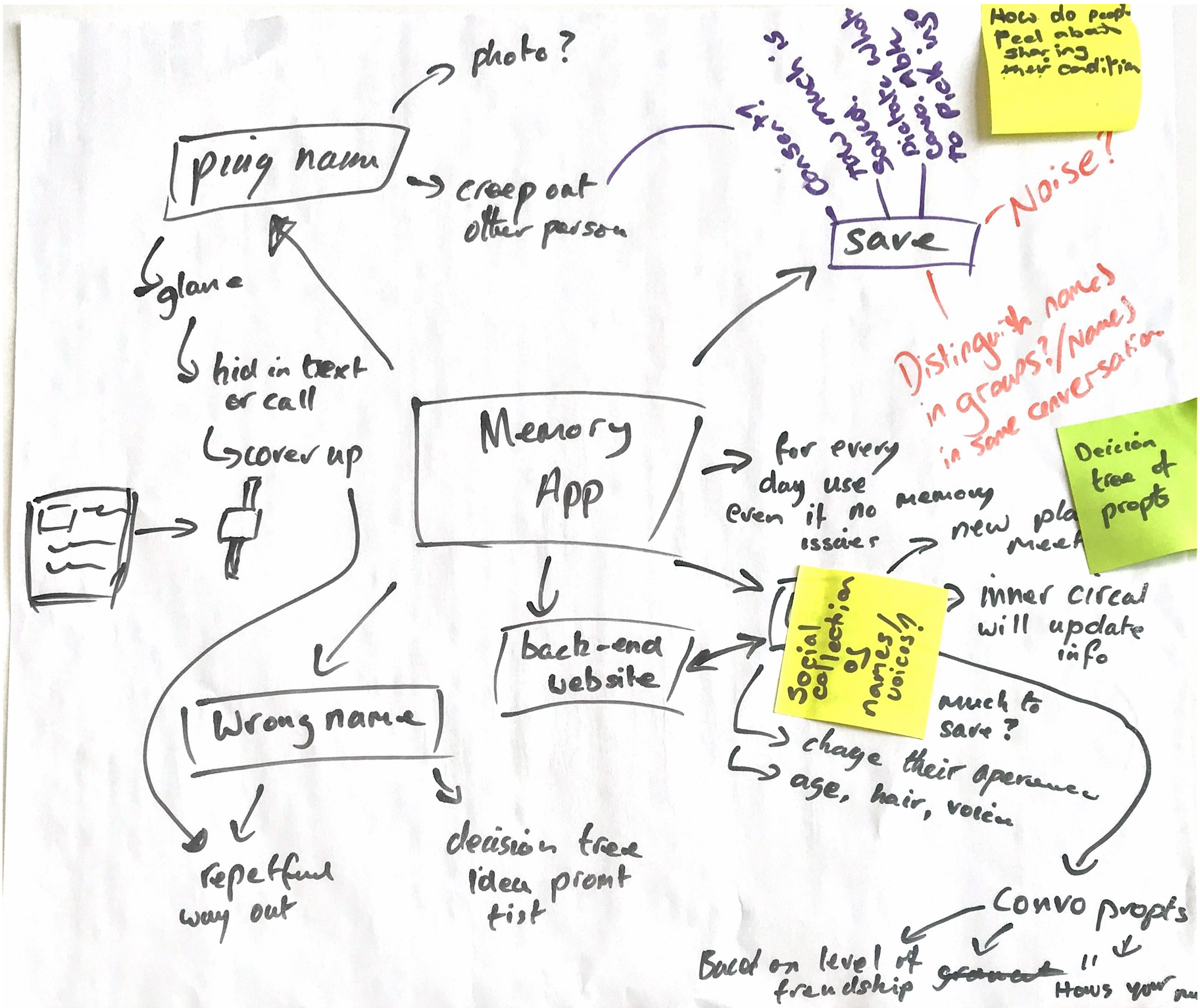


Fail

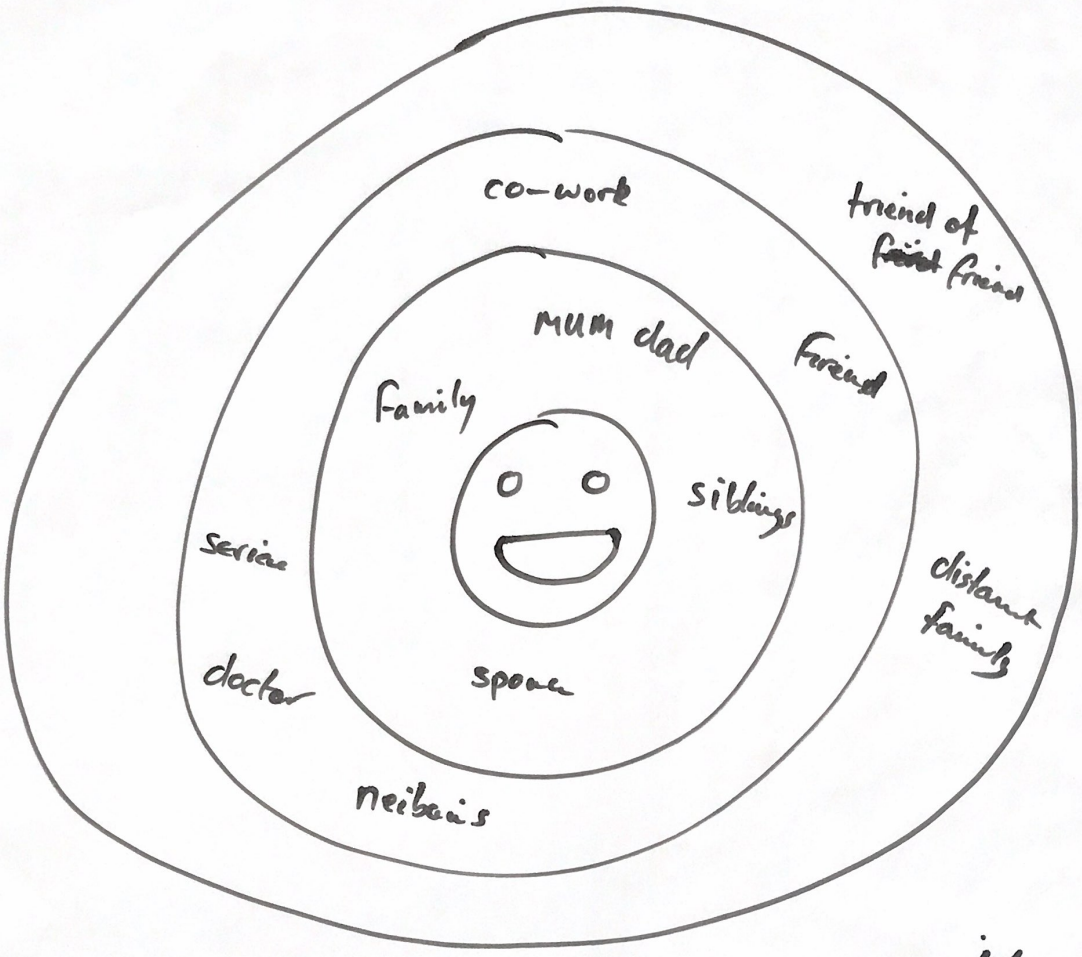




+	+ save convo to refrence to + hid the notifcate as a text.	+ respetful way to ask name
- noice interference (award loan)	- Someone seeing name - to many people in the convo	



photo?
ping name
creep out other person
glance
hid in text or call
cover up
Memory App
save
Noise?
Distinguish names in groups?/Names in same conversation
How do people feel about sharing their condition
Decision tree of propts
for every day use even if no memory issues
new pla meet
inner circal will update info
Social collection of names/voices?
much to save?
change their aperence age, hair, voice
back-end website
Wrong name
repetful way out
decision tree idea promt list
Convo propts
Based on level of friendship
Hows your


co-work
friend of friend
MUM dad
friend
family
siblings
service
distant family
doctor
spouse
neibours
hard to explain
Sisan
individuum
confidence

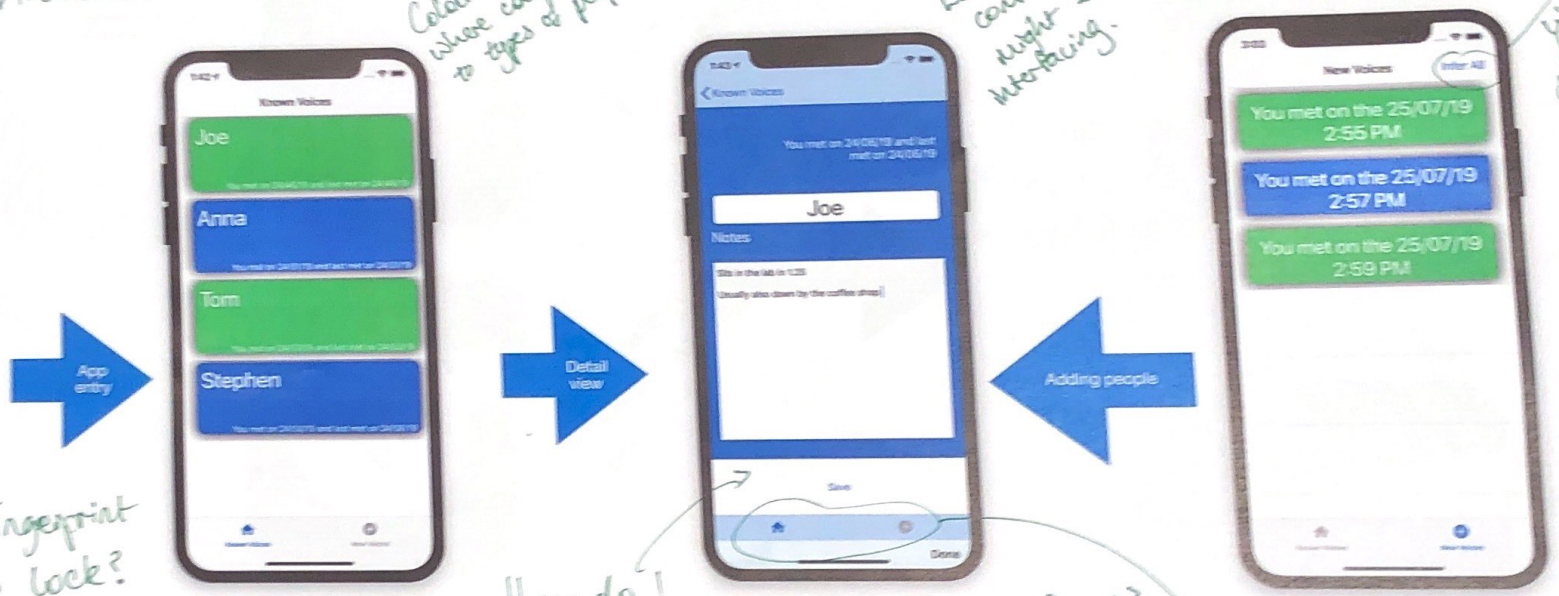
F.0.2 People Who Find Socialising Difficult Design Exhibits

#1

Is it available on Android?

Colour customisation, where colours correspond to types of people

Link to real context? That might enable interfacing.



Home Screen

More information

Adding people

Fingerprint to lock?

How do I delete the 'voice'? What if I don't want it anymore?

Is there a way to view recordings - if they are stored? There should be an option to delete old recordings — if they are stored, anyway.

Why are these options still here?

?. What does this do?

29 of July 2019

Version 2.1

Version 2.0 + ?

A
Listening Button similar to a record button

This looks creepy. It's a pretty stark colour.

B
A button without any text - can be more discreet

Large buttons just attract attention, I feel.

C
Similar to design C however with text stating what it says

Writing just attracts attention, too

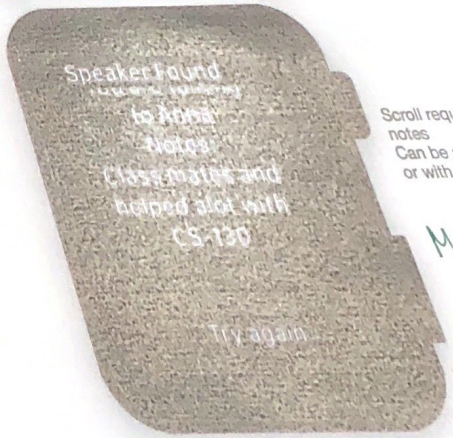
Home Screen

Listning

Trying again

If someone has bad eyesight, then large writing may be needed.

There need to be an option to expand this information, as there is too much writing at the moment!



Scroll required
notes
Can be achie
or with the

Speaker Found Screen

29 of July 2019

A
Listening Button similar to a record button

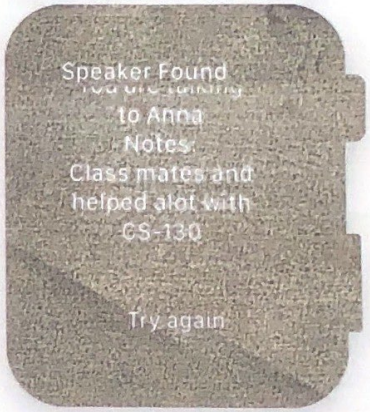
B
A button without any text - can be more discreet

C
Similar to design C however with text stating what it says

Listning

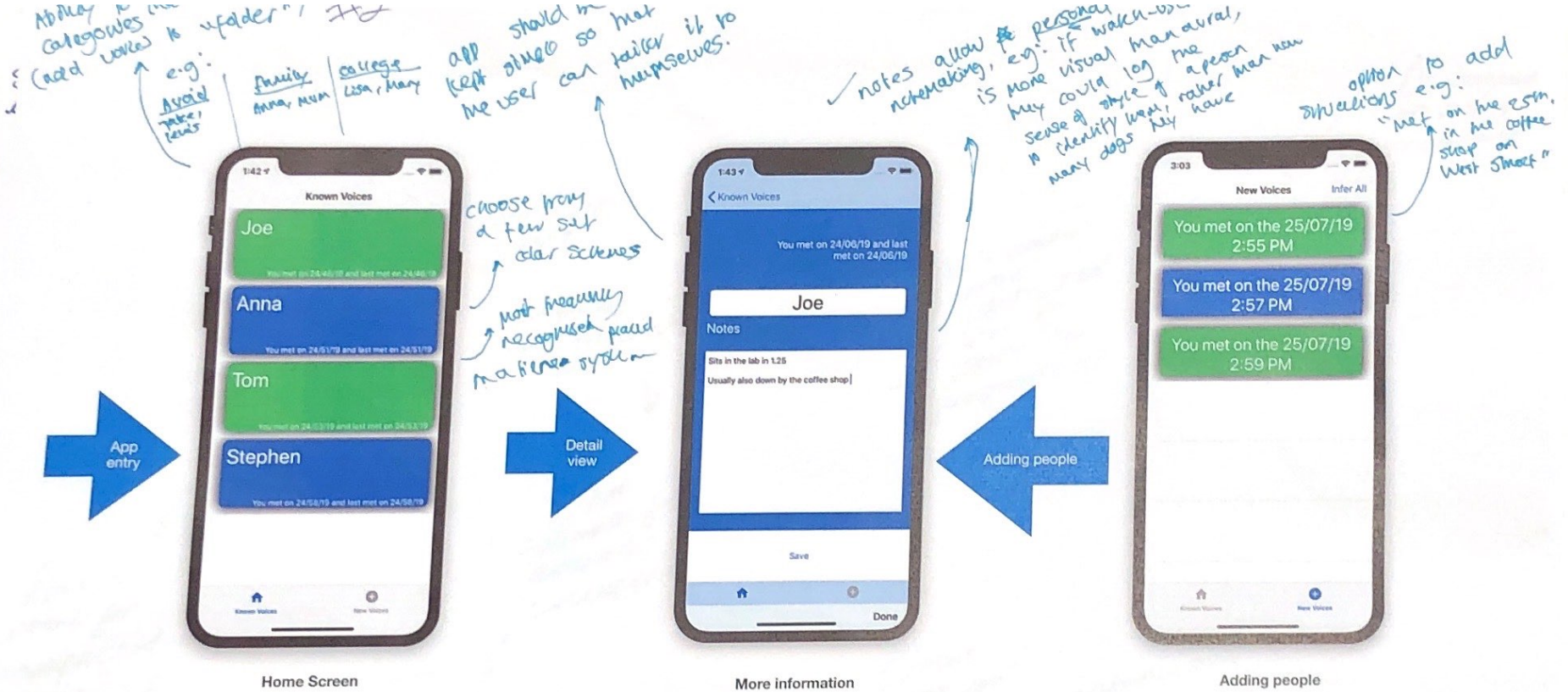
Trying again

Home Screen



Scroll required to read notes
Can be achieved on screen or with the crown

Speaker Found Screen



Home Screen

More information

Adding people

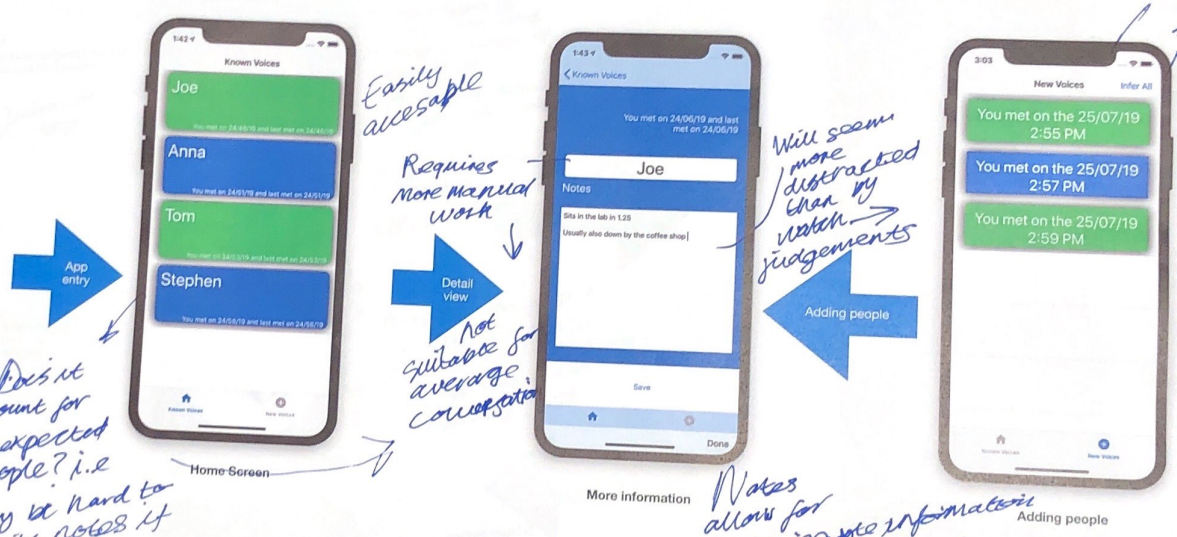
✓ large text, simple interface

- ability to personalise is key - every experience with a disorder is different.

[what are the consequences if user doesn't abide by a request to remove stored data?]

#3

Seen intro



Easily accessible

Requires more manual work

Not suitable for average conversation

Will seem more distracted than by watch judgements

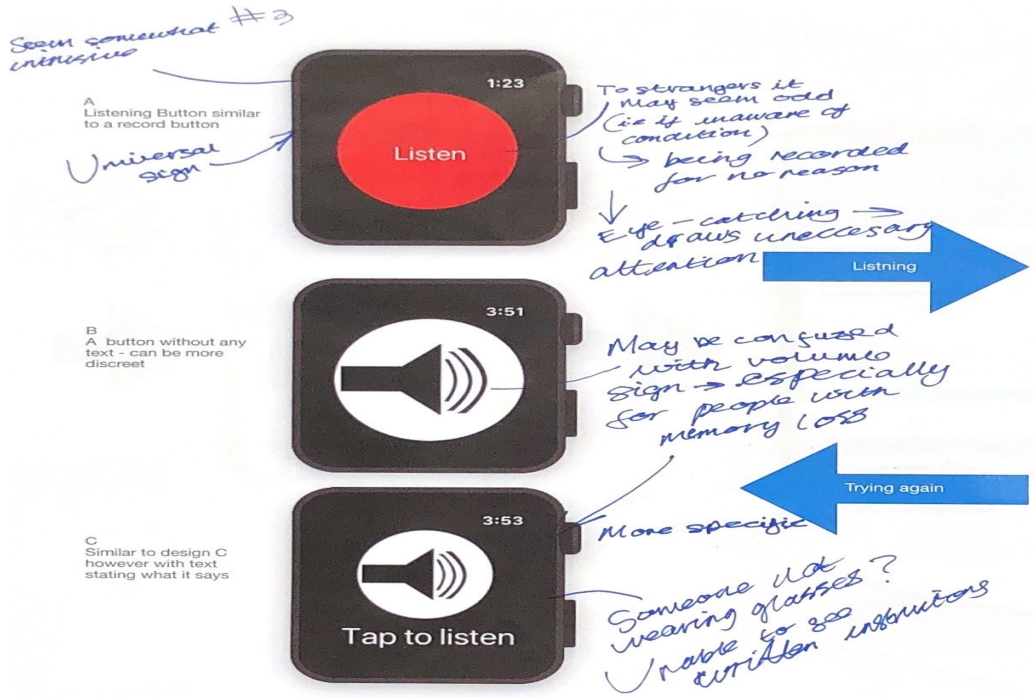
Discomfort by being recorded

Does it account for unexpected people? i.e May be hard to write notes if phone isn't on hand

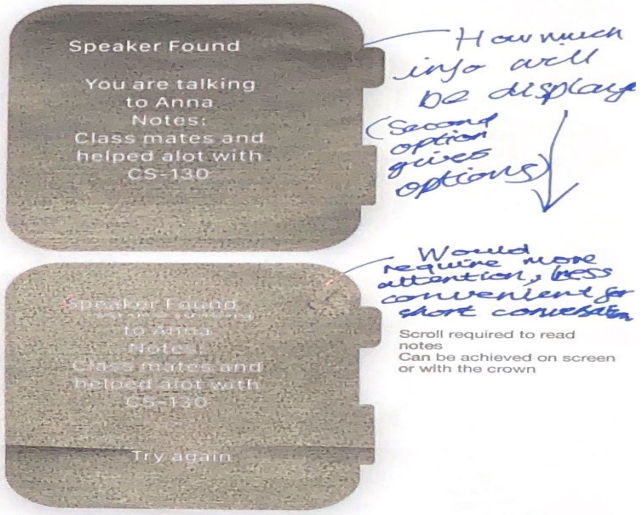
Notes allows for customisable ~~custom~~ information on specific person

29 of July 2019

Version 2.1



Home Screen

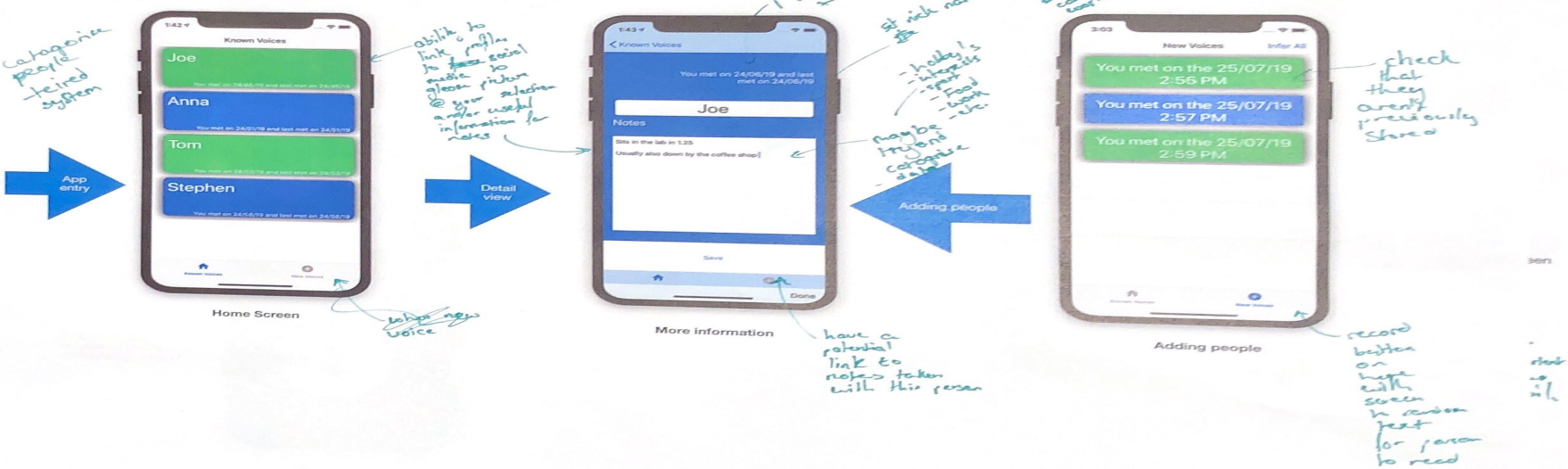


Speaker Found Screen

29 of July 2019

Version 2.1

#4


Known Voices
Joe
Anna
Tom
Stephen
App entry
Home Screen
Detail view
Known Voices
You met on 24/06/19 and last met on 24/06/19
Joe
Notes
Sits in the lab in 1.25
Usually also down by the coffee shop!
Save
Done
More information
Adding people
New Voices
Infer All
You met on the 25/07/19 2:56 PM
You met on the 25/07/19 2:57 PM
You met on the 25/07/19 2:59 PM
Adding people

#4

A
Listening Button similar to a record button

less intimidating

Customisable watch face

B
A button without any text - can be more discreet

C
Similar to design C however with text stating what it says

Home Screen

Listning

Trying again

Scroll required to read notes
Can be achieved on screen or with the crown

a swipe base more or less information. With more important information shown with less detail.

Speaker Found Screen

29 of July 2019

Version 2.1

#5

Allow update / of nicknames

Ensure they aren't already stored

Allows manual input of notes that may have been missed by app

#5

A
Listening Button similar to a record button

Button could be more discreet

B
A button without any text - can be more discreet

C
Similar to design C however with text stating what it says

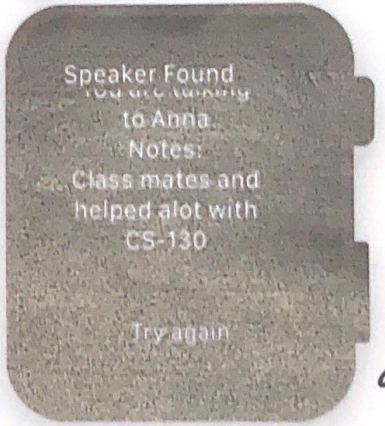
Most subtle

Home Screen

Listning

Trying again

Can be construed as 'stalker-ish'



Scroll required to read notes
Can be achieved on screen or with the crown

Stored person could demand the removal of their data

A wrong name could be insulting

Speaker Found Screen

29 of July 2019

Version 2.1

F.1 Raw Results From evlaution of delay

F.1.1 2 seconds delay

Table F.1: Raw Data of 2 seconds delay

How comfortable did you find the conversation	How useful was the prompts on the watch to the conversation	Was the information on the prompts timely or not?
5	6	8
9	6	6
8	8	10
8	8	9
9	9	8
8	10	10
7	8	8
9	10	10
7	8	6
8	8	8
10	10	10
8	8	9
7	8	4
7	5	2
9	9	8
10	9	7
9	9	9
9	10	10
8	9	7
10	9	7
9	9	9

F.1.2 4 seconds delay

Table F.2: Raw Data of 4 seconds delay

How comfortable did you find the conversation	How useful was the prompts on the watch to the conversation	Was the information on the prompts timely or not?
9	10	7
8	8	8
9	8	9
7	8	4
8	4	3
5	7	4
8	8	7
10	10	10
6	3	7
10	9	8
9	9	8
10	10	9
6	6	7
8	9	6
8	9	7
8	10	6

Table F.3: Raw Data of 8 seconds delay

How comfortable did you find the conversation	How useful was the prompts on the watch to the conversation	Was the information on the prompts timely or not?
8	8	8
7	7	7
9	0	1
6	8	5
8	8	3
4	7	7
8	4	3
7	6	5
9	10	8
5	3	5
6	0	3
8	5	3
8	8	6
9	10	10
9	10	6
9	7	9
9	9	6
7	8	9
6	7	5
10	8	7

G Code for Evlaution of Algorithm (WaveNet and CatBoost)

CatBoost

September 21, 2019

[ ]:
```
print("Importing")
import os, time
#os.environ['CUDA_VISIBLE_DEVICES'] = '0'
from glob import glob
from pathlib import Path

import numpy as np
import tensorflow as tf
import h5py
from scipy.io.wavfile import read as read_wav
from IPython.display import Audio, display

import matplotlib.pyplot as plt
def set_fig_size(figw=1500, figh=1000, figdpi=100):
    fig = plt.figure(facecolor='white', figsize=(figw/figdpi, figh/figdpi),␣
 ↪dpi=figdpi)

import tensorflow.keras.backend as K
from tensorflow.keras import Model
from tensorflow.keras.optimizers import Adam, SGD
from tensorflow.keras.losses import sparse_categorical_crossentropy
from tensorflow.keras.metrics import sparse_categorical_accuracy,␣
 ↪sparse_top_k_categorical_accuracy
from tensorflow.keras.layers import Input, Conv1D, Dropout, BatchNormalization,␣
 ↪LeakyReLU
from tensorflow.keras.callbacks import TensorBoard, ModelCheckpoint
from keras import models

from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification

#dataset for testing random forrests without able to use wavnet
from sklearn.datasets import load_iris
from sklearn import datasets
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

```
print("Importing completed")
speakers = ["205","223","205","250","256","263","264","271","272","326"]
datalocation = '/Users/osiansmith/Documents/University/MRES/Application/Test␣
 ↪code/Data/'


print("We have " + str(len(speakers)) + " speakers")
```

[ ]:
```
#this file loads in all the audio data
def loadFileNames():
    #extracts data for each speaker
    speakeresList = []
    for voiceFile in speakers:
        #gets the first speaker
        fileURL = (datalocation + voiceFile)
        rawFiles = []
        for filename in sorted(glob(fileURL + "/*.wav")):
            rawFiles.append(filename)
        speakeresList.append(rawFiles)
    print("Speakers loaded in")
    return speakeresList


loadFileNames()
```

[ ]:
```
def getWav(file_path):
    print(np.array(file_path).shape)
    sample_rate, waveform = read_wav(file_path)
    print(waveform)
    print(waveform.shape)
    #having to resize the data here
    chunk_size = sample_rate * 30
    file_name  = file_path[:-4].rsplit('/',1)[-1]
    print(file_path, 'sample_rate', sample_rate, 'waveform', waveform.dtype,␣
 ↪waveform.shape, 'chunk_size', chunk_size)

    if len(waveform.shape) > 1:
        print('skipping')
    else:
        #archiveLocation =   datalocation +  speakers[count]
        with h5py.File(file_path +'.h5', mode='w') as archive:
                archive.create_dataset('sample_rate', data=sample_rate)
```

```
                archive.create_dataset('waveform', data=waveform)

def convertFile(files):
    print(len(files))
    for i in files:
        getWav(i)

def readfiles():
        wav_files = loadFileNames()
        for i in wav_files:
            convertFile(i)
#             print("convering = " + str(i))
readfiles()
print("done")
```

[ ]:
```
#this file loads in all the audio data
def loadh5FileNames():
    #extracts data for each speaker
    speakeresList = []
    for voiceFile in speakers:
        #gets the first speaker
        fileURL = (datalocation + voiceFile)
        rawFiles = []
        for filename in sorted(glob(fileURL + "/*.h5")):
            rawFiles.append(filename)
        speakeresList.append(rawFiles)
    print("Speakers loaded in")
    return speakeresList


data = loadh5FileNames()
for i in data:
    print(i)
    print("\n")
```

[ ]:
```
def model(num_inputs, num_filters, num_layers, num_classes):

    if (type(num_filters) is int):
        num_filters = [num_filters]*num_layers
    assert type(num_filters) == list

    window = (2**(num_layers+1))-1
    print('window: ', window)

    inputs = Input(shape=(None, num_inputs))

    x = inputs
    for lvl in range(0, num_layers):
```

```
        rate   = 1
        stride = 2
        field = (2**(lvl+1))
        #print('lvl: %2d rate: %6d field: %6d' % (lvl, rate, field))
        x = Conv1D(num_filters[lvl], 2, strides=stride, dilation_rate=rate,␣
 ↪padding='causal')(x)
        x = LeakyReLU(0.1)(x)

    #outputs = Conv1D(num_classes, 1, padding='same', activation='softmax')(x)

    model_obj = Model(inputs=inputs, outputs=x)
    return model_obj
```

[ ]:
```
def time_to_meta_batch(y_true, y_pred):
    return K.reshape(y_true, (-1, 1)), K.reshape(y_pred, (-1, y_pred.shape[-1]))

def loss(y_true, y_pred):
    return sparse_categorical_crossentropy(*time_to_meta_batch(y_true, y_pred))

def accuracy(y_true, y_pred):
    return sparse_categorical_accuracy(*time_to_meta_batch(y_true, y_pred))

def top_k(y_true, y_pred):
    return sparse_top_k_categorical_accuracy(*time_to_meta_batch(y_true,␣
 ↪y_pred))
```

[ ]:
```
def createRandomForrests(X_train,y_train):
    #Creates a radnom forests off with n classifiers
    randForest = RandomForestClassifier(n_estimators=150)
    #fits tge rabdin firsest
    randForest.fit(X_train,y_train)
    return randForest
#test data
iris = datasets.load_iris()
#organise data
data=pd.DataFrame({
    'sepal length':iris.data[:,0],
    'sepal width':iris.data[:,1],
    'petal length':iris.data[:,2],
    'petal width':iris.data[:,3],
    'species':iris.target
})

X=data[['sepal length', 'sepal width', 'petal length', 'petal width']]  #␣
 ↪Features
y=data['species']  # Labels
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3) # 70%␣
 ↪training and 30% test
```

```
data.head()

forest = createRandomForrests(X_train,y_train)
y_pred = forest.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

[ ]:
```
def loadWaveNet():

    #Loads in the models and generates
    h5_files       = sorted(glob('hdd-data/16k/h5/*.h5'))
    sample_rate    = 16000 #22050
    window         = 10 # seconds
    window_samples = window * sample_rate

    num_epochs     = 128
    num_steps      = 128
    batch_size     = 32
    num_filters    = 128
    num_layers     = 10
    num_classes    = len(h5_files)

    #loads model into function
    WaveNet  = model(128, num_filters, num_layers, num_classes)

    optimizer = Adam()
    metrics   = [accuracy, top_k]
    WaveNet.compile(optimizer, loss='sparse_categorical_crossentropy')

    print(WaveNet.summary())

    WaveNet.load_weights('weights.h5', by_name=True)
    return WaveNet
#     data =  loadh5FileNames()
#     testdata = data[0]
#     print(testdata)
#     perecitions = model.predict_classes(testdata)

    #freeze(model_to_convert)

    #model_to_convert.save("fullModel.h5")

   # converter = tf.lite.TFLiteConverter.from_keras_model_file("model.h5")

    #tflite_model = converter.convert()
    #open("converted_model.tflite", "wb").write(tflite_model)
```

```
loadWaveNet()
```

[ ]:
```
num_filters    = 128
num_layers     = 10
num_classes    = loadh5FileNames()
model = model(128, num_filters, num_layers, num_classes) #loadWaveNet()# loads␣
↪model
```

[ ]:
```
def make_dataset(data_paths, batch_size, window_samples, assert_sample_rate,␣
↪device_num=0,
                 prefetch_size=tf.contrib.data.AUTOTUNE, num_parallel_calls=16,␣
↪name='dataset'):
    with tf.variable_scope(name):
        num_speakers = len(data_paths)

        sample_rates = [h5py.File(file_path, mode='r')['sample_rate'].value for␣
↪file_path in data_paths]
        assert len(set(sample_rates)) == 1
        sample_rate = sample_rates[0]
        #print('sample_rate :', sample_rate)
        assert sample_rate == assert_sample_rate

        def decode_example(speaker_idx):
            with h5py.File(data_paths[speaker_idx], mode='r') as archive:
                length = archive['waveform'].shape[0]
                idx    = np.random.randint(length-window_samples)
                return archive['waveform'][idx:idx+window_samples]

        def sample_example(speaker_idx):
            waveform, = tf.py_func(decode_example, (speaker_idx,), (tf.int16,))
            waveform.set_shape((window_samples,))
            waveform = tf.cast(waveform, tf.float32) / 32767.
            waveform = tf.reshape(waveform, shape=(-1,))
            return waveform, tf.cast(speaker_idx, tf.int64)

        data       = tf.data.Dataset.
↪from_tensor_slices(list(range(num_speakers)))
        data       = data.apply(tf.data.experimental.
↪shuffle_and_repeat(buffer_size=num_speakers))
        data       = data.map(sample_example,␣
↪num_parallel_calls=num_parallel_calls)
        data       = data.batch(batch_size, drop_remainder=True)
        data       = data.prefetch(prefetch_size)
```

```
        data       = data.apply(tf.data.experimental.prefetch_to_device('/
↪device:GPU:%d' % (device_num)))
        iterator   = data.make_one_shot_iterator()
        return iterator.get_next()
```

[ ]:
```
def waveform_to_log_mel_spectrogram(waveforms, sample_rate, num_mel_bins=128,
                                    lower_edge_hertz=20.0,␣
↪upper_edge_hertz=4000.0,
                                    frame_length=1024, frame_step=2**4,␣
↪log_eps=1e-6,
                                    name='waveform_to_log_mel_spectrogram'):
    with tf.variable_scope(name, reuse=tf.AUTO_REUSE):
        stft       = tf.contrib.signal.stft
        mel_matrix = tf.contrib.signal.linear_to_mel_weight_matrix
        magnitude_spectrograms = tf.abs(stft(waveforms, frame_step=frame_step,
                                             frame_length=frame_length,␣
↪fft_length=frame_length))

        num_spectrogram_bins = magnitude_spectrograms.shape[-1].value
        mel_weight_matrix    = mel_matrix(num_mel_bins, num_spectrogram_bins,␣
↪sample_rate,
                                          lower_edge_hertz, upper_edge_hertz)
        mel_spectrograms     = tf.tensordot(magnitude_spectrograms,␣
↪mel_weight_matrix, 1)
        log_mel_spectrograms = tf.log(mel_spectrograms + log_eps)

        return log_mel_spectrograms
```

[ ]:
```
def makeDataset(Speaker):
        sample_rate = 16000 #22050
        batch_size  = 1
        window      = 1 # seconds
        window_samples = window * sample_rate


        print("Speaker =" + str(Speaker))
        for sample in Speaker:
            sample_waveforms, sample_idxs = make_dataset(Speaker, batch_size,␣
↪window_samples, sample_rate, prefetch_size=1)
            log_mel_spectrograms =␣
↪waveform_to_log_mel_spectrogram(sample_waveforms, sample_rate)
            #print(log_mel_spectrograms)
```

[ ]:
```
import pickle, os
from sklearn.neighbors import NearestNeighbors
from sklearn.neighbors.nearest_centroid import NearestCentroid
from catboost import CatBoostClassifier, Pool
```

[ ]:
```
#Here we load in the raw audio into the program and get temporal data
def getTemporalData(FileData):
    sample_rate = 16000 #22050
    batch_size  = 1
    window      = 1 # seconds
    window_samples = window * sample_rate
    rawAudio = []
    model = loadWaveNet()
    for speaker in FileData:
        speakerSamples = []
        #print(speaker)
        for sample in speaker:
            globSample = glob(sample)
            print(globSample)

            pklSample = globSample[0] + '.pkl'
            if os.path.isfile(pklSample):
                #print('loading from pickle')
                with open(pklSample, 'rb') as f:
                    results = pickle.load(f)

            else:
                #print('caching to pickle')
                #print("Glob = " + str(globSample)+ "\n")
                sample_waveforms, sample_idxs = make_dataset(globSample,␣
 ↪batch_size, window_samples, sample_rate, prefetch_size=1)

               # print(sample_waveforms.shape)

                log_mel_spectrograms =␣
 ↪waveform_to_log_mel_spectrogram(sample_waveforms, sample_rate)

                #print(log_mel_spectrograms.shape)
                results = []
                for _ in range(100):
                    results += [model.predict(log_mel_spectrograms, steps=1)]

                with open(pklSample, 'wb') as f:
                    pickle.dump(results, f)

#             print(results.shape)
#             plt.figure()
```

```
#             plt.hist(np.reshape(results, (-1,)))

#             plt.show()

            speakerSamples.append(results)
      #  print("speakerSamples = " + str(len(speakerSamples)) + "\n\n new␣
 ↪sample \n\n")
        rawAudio.append(speakerSamples)
        #print("Lenght =" + str(len(speakerSamples)))
   # print("RawAudio = ")
    #print("rawAudio = " + str(len(rawAudio)))
    return rawAudio
```

[ ]:
```
sample_rate = 16000 #22050
batch_size  = 1
window      = 1 # seconds
window_samples = window * sample_rate
#fileData =  glob()#loadh5FileNames()
rawData = getTemporalData(loadh5FileNames())
print("******")
print("Temporal Data generated")
print("Generating Testing data")

x_dataLoading = []
acc = 0

train_env_idxs = [0,1]#[3,4,5]

y_data = []
for idx, i in list(enumerate(rawData)):
    train_envs = [i[j] for j in train_env_idxs]
    for sample in train_envs:
        x_dataLoading += sample
        y_data        += [idx] * len(sample)

    #x_data[acc].pop(0)
    acc = acc + 1
x_data = np.concatenate(x_dataLoading, axis=0)
x_data = x_data.reshape((-1, x_data.shape[-1]))

print(np.array(x_dataLoading).shape)

x_data = np.concatenate(x_dataLoading, axis=0)
x_data = x_data.reshape((-1, x_data.shape[-1]))
print(x_data.shape)
```

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
x_pca = pca.fit_transform(x_data)

plt.figure()
print(x_pca.shape)
plt.scatter(x_pca[:,0], x_pca[:,1], c=y_data)

plt.show()

augment_data = True #Just set this to true to see if it makes a diffrence

if augment_data:
    x_data = np.concatenate([x_data for _ in range(100)], axis=0)
    x_data += np.random.normal(scale=1e-4, size=x_data.shape)
    y_data = np.concatenate([y_data for _ in range(100)], axis=0)

#randForest = RandomForestClassifier(n_estimators=100, max_features=99,␣
 ↪max_depth= 150)
randForest = None #CatBoostClassifier()
pklSample = 'model.pkl'
if False: #we don't want to load from file just yet
    if os.path.isfile(pklSample):
        print('loading model from pickle')
        with open(pklSample, 'rb') as f:
                    randForest = pickle.load(f)
else:
    print("Training from raw data")
    randForest = CatBoostClassifier(num_trees=1000, max_depth = 10,verbose =␣
 ↪True)
    randForest.fit(x_data, y_data)
with open(pklSample, 'wb') as f:
            pickle.dump(randForest, f)

env_names = ['Clean', 'Beach', 'Pub', 'Nightclub', 'Coffee Shop', 'Train']

for env_idx, env_name in enumerate(env_names):
    x_test = []
    y_test = []
    for idx, i in enumerate(rawData):
        trainingData = i[env_idx]
        x_test.append(trainingData)
        y_test += [idx] * len(trainingData)
        #x_data[acc].pop(0)
        acc = acc + 1
```

```
    x_test = np.concatenate(x_test, axis=0)
    x_test = x_test.reshape((-1, x_test.shape[-1]))
    print(x_test.shape)

    print(env_name)
    y_pred = randForest.predict(x_test)
    print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

[ ]:
```
sample_rate = 16000 #22050
batch_size  = 1
window      = 1 # seconds
window_samples = window * sample_rate
#fileData =  glob()#loadh5FileNames()
rawData = getTemporalData(loadh5FileNames())
print("******")
print("Temporal Data generated")
print("Generating Testing data")

x_dataLoading = []
acc = 0
allSamples = [0,1,2,3,4,5,6,7,8,9]


trainingDataPositions = [0,3]

testingData = [1,2,4,5]#[x for x in allSamples if x not in␣
 ↪trainingDataPositions]#trainingData - allSamples



y_data = []
for idx, i in list(enumerate(rawData)):
    train_envs = [i[j] for j in trainingDataPositions]
    for sample in train_envs:
        x_dataLoading += sample
        y_data        += [idx] * len(sample)

    #x_data[acc].pop(0)
    acc = acc + 1
x_data = np.concatenate(x_dataLoading, axis=0)
x_data = x_data.reshape((-1, x_data.shape[-1]))

print(np.array(x_dataLoading).shape)

x_data = np.concatenate(x_dataLoading, axis=0)
x_data = x_data.reshape((-1, x_data.shape[-1]))
print(x_data.shape)
```

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
x_pca = pca.fit_transform(x_data)

plt.figure()
print(x_pca.shape)
plt.scatter(x_pca[:,0], x_pca[:,1], c=y_data)

plt.show()

augment_data = True #Just set this to true to see if it makes a diffrence

if augment_data:
    x_data = np.concatenate([x_data for _ in range(200)], axis=0)
    x_data += np.random.normal(scale=1e-4, size=x_data.shape)
    y_data = np.concatenate([y_data for _ in range(200)], axis=0)

#randForest = RandomForestClassifier(n_estimators=100, max_features=99,
 ↪max_depth= 150)
randForest = None #CatBoostClassifier()
pklSample = 'model.pkl'
if False: #we don't want to load from file just yet
    if os.path.isfile(pklSample):
        print('loading model from pickle')
        with open(pklSample, 'rb') as f:
                    randForest = pickle.load(f)
else:
    print("Training from raw data")
    randForest = CatBoostClassifier(num_trees=200, max_depth = 12,verbose =
 ↪True)
    randForest.fit(x_data, y_data)
with open(pklSample, 'wb') as f:
            pickle.dump(randForest, f)



predictionResults = []
env_names = ['Clean', 'Beach', 'Pub', 'Nightclub', 'Coffee Shop', 'Train']

for env_idx, env_name in enumerate(env_names):
    x_test = []
    y_test = []
    for idx, i in enumerate(rawData):
        trainingData = i[env_idx]
```

```
        x_test.append(trainingData)
        y_test += [idx] * len(trainingData)
        #x_data[acc].pop(0)
        acc = acc + 1
    x_test = np.concatenate(x_test, axis=0)
    x_test = x_test.reshape((-1, x_test.shape[-1]))
    print(x_test.shape)

    print(env_name)
    y_pred = randForest.predict(x_test)
    predictionResults.append(metrics.accuracy_score(y_test, y_pred))
    print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
print(predictionResults)

tranedDataAverage = 0
for i in trainingDataPositions:
    tranedDataAverage += predictionResults[i]
tranedDataAverage = tranedDataAverage / len(trainingDataPositions)
print("Trained data average = " + str(tranedDataAverage))

testDataAverage = 0
for i in testingData:
    testDataAverage += predictionResults[i]

testDataAverage = testDataAverage / len(testingData)
print("Test data average = " + str(testDataAverage))
```

[ ]: