DESIGNING FOR QUALITY IN REAL-WORLD MOBILE CROWDSOURCING SYSTEMS

Mohammad T Othman

School of Computing

Newcastle University

This dissertation is submitted for the degree of Doctor of Philosophy

January 2021

BIOGRAPHICAL SKETCH

Mohammad T Othman received a B.S. in Computer Science from Applied Science University in Jordan in 2005, and an MSc in Network Systems from the University of Sunderland in 2007. Since 2007, he has worked in the industry as a software architect building enterprise level web and mobile applications for various purposes. Currently he is applying the research skills learnt from this study to manage the research and development of virtual and augmented reality training solutions for the petrol and gas industry.

Publications

Othman, M., Amaral, T., Mcnaney, R., Smeddinck, J. D., Vines, J., & Olivier, P. (2017). CrowdEyes : Crowdsourcing for Robust Real - World Mobile Eye Tracking. 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017.

Mcnaney, R., Othman, M., Richardson, D., Dunphy, P., Amaral, T., Miller, N., Stringer, H., Olivier, P. & Vines, J. (2016) Speeching: Mobile Crowdsourced Speech Assessment to Support Self - Monitoring and Management for People with Parkinson’s. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 4464–4476.

Pennington, L., Stamp, E., Smith, J., Kelly, H., Parker, N., Stockwell, K., Aluko, P., Othman, M., Brittain, K. & Vale, L. (2019) Internet delivery of intensive speech and language therapy for children with cerebral palsy: a pilot randomised controlled trial. BMJ Open. 9 (1)

ACKNOWLEDGEMENT

I would like to start by thanking my supervisors Prof. Patrick Olivier for his great inspirational directions, Prof. John Vines for his terrific knowledge in the design domain, and Prof. Peter Wright for his much-appreciated constant support during this research. I would also like to thank Prof David Kirk for his support in finishing this thesis, and my colleagues; Dr. Róisín Mcnaney, Dr. Telmo Amaral, Dr Lindsay Pennington and others for collaborating with me on this research.

To everybody in Open Lab; thank you for sharing your time with me and inspiring me with your research.

My wife Iris who has had to endure listening to years of PhD woes, and who has unquestionably been my source of enthusiasm over past many years, and especially during this PhD study. And a special thanks to my young son Hamza who kept me awake most nights first two years of this PhD when he was a baby, and my daughter Polina, the shiny star of my life.

Finally, I would like to thank my father and my mother for their steadfast support throughout my studies and indeed throughout my life and who are always there to offer their kind words of encouragement in times of need.

ABSTRACT

Crowdsourcing has emerged as a popular means to collect and analyse data on a scale for problems that require human intelligence to resolve. Its prompt response and low cost have made it attractive to businesses and academic institutions. In response, various online crowdsourcing platforms, such as Amazon MTurk, Figure Eight and Prolific have successfully emerged to facilitate the entire crowdsourcing process. However, the quality of results has been a major concern in crowdsourcing literature. Previous work has identified various key factors that contribute to issues of quality and need to be addressed in order to produce high quality results. Crowd tasks design, in particular, is a major key factor that impacts the efficiency and effectiveness of crowd workers as well as the entire crowdsourcing process.

This research investigates crowdsourcing task designs to collect and analyse two distinct types of data, and examines the value of creating high-quality crowdwork activities on new crowdsource enabled systems for end-users. The main contribution of this research includes 1) a set of guidelines for designing crowdsourcing tasks that support quality collection, analysis and translation of speech and eye tracking data in real-world scenarios; and 2) Crowdsourcing applications that capture real-world data and coordinate the entire crowdsourcing process to analyse and feed quality results back. Furthermore, this research proposes a new quality control method based on workers trust and self-verification. To achieve this, the research follows the case study approach with a focus on two real-world data collection and analysis case studies. The first case study, Speeching, explores real-world speech data collection, analysis, and feedback for people with speech disorder, particularly with Parkinson's. The second case study, CrowdEyes, examines the development and use of a hybrid system combined of crowdsourcing and low-cost DIY mobile eye trackers for real-world visual data collection, analysis, and feedback. Both case studies have established the capability of crowdsourcing to obtain high quality responses comparable to that of an expert. The Speeching app, and the provision of feedback in particular were well perceived by the participants. This opens up new opportunities in digital health and wellbeing. Besides, the proposed crowd-powered eye tracker is fully functional under real-world settings. The results showed how this approach outperforms all current state-of-the-art algorithms under all conditions, which opens up the technology for wide variety of eye tracking applications in real-world settings.

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by my advisors Professor Patrick Olivier (Monash University), Professor John Vines (Northumbria University), Professor Peter Wright (Newcastle University) and Professor Dave Kirk (Newcastle University).

I have also worked with a number of doctoral trainees and research associates, each of whom have made considerable contributions to the projects presented in this dissertation. Dr. Telmo Amaral contributed to both Speeching and CrowdEyes projects with his math and analytical skills, Dr. Roisin McNaney mentored the development of Speeching app and led the trial study with participants with Parkinson's, Dan Richardson contributed to the engineering underlying Speeching app. Speeching study was also mentored by Professor Nick Miller, while CrowdEyes was mentored by Dr. Jan D Smeddinck.

My work was generously funded by EPSRC Grant No: EP/K037366/1 *MyPLACE: Mobility and Place for the Age Friendly City*. Data supporting this publication is openly available under an 'Open Data Commons Open Database License'. Additional metadata are available at: http://dx.doi.org/10.17634/122839-1 and http://dx.doi.org/10.17634/141304-1. Please contact Newcastle Research Data Service at rdm@ncl.ac.uk for access instructions.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

1 Introduction

Crowdsourcing is the act of outsourcing tasks to an undefined list of people via an open-call for participation (Howe, 2006). Crowdsourcing is facilitated by crowdsourcing online *platforms* (e.g., Amazon MTurk) where a group of people, known as *requesters,* call-on another unspecified group of people (known as *Workers*) to solve a defined problem.

Crowdsourcing as a concept is by no means a recent idea. History shows that in 1714 the British government commissioned a public competition to find a solution for "*the longitude problem*" which caused the deaths of thousands of sailors and ship passengers each year (Saxton et al., 2013). The British government back then sought innovative solutions from the public in an open call contest, and offered £20,000 in return. Perhaps, this is the first case of crowdsourcing. This example of crowdsourcing is significant, given how a problem like this, almost unsolvable, was solved by a member of the public. Yet, the ethical dilemma was evident back then, when the government was hesitant to award the solution founder, John Harrison, as he was the son of a carpenter. Later on, in 1783 the King of France offered a prize in an open call contest to separate alkali from salt (Halder, 2014). While these examples highlight how crowdsourcing can be conceived as a way of collecting many ideas or potential innovations, historical examples also exist of crowdsourcing as a form of gathering information or 'data' from a large number of people. For example, the Oxford English Dictionary was initiated via an open call for members of the public to submit words and examples of how they are used. Further, in the field of astronomy, crowdsourcing has been used since the early 19th century as a way of collecting observations and sightings of stars and meteors—and similarly in journalism as a way of crosschecking facts. However, Brabham (Brabham, 2013) doesn't consider such examples as crowdsourcing since participants were only bounty hunters and the contest was not internet based. For Brabham crowdsourcing is an online model aims to solve problems by gathering a high number of shared resources.

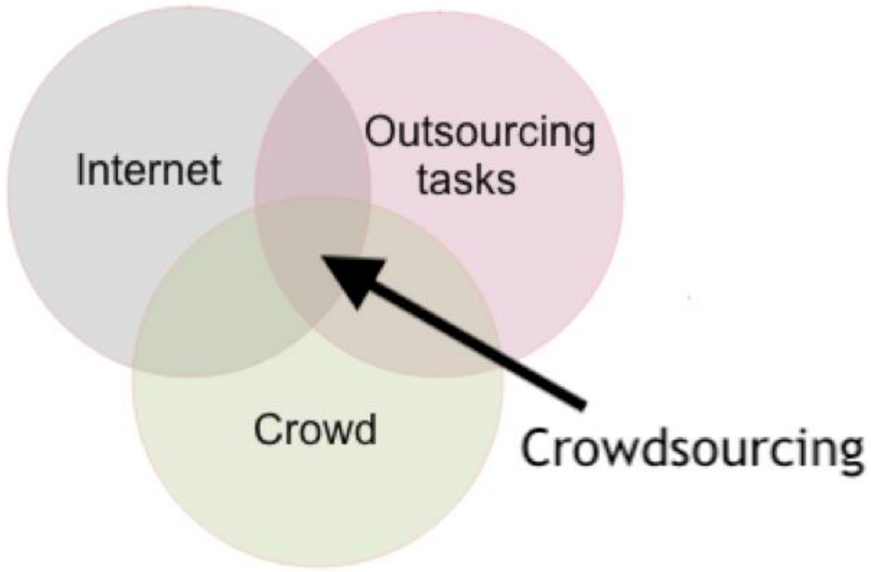


Figure 1-1 The three building blocks of Crowdsourcing

Crowdsourcing model has come to greater attention over this century through its application in a range of Internet-based and -mediated services. In this context, the concept came to popular attention through a 2006 Wired magazine article by Jeff Howe and Mark Robinson (Howe, 2008; Schenk & Guittard, 2011). Howe noted that:

> *'... Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.' (Howe, 2006)*

Both crowdsourcing and outsourcing here refer to the operation of taking a task and performing it off-shore—where off-shore refers to having third parties or individuals and companies external to an organisation complete tasks on their behalf. However, while outsourcing requires tasks to be performed by a fixed number of contracted trusted and accountable professionals, crowdsourcing sends these tasks to be completed by a larger number of potentially anonymous people with different skill sets. Hence the term *Crowd*sourcing. As such, crowdsourcing is often seen as a way of quickly and competitively developing solutions and providing services (Balicki et al., 2014), and those who set tasks typically only pay for what meet their expectations (Schenk & Guittard, 2009; Barbier et al., 2012).

The way in which the crowdsourcing approach is defined typically depends upon its application. For example, Brabham (Brabham, 2008) and Doan *et al.* (Doan et al., 2011) define crowdsourcing as a model for problem solving, while Chanal (Chanal, 2008) describes it as a way for businesses to obtain access to outside skills and experiences. Others have simply identified crowdsourcing as the means for allowing people to complete simple tasks that machines, algorithms and computers are still unable to perform (Kittur et al., 2013; Saxton et al., 2013). In this context, Estellés-Arolas and González-Ladrón offer a more exhaustive definition of crowdsourcing that integrates many of these perspectives. They defined crowdsourcing to be a form of participative online activity (Estellés-Arolas & González-Ladrón-de-Guevara, 2012) where:

> *'An individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task.' (Estellés-Arolas & González-Ladrón-de-Guevara, 2012)*

Notably, Estellés-Arolas and González-Ladrón suggest that crowdsroucing should always entail 'mutual benefits'. Workers will receive extrinsic (be it financial) and/or intrinsic (i.e. self-esteem) reward (Minder & Bernstein, 2012) for solving tasks, whereas requesters will take advantage of what workers have brought to the venture.

Requesters may use their own quality control measures to evaluate the quality of the obtained responses before rewarding workers for their effort. This is often the case when crowdsourcing uncommon tasks via crowdsourcing platforms or demanding sophisticated quality control measures. Alternatively, requesters may rely solely on the crowdsourcing platform quality control measures to accept tasks and reward workers for their responses.

Depending on the nature of the task and the nature of the reward, four different models have emerged to map crowdsourcing tasks to workers. In no particular order, the first model is the *Marketplace model* described by (Ipeirotis, 2010) focuses on tasks of low complexity that require no specialised skills, such as image labelling or text extraction, for which the requester typically demands three or more responses (e.g., three tags per image) for a small reward. All workers who satisfy the requester's acceptance criteria will receive the predefined reward. Such model is implemented in many crowdsourcing platforms, such as Amazon MTurk and Figure Eight (formerly known as CrowdFlower). Second is the *Contest model* described by (Cavallo & Jain, 2012) focuses more on creative tasks that require talents and special skills, like designing a logo or building a prototype for an idea, for which the requester typically demands one best solution for a fixed reward. All workers will compete for the reward but only those (typically one) with the best accepted solution will receive the agreed reward. Prominent examples are InnoCentive and 99designs. Third is the *Auction model* presented by (Satzger et al., 2013) focuses on complex tasks that require special expertise, such as building a website or writing a blog post. In this model, the requester defines the task and workers bid for the task. The task will then be assigned to the winning bidder who will receive the mutually agreed reward if their solution meets requester's acceptance criteria. One famous platform that implements the auction model is UpWork. And finally, the *Volunteering model* (Mao et al., 2013; Hosseini et al., 2014) focuses on unstructured and organic tasks, like sharing knowledge by voluntarily contributing to a topic in Wikipedia or answering an engineering question on StackOverflow. Such contribution is driven by reasons beyond financial, such as knowledge sharing and for the love of the community.

In the first three crowdsourcing models, unlike the volunteering model, requesters and crowdsourcing platforms typically have predetermined quality control measures to verify workers responses. This is essential to count for workers anonymity, skills discrepancies, and motivations (Minder & Bernstein, 2012). Whereas in the volunteering model responses' validation follows trials and errors methods (e.g., by the person who asked the question or others with similar engineering problem in StackOverflow) or opens for judgements and interpretation of unrelated people (e.g., readers in Wikipedia). The quality of crowdsourcing output depends on multiple factors including workers selection process, responses aggregation method, and tasks creation process governance. A large volume of published studies established that quality measures and process governance in existing crowdsourcing platforms and their defence against malicious attacks (e.g., cheating, stealing sensitive data) are insufficient (Kritikos et al., 2013). There has, also been growing concerns over the unintended consequences of inadequate quality control measures, including privacy risks, additional unnecessary costs, and malicious attacks (Walter S Lasecki, Teevan, et al., 2014; Naroditskiy et al., 2014; Abeliuk & Masuda, 2014). These concerns are still under explored (Hu et al., 2017) by researchers, and due to the various nature of crowdsourcing tasks, platform providers may not have adequate information to address them appropriately. In marketplace platforms (e.g., Amazon MTurk) requesters are fully responsible for designing and creating their tasks of any nature (e.g., image tagging, audio transcription). As such, the platform that execute these tasks is typically unaware whether such tasks raise any privacy or other related issues. Thus, beside designing and creating their tasks, requesters are encouraged to develop and carry out relevant quality controls externally to check whether workers responses meet expected levels of quality.

On the other hand, in platforms that specialise in one or a few types of tasks, like *99designs* that focus on graphic design, issues like, intellectually property are identified and well addressed.

Despite the increasing importance and benefits of crowdsourcing services, many quality issues have not been addressed yet. Ideally, crowdsourcing platforms should offer requesters relevant and effective quality control methods, and tools to configure and adjust these methods to meet their criteria. Instead, crowdsourcing platforms (marketplace in particular) offer requesters narrow quality controls that focus on things like worker reputation. The lack of clear crowdsourcing quality measures and restricted control over them to promptly respond to quality issues often lead to additional cost (Stol & Fitzgerald, 2014). Furthermore, crowd-powered systems are vulnerable to active attacks (Walter S Lasecki, Teevan, et al., 2014), and the lack of adequate quality controls attract malicious workers to cause more harm (Gadiraju, Kawase, et al., 2015).

Artificial Intelligence (AI), on the other hand, have gained a lot of benefits from Crowdsourcing. Since the emergence of AI, considerable effort has been invested to simulate human behaviour. At its very early stage in 1950 Alan Turing envisioned AI and questioned "Can a machine be made to be super-critical?" He introduced human and digital computers model where computable problems were computable by human means. Later on (Licklider, 1960) teamed up human and computer in order to achieve a machine that not only performs arithmetic operations but also facilitates formulative thinking. Since then and until now, despite the rapid and wide advances in the domain of AI, the human contribution is still a key for its success. Which led to the rise of Human Computation approach that was first introduced in a PhD thesis in 2005 by (Von Ahn, 2005). Ahn developed games that are geared by human to solve problems that cannot be solved by computers back then, such as image labelling, object tagging and text extraction from arbitrary images. Ahn defined Human Computations as *"a paradigm for utilizing human processing power to solve problems that computers cannot yet solve"*, and was followed by a range of various research (Yang et al., 2008; Chan et al., 2009; Quinn & Bederson, n.d.; Schall et al., 2008) confirming the computational problems and presenting human computation as the interim solution. Before Crowdsourcing, Human Computations required building up specific communities to sustain it. However, at the emergence of Crowdsourcing Human Computation grew faster by leveraging greater communities of online workers that can be trained and recruited any time for any computational task for the fraction of cost and time. For example, Lasecki et al. introduced a video coding system powered by online recruited crowd workers (Walter S Lasecki, Gordon, et al., 2014), and reviewed the design of crowd systems that perform complex real-time tasks (Walter S. Lasecki, Homan, & Bigham, 2014) taking advantage of various available crowdsourcing technologies.

In my research, I recognise crowdsourcing as “a method to outsource any task to an undefined list of participants, which can be performed remotely and submitted online to support or solve current problems for higher quality and lower costs than experts and current state-of-the-art automated methods”.

1.1 Problem statement

While crowdsourcing has traditionally introduced challenging computer’s tasks to be solved by undefined network of people, many problems are currently expensive and time consuming when quality crowd input is inevitable (Fung, 2011; Sheng et al., 2008; Kittur et al., 2013). For instance, the accurate localisation and annotation of objects in large-scale image data sets is still difficult for computers due to the variety of object types and shapes, or image quality, but is essential for AI-powered real-world mobile applications. Besides, preparing row data to obtain labels (via crowdsourcing) for training such AI applications can become considerably expensive (Sheng et al., 2008) Furthermore, applications for assessing intelligibility or producing experts like ratings are still not achievable following AI approaches independently. Whereas many people find such task easy to complete, but they may lack the means and motivation to put enough effort and generate accurate responses (Liang et al., 2018), which tasks design should address. Typically, requesters demand extra crowd responses per task, or offer higher payment while looking for quality responses. However, such factors alone do not guarantee higher quality responses, and even when they do the overall cost is usually too high. Which makes crowdsourcing methods unsuitable for the use in day-to-day real world mobile applications. That being said, crowdsourcing tasks if well-designed result in higher quality input with low cost and better crowdsourcing effort. This thesis demonstrates that by solving some issues of task design, problems like these can be resolved effectively and efficiently. Which opens the doors towards building economical and robust crowd-powered solutions that could not, otherwise operate solely by AI approaches.

This thesis sets out to investigate and propose solutions to two critical limitations of current crowdsourcing approaches in supporting personal mobile systems: maintaining low completion-time and cost when processing data, and obtaining experts like quality crowd responses without increasing the overall cost. The approach described in the thesis advances existing knowledge in the fields of human computation and crowdsourcing systems. It opens up new opportunities for systems in various domains such as, but not limited to, self-monitoring and self-directed practices where human judgement is inevitable. During this course of research, I focused on two types of data: speech audio in natural settings, and visual eye tracking recordings in natural settings too. The effectiveness of my approach is demonstrated in two separate case studies by two large working crowdsourcing applications that I developed to leverage the power of the crowd and solve key, but underexplored, problems in crowdsourced mobile systems.

1.2 Research questions

This research aims to address the design of crowdsourcing speech and visual solutions that maximise the quality of crowd responses, while minimising cost and crowd effort.

Q1: What is the implication of self-verification as a quality control method on improving accuracy with no additional costs?

Imposed standard quality measures often result in expelling, not only unsatisfactory, but sometimes quality honest workers from the job and consequent tasks. For example, a worker provides quality responses to most of the tasks but fails on some of the quality check tasks (i.e., gold standards), their effort will not be compensated, and their work will often be rejected. Such outcome will demotivate honest workers and often negatively influence their performance in consequent tasks (Mcinnis et al., 2016). Mcinnis et al. also suggested that such workers are more likely to refuse more work from these requesters. with Consequently, it is essential to understand how quality measures can be designed, not only to ensure quality responses but retain workers and guarantee fair compensation. This thesis investigates the implication of applying self-verification quality measure that is based on trust to motivate workers to complete their tasks with higher accuracy but no additional cost.

Q2: How to design crowdsourcing tasks to achieve expert-comparable input when working with speech and visual data?

Since Crowd workers perform tasks in isolation, it is essential to curate and aggregate their responses in order to achieve a required level of quality. Considering workers motivation is usually financial, workers often multitask (Chandler et al., 2014) and tend to complete tasks as quick as possible, which often lead to low data quality just enough to be rewarded (Meade &

Craig, 2012). Crowd tasks should be designed in a way that ensures higher engagement and stronger commitment. However, it is often unclear to requesters how to design their tasks and achieve quality responses comparable to that of an expert. This thesis gives guidelines on how to design two types of tasks, one for speech analysis and another for eye tracking, to achieve expert-comparable responses.

Q3: How to develop low-cost crowd-powered solutions that directly benefit end-users?

Currently, crowdsourcing platforms are greatly utilised by product or service providers seeking to improve their offering. Whereas the benefits of utilising crowdsourcing in end-users applications, such as VizWiz (Bigham, White, et al., 2010) (an app that enables visually impaired and blind people make sense of their surroundings), are still underexplored. While automated approaches are far from supporting many applications like VizWiz, human computation in the form of crowdsourcing holds the promise to do so. However, crowdsourcing, can result in higher costs than automated approaches, especially when quality is a concern. Therefore, this research investigates and presents tasks design that ensure not only expert-comparable quality, but low cost to benefit wider group of users via two crowdsourcing applications.

1.3 Methodology

To achieve the aim of this study and address the research questions, I followed the case study methodology. This method enabled me to empirically demonstrate and evaluate various crowd task's design variables for quality through real-world mobile crowdsourcing systems. Throughout two case studies, I employed crowdsourcing techniques to i) derive insights into quality in crowdsourcing the collection and analysis of real-world data, and to ii) examine the influence of task's design and workers' self-verification on achieving the desired quality and valuable feedback.

The first case study, described in chapter 3 focuses on people with Parkinson's (PwP's) who developed speech disorders (e.g., stammering, dysarthria), and addresses the research questions by developing, deploying, and evaluating a novel crowd-powered mobile app. The mobile app was designed to enable PwP's monitor and manage their speech condition via in-app speech exercising tools and crowd generated intelligibility ratings. The complexity of this work resides in processing the data (speech), which is normally processed and rated by a Speech and Language therapist. Participants with Parkinson's were recruited through Parkinson's UK research and support charity, and were interviewed by Róisín Mcnaney, my research collaborator, before and after deploying the crowdsourcing solution. The initial (pre- app- deployment) interviews were used to evaluate the quality of crowd judgements in comparison to users' perception of their speech condition and against ratings of a Speech and Language therapist. The second set of interviews (post- app- deployment) offered insights into the value of such crowd-powered solutions, particularly the crowd generated feedback on intelligibility to people with speech disorder. Chapter 3 demonstrates the feasibility of this approach and reports on various task design considerations to achieve expert-like quality intelligibility ratings by anonymous naïve listers (the crowd workers).

The second case study, described in chapter 4 focuses on technologies that drastically fail when used in natural settings (e.g., outdoor), and addresses the research questions via a novel crowd-powered mobile eye-tracking system. The system was used to evaluate crowdsourcing approaches in delivering quality eye-tracking technology and overcoming its major challenges. Mobile eye tracking technologies are essential in many usability studies (Goldberg & Wichansky, 2003), internet of things (Klaib et al., 2019), and potential day-to-day solutions (Krafka et al., 2016), but their limitations and cost keep them out of reach of the vast majority of people. When recording an eye-tracking session, the eye tracker generates large quantity of captured images of various quality (e.g., blurred, in motion, partially obscured target) which are difficult to process automatically and costly to annotate or label by crowd workers. The high cost of robustly crowdsourcing eye tracking captured images can undermine the benefits of using crowdsourcing. Consequently, this case study examined various methods to collate quality crowd responses and produce quality and robust eye tracking experience in natural settings for a small cost. Chapter 4 details quantitative results of this research. It demonstrates the accuracy of the crowd approach in comparison to five current state-of-the-art algorithms when localising the centre of eye pupil. The proposed solution is evaluated through heterogeneous pre-annotated eye tracking data collected in the wild (Tonsen et al., 2016). The results suggest the proposed crowdsourcing solution outperforms all five algorithms under all conditions (e.g., outdoor, wearing eye make-up or spectacles) for a fraction of the price of using commercial eye trackers. Finally, to demonstrate the usability of this solution and the crowdsourcing quality it yields when used under natural settings, I recruited participants through Newcastle University to use the developed eye tracker during lunch purchase activity. The crowd then helped achieving accurate eye tracking data, and labelled all objects that a user fixated upon while selecting what to have for lunch. The quality of this approach is described in-depth in chapter 4.

1.4 Terminology

This thesis uses the following terminology across its content:

Crowdsourcing: A form of participative online activity responding to an open call for work.

Requester: An individual, an institution, a non-profit organization, or company proposes the undertaking of tasks.

Worker: A person carrying out small tasks as part of their paid or voluntary job.

Task: Also known as Microtask or Human Intelligence Task (HIT), the smallest piece of work to be done by a worker, which is used in a well-defined and structured process.

Human Computation: Human input to resolve certain computational tasks to enable the machine to complete its function.

1.5 Summary of contribution

This research contributes to the field of HCI, and particularly the Crowdsourcing community in several ways, including crowdsourcing tasks design consideration that support quality collection and translation of speech as well as visual perception data in real-world scenarios. It also delivers two Crowdsourcing applications that capture real-world data and manage the entire crowdsourcing workflow for the analysis of data and the provision of meaningful feedback to data owners. Through these two crowdsourcing applications, this research establishes the capability of the crowd to provide expert-like quality to be used in complex domains (e.g. impaired speech assessment). Furthermore, this research presents and evaluates a new quality control method based on workers trust and self-verification, which encourages workers to improve their responses and return to complete more tasks. Finally, this research also contributes to the research community of eye tracking technologies by delivering large quantity of crowd annotated eye tracking data collected in real-world environments. The data contains accurate pupil and calibration marker localised centre as well as descriptive labels on what participants gazed on during the study. Likewise, this work also delivers annotated in-the-wild speech data with expert-comparable crowd-measured speech intelligibility.

1.6 Thesis outline

Chapter 2 begins by outlining the growing research in this space, which focuses mainly on the quality of crowd responses and the implications of tasks design on participation and final results. The remaining of this thesis is organized in the following manner:

Chapter 3 introduces techniques for crowdsourcing the collection and translation of impaired speech and voice data from and to end users, maintaining expert-comparable crowd feedback. This has been implemented in the first case study, Speeching (Mcnaney et al., 2016) a crowd-powered self-monitoring system that enables users to practice and manage their speech and voice without resorting to speech and language therapists. Further in this chapter, I evaluate the quality of crowd responses in comparison to experts and how the end users value the provision of crowd feedback, using Speeching.

Chapter 4 describes a process of designing and implementing a crowd-powered mobile low costs eye tracker system, which streamlines the eye-tracking data collection and processing via the crowd. This study demonstrates how crowdsourcing methods could overcome eye tracking real-world challenges and offer research community and other potential users a low-cost and robust eye tracker. The quality (in comparison to the state-of-the-art algorithms and to expert annotators) and robustness of this approach, in addition to the affordability and usability of the solution are also evaluated in this chapter. Furthermore, this chapter demonstrates how workers are capable to self-verify their responses and yield much better accuracy for no additional costs. Such approach helped reduce costs by demanding fewer crowd responses, in fact in this study, only one worker response is required to accurately complete each crowd task.

Chapter 5 I finally conclude by summarising the research contribution and the importance of this work, and end by discussing limitations as well as directions for future research.

2 BACKGROUND

Almost in all computing disciplines, including software engineering (Nordin et al., 2017; Terragni et al., 2020), user interfaces (Riegler & Holzmann, 2018) and e-commerce services (Sari et al., 2018), quality is a major factor to measure success. In addition to crowdsourcing quality research, some quality research from other computing disciplines can be leveraged to control quality in crowdsourcing solutions. For example, measuring software testability quality (Terragni et al., 2020) can be utilised in software testing crowdsourcing tasks. Having said that, the unique characteristics of crowdsourcing demand overcoming novel crowdsourcing quality issues that are rising, and perhaps consider including requesters and workers more in the execution of crowd tasks, like in data preparation and responses evaluation.

The challenge in achieving quality output lies in various factors. For example, workers who execute crowdsourcing tasks are heterogeneous with various skills level (Gadiraju, Fetahu, et al., 2017) and their experience is sometime inadequate or irrelevant to successfully complete their tasks (Minder & Bernstein, 2012; Malone et al., 2010). Depending on their demographics (e.g., education, location) workers may be biased (Difallah et al., 2018), and have various interests and motives (Goncalves et al., 2015; Posch et al., 2017; Eickhoff & de Vries, 2013). To date, crowdsourcing has been subjected to malicious workers (e.g., cheating, or breaching privacy) and activities (e.g., posting online fake reviews) (Gadiraju, Kawase, et al., 2015). Thus, poorly defined and designed tasks (e.g., ambiguous description, unfriendly user interface) confuse workers and alienate some of them, which result in reduced quality outcomes and fewer workers to participate.

Requesters are expected to make a variety of task design decisions when creating their crowdsourcing tasks. Appropriate task design is a key to achieve quality output and higher inter-rater reliability between crowd workers (Garcia-Molina et al., 2016). A design taxonomy was defined by (Catallo & Martinenghi, 2017) based on four design considerations described as *What is the nature of the task to be resolved*; *Who is going to resolve it*; *Why would anyone participate in resolving it*; and *How to execute these tasks*. Such classification summarises the major dimensions that are involved in designing clear and quality crowdsourcing tasks. The *What* dimension shall define the following task's properties:

– Task type: requesters should clarify the task type, is it image labelling, survey, text extraction, or something else. Task type could be anything that machines cannot complete yet.

– Task features: requesters should clearly communicate to the crowd the required skills (e.g., fluent English writing skills), the definition of complete (e.g., workers may give up any time and get partial reward), and the task significance (e.g., this task supports people with speech disorder). As well, requesters better identify the required effort to complete their tasks and communicate this back to the crowd (e.g., you will listen to three "2-minute" audio recordings). This all ensure transparency between workers and requesters, which consequently increases workers' interest.

– Task output: A task could be deterministic, it accepts one correct answer (e.g., what object is this?), or nondeterministic, it accepts one or more correct answers (e.g., paraphrase this paragraph).

The *Who* dimension refers to the people (workers) who will execute the tasks and their characteristics, such as diversity (e.g., gender, location, qualifications), anonymity (depending on the crowdsourcing platform, requesters and workers may hide their identity including their past performance), and hierarchy. Although hierarchy in most of crowdsourcing solutions is flat (i.e., peer workers with no hierarchical roles), some solutions may benefit from two- or more- level hierarchy (e.g., assign highly reliable workers as group leaders for workers with lower reliability, perhaps to evaluate their work or encourage them deliver higher quality output).

Finally, the *Why* dimension is concerned with the motive that derives participation in crowdsourcing, which are identified as extrinsic (e.g., financial reward) or intrinsic (e.g., killing time). The identification of workers motivation is key for successful task design (Malone et al., 2011; Antin & Shaw, 2012).

The following sections of this chapter offer further analysis of these dimensions in relation to my work. Chapter 3 and 4 of this thesis also demonstrate a practical example of how these dimensions were implemented when designing the tasks that led to high quality output.

Quality control methods, on the other hand where categorised by (Allahbakhsh et al., 2013) into two groups based on their execution time: those that take place at design-time and others at runtime. Design-time approach is concerned about key elements that clarify a task and its objectives and encourage participation prior to executing the crowd task. Such as determining the compensation criteria; writing clear description; or selecting workers based on their ranking. Whereas quality control at runtime approach is concerned about governing the execution of tasks and responses, as well as monitoring the crowd performance. Such as comparing responses to ground truth (widely used method); peer review (Hansen et al., 2013; Huang & Fu, 2013); accepting majority consensus (most common when aggregating workers responses); or providing real-time support to workers to increase quality. Both runtime and design-time approaches are not mutually exclusive, and a task may implement elements of both approaches to aim for higher quality outcomes.

Addressing quality issues in crowdsourcing demands thorough understanding of the quality control methods, as well as the other factors that influence the quality of the output (e.g., task design, demographics, motive).

This chapter begins by summarising crowdsourcing technologies into three classifications: crowd work, crowd innovation and volunteer-based crowdsourcing. Then it analyses factors that influence the quality of task design, such as, workers' motivations, the associated ethical dilemma, participation, and other task design factors and to demonstrate their influence on the quality of crowdsourcing output.

2.1 Crowdsourcing technologies

Since crowdsourcing as a concept has grown in popularity, the last decade has seen an enormous amount of development of technological platforms that support various forms of crowdsourcing. This section is organised around three main areas: crowd work, crowd innovation, and volunteer-based crowd work.

2.1.1 Crowd work

Crowd work refers to an emerging industry where workers receive a financial reward for completing tasks and work. In theory, crowd work platforms enable anyone with Internet connection is able to complete tasks, which often can be completed in seconds, and be paid a fractional amount for completing tasks accurately (Kittur, 2010; Kittur et al., 2008). A number of crowdsourcing platforms for crowd work have emerged in recent years. Two widely used and studied crowdsourcing marketplace platforms are Amazon Mechanical Turk (MTurk) and Figure Eight (Alonso & Lease, 2011). In such platforms requesters are expected to design their tasks, splitting large tasks into micro tasks, often known as human intelligent tasks (HITs), in order for crowd of workers to complete. Although task design is important (Gadiraju, Yang, et al., 2017; Jain et al., 2017), it is also challenging (Alonso & Mizzaro, 2012; Bragg et al., 2018) for requesters to accomplish, for many reasons. Such as, determining the output type, adequately describing the task and convey it to workers, and fine-tuning quality control methods. A process that is iterative and costly.

Tasks on crowdsourcing platforms are usually conducted to solve computationally complex problems (Gurari & Grauman, 2016; Kaspar et al., 2018; Cooper et al., 2010), generating large amounts of data (Dergousoff & Mandryk, 2015), and performing large-scale experiments (Komarov et al., 2013; Alallah et al., 2018). Crowdsourcing platforms have opened up wide opportunities to researchers who found Crowdsourcing an attractive alternative to laboratory-based studies— not only giving researchers access to potentially enormous pools of diverse participants at reduced operational costs (Vaish et al., 2017); but also allowing them to streamline the experimental analysis by cutting down development and administrative time. Researchers can economically design analysis tasks such as audio transcription, image labelling, and video coding in such a way as they can be done by the crowd in short time. Some researchers, on the other hand, have developed purpose-built tools to enable others to make more effective use of these platforms. For example, Lasecki *et al.* (Walter S Lasecki, Gordon, et al., 2014) developed *Glance*, a tool to code and analyse behavioural events in large video datasets. Glance enables researchers to design analysis tasks for the crowd by simply uploading videos and proposing questions to workers about events that may occur in their video and get (nearly) immediate responses from the crowd for a small price. However, Glance research focused primarily on maintaining context to overcome high workers turnover but neglected advanced quality measures as well as the total costs. Glance suffered from malicious workers (e.g., careless answers, cheating) as quality controls were not sufficient to deal with malicious workers at runtime. While this thesis also explores crowd data analysis techniques, it focuses primarily on achieving expert's like quality crowd responses for much lower cost.

2.1.2 Crowd innovation

Another popular area of growth in internet-mediated crowdsourcing has been in firms and organisations setting competitions or challenges to gather creative and innovative solutions and ideas. In examples like InnoCentive, a solutions provider, the crowd focus on generating business, social, technical, policy and scientific ideas and solutions. The crowd members, individually or among agreed-upon team members, submit their ideas and solutions for the requester to choose from (Saxton et al., 2013). Crowd innovation has also been used to generate a large number of creative products and graphic designs. For example, clothing company Threadless has created an online designers' community where crowd workers (known as community members on Threadless) compete in producing the best T-shirt design for themes requested by Threadless in an on-going open call for design submission. Once designs are submitted the community members vote, score and flag T-shirt with "I would buy it" in order to find the most popular designs (Brabham, 2008; Saxton et al., 2013).

Unlike the micro-tasks markets where the crowd workers receive small amounts of money, usually a few cents per tasks (Brabham, 2008; Norcie, 2011), both InnoCentive and Threadless offer large rewards for accepted ideas (Saxton et al., 2013; Brabham, 2008). In Threadless, winners receive $1500 USD in cash plus $500 USD Threadless gift voucher (Brabham, 2008). Rewards in InnoCentive ranging from tens of thousands to hundreds of thousands of USD. In 2008 a challenge placed on InnoCentive to identify a biomarker for measuring progression in a neurodegenerative disease, amyotrophic lateral sclerosis (ALS), offered one million USD as a reward for those who solve it (Brabham, 2008).

Arguably, the large reward to the accepted innovative solution does not justify the compensation refusal to other workers who participated and contributed to find a solution. Perhaps many refused solutions once influenced the winning one. Similarly, tight quality measures, vague instructions and or poor task design in paid crowd work markets contribute to compensation refusal, even to honest workers (Mcinnis et al., 2016). As such, reward refusal may contribute to discourage online crowd workers from taking tasks from requesters seen as unjust or accurately completing them (Johnstone et al., 2018).

2.1.3 Volunteer-based crowd work

Volunteer-based crowd work refers to crowdsourcing but in the absence of financial incentives. It focuses more on people's intrinsic motivation (i.e., entertainment, self-esteem, knowledge sharing). While some voluntary crowdsourcing is unstructured and organic (thus requiring skilful members), like in Wikipedia and open-source software projects, new organised crowdsourcing platforms have emerged to provide a structured environment that connects members with micro tasks. One prominent example is Zooniverse[1]—an online citizen science platform that links scientists seeking public (often referred to as citizen scientists) involvement in inquiry and unearthing new scientific knowledge. One such example of a Zooniverse project is Cell Slider (Candido dos Reis et al., 2015), where any member of the public can participate and receive training in order to then identify the presence of cancer cells and their number in images presented to them online. Another example is Galaxy Zoo (Lintott et al., 2008), a project which to date has involved volunteers in classifying millions of galaxies morphologically. Volunteers in unpaid crowd work are not driven by financial reward, but rather different objectives and interest in the topics (Mao et al., 2013; Balicki et al., 2014); suggesting the volunteers may have more knowledge and interest in the topic than most workers in paid crowd work. However, this has come at the cost of performance in comparison to paid crowd workers, with volunteers scoring lower precision as reported in (Mao et al., 2013). This is, possibly, due to different objectives, motivations, and commitments to the accurate completion of tasks. Unlike unpaid crowd work, in paid crowd work workers are bounded by the task design, the necessary financial reward as well as the crowdsourcing platform standards, where low scoring workers may eventually be prevented from completing any tasks (Shaw et al., 2011; Harris, 2011; Mason & Watts, 2009).

This leads us to another form of unpaid crowd work that utilise gamification techniques. Gamification refers to the addition of game design elements in a non-gaming context in order to create joyful user experiences, motivate participation, and increase engagement and loyalty (Deterding et al., 2011). Prior research shows the effectiveness of gamification in harnessing wider public participation in complex tasks, such as OCR results verification (Jovian, 2011), geographic data collection (Odobasic et al., 2013) and landmarks identification (Bockes et al., 2015). However, gamification typically requires special game development to answer specific research questions. To address this limitation, Dergousoff *et al* (Dergousoff & Mandryk, 2015) introduced a model based on leveraging freemium models, where players receive free in-games rewards (i.e. unlock premium features) by occasionally completing specially designed tasks embedded in popular games. Further, Dergousoff *et al* evaluated their approach and reported

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

anonymous players performed equally well in motor tasks in compare to participants in the controlled lab environments, though anonymous players performed worse in cognitive tasks. This could be attributed down to task design or the motivation behind undertaking such task, which the rest of this chapter thoroughly reviews.

This thesis and the research demonstrated in it focus on Crowd Work technologies. All the work was carried out via Amazon MTurk and Crowdflower (now known as Figure Eight) platforms.

2.2 Worker Motivations

The above sections not only highlight the different ways in which crowdsourcing has been applied and grown in popularity in recent years, but also indicate a range of different ways people become motivated to perform crowd 'work'. Ryan and Deci in their development of Self Determination Theory (Ryan & Deci, 2000) split motivation into two categories, intrinsic and extrinsic, based on the individual's goals and attitudes that cause an action. Research on crowdsourcing has explored the distinction between intrinsic and extrinsic factors that motivate people to participate (Kaufmann & Veit, 2011; Leimeister et al., 2009; Zheng et al., 2011). In 2011, Kaufmann *et al.* (Kaufmann & Veit, 2011) offered a theoretical classification to the intrinsic and extrinsic motivations in crowdsourcing environments. They broke down intrinsic motivation into two groups - community based motivation and enjoyment - and the extrinsic motivation into three groups: (i) immediate payoff; (ii) delayed payoff and (iii) social motivation.

Intrinsic motivation refers to performing tasks that are interesting, enjoyable and self-satisfactory (Naderi et al., 2014; Ryan & Deci, 2000). Kaufmann *et al.* (Kaufmann & Veit, 2011) further explained how 'community based motivation' comes from a love of the community within which users participate, and this becomes a major driver of motivation to participate in a crowd activity (see also (Brabham, 2010; Gerber & Hui, 2013)). On the other hand, 'enjoyment' based motivations are where crowd workers perform the tasks for fun and to fill in their free times (Ke & Zhang, 2008; Zheng et al., 2011).

In contrast, extrinsic motivations refers to tasks that are completed for financial rewards or social reputation (Naderi et al., 2014). Kaufmann *et al.* (Kaufmann & Veit, 2011) justified how 'immediate payoff's' refer to workers being motivated through the receipt of payments as soon as they complete a task (e.g., as in (Rogstadius et al., 2011)). On the other hand, 'Delayed payoffs' is another form of extrinsic motivation where workers complete tasks to build a portfolio and gain extra skills important for their career. Whereas finally, ‘social motivations’ come from the extrinsic desire to gain publicity and recognition for contributions among a community, and to grow one’s own reputation (e.g., as in (Kittur et al., 2013)).

However, (Zheng et al., 2011; Leimeister et al., 2009) have argued that public recognition and personal benefits are far more influential on the motivation to perform crowd work than monetary compensation. Likewise, (Kaufmann & Veit, 2011) conclude, from their study of 431 crowd workers on MTurk, that very often intrinsic motivation overtakes extrinsic financial motivation. In addition, Rogstadius *et el.* (Rogstadius et al., 2011) investigated the relationship between motivation and task performance in crowdsourcing markets and made two key conclusions: first, that higher paid tasks lead to faster completion but not necessary better quality; while increasing the intrinsic motivation factors lead to better work quality.

Having said that, my research focuses on speech and eye tracking data collected in real-world environments, which normally yield large quantity of audio recordings and eye tracking images to crowdsource. This type of data usually includes noise that is costly and time consuming to crowdsource. Thus, this research proposes methods to overcome such hurdle, and increase quality participation in crowdsourcing by the self-verification method evaluated in chapter 4, while maintaining costs to minimum.

2.2.1 Increasing participation in crowdsourcing

While there has been research conducted on the motivations of crowd workers, there have also been many developments in recent years to adapt crowdsourcing approaches to increase the level of participation in such activities. One of the popular examples of this is in how gamification mechanics have been used at length in crowdsourced citizen science derived by enjoyment and community-based motivations. While Gamification is a key motivation method, it is also a key factor to increase participation—as is illustrated in games such as Galaxy Zoo and FoldIt. In FoldIt (Cooper et al., 2010) players receive scores to open further game levels depending on how well they fold images of proteins (puzzles). In addition to gamifications, GalaxyZoo offers top volunteers who successfully classify galaxies in numerous images a public community recognition (Lintott et al., 2008; Eveleigh et al., 2014; Prestopnik & Crowston, 2012).

Another strategy to increase participation in crowdsourcing targets people with mobile devices (e.g., phones, tablets). (Allen, 2015) explored the use of crowdsourcing techniques to collect data while on the go. He introduced a mobile app called *FixTheCity*. Here, those with the app installed receive notifications or questions about the area where they are in at that particular moment and respond with an answer. For example, they receive a question about whether a streetlight, close to where they are, is broken or not. The collected data is then used by the local authority for neighbourhood maintenance scheduling. Similarly, (Ching et al., 2012) explored how crowdsourcing via smart phones could help in collecting data about bus location and crowding, as well as the riders' satisfactions and experiences to improve the public transport service. Furthermore, the same collected data about bus locations was then used to build the first geo-coded bus routes for the city main two bus lines. Going further, Elaine and Chris (Massung & Preist, 2013) developed Close The Door (CTD), which sought to reduce carbon emissions by encouraging shop owners to keep their doors shut when running their air conditioning. To achieve this, a mobile app with a map was developed to allow users (the crowd) collect and submit data about any shop with open or closed doors by tagging them on the map. The collected data was initially used to reward shops that keep doors shut, and was later used chastise those who do not.

Furthermore, (von Ahn, 2013) took advantage of providing free online foreign languages learning service to increase participation in translating the web, and introduced Duolingo—Embedding crowdsourcing complex foreign languages translation tasks into an online learning environment. I argue that systems, which directly benefit end-users (i.e., Duolingo), open up new sustainable opportunities to complete crowdsourcing tasks.

Building on that, this PhD merges between the aforementioned techniques and introduces new opportunities where there are two separate crowdsourcing groups (users and supporters), both benefiting from using and supporting the system. Whereas users (people with Parkinson's in chapter 3) collect data and answer initial questions, supporters (traditional crowd workers) analyse and make sense of the data. Supporters are then rewarded (financially) and the data is translated and fed back to users for their personal benefits.

2.3 Quality in Crowdsourcing

Quality in crowdsourcing is influenced by three crowdsourcing core elements: data, users, and task design.

Data: is a task's content (the *input*) that is required by workers to carry out that task. It is also the result of completing the task (the *output*). For example, in "Extract text from image" tasks, images are the input provided by a requester, while the extracted text by workers is the output.

However, achieving high quality output data is challenging and is still a barrier for mass adoption of crowdsourcing solution (Niu et al., 2019; Kittur et al., 2013)

Users: In crowdsourcing platforms users are divided into two categories, *Requesters* and *Workers*. Where *requesters* represent individuals or businesses who prepare crowdsource tasks and data, configure crowdsource work, validate output quality and reward successful workers. While *workers* represent a workforce (often called crowd pool) that is available at any time to complete requested tasks on a particular crowdsourcing platform. Integrity and good interaction skills are two qualities of a successful requester (Irani & Silberman, 2013). From the abundant literature, one could distinguish between good and bad workers by understanding three workers attributes: behaviour, experience, and demographic and their influence on the output quality (analysed thoroughly later in this section).

Task design: Task's quality is influenced by its design, which reflect on workers participation level and the quality of the outputs. This chapter offers an analysis of what make a good task design to achieve quality results, including tasks complexity, workflow, user interface and incentives.

2.3.1 Data quality

Crowdsourcing does not guarantee quality results when input data is of poor quality (Khazankin et al., 2012). For example, workers may struggle to identify objects in image tagging tasks, especially when images are of low resolution or objects are partially presented in the provided images. Thus, it is vital that requesters improve the quality of their data, and perhaps they start by crowdsourcing a data subset and evaluate workers responses before crowdsourcing the entire dataset (Brambilla et al., 2015). A few strategies can be followed to improve input and output data:

Input data cleansing: To increase the chance of procuring quality output in crowdsourcing, input data should be of good quality. Crowdsourcing platforms do not take responsibility for the quality of its workers responses when the input data is inaccurate or noisy, but delegate the full responsibility to requesters instead (Khazankin et al., 2012). Since workers in marketplace platforms are, typically, after financial reward, low data quality input (e.g., unintelligible or too noisy audio to transcribe) reduces their chances to successfully complete the task, and potentially miss out on the reward (Schulze et al., 2013). To overcome such hurdle, requesters should pre-process their data, when possible. For example, (Bigham, White, et al., 2010) introduced VizWiz, a crowd-powered application to enable blind people make sense of their surroundings. VizWiz uses computer vision algorithms to enhance the quality of pictures taken by its users (e.g., sharpening, brightening and colour contrast) before crowdsourcing these pictures. In chapter 4 of this thesis, I use multiscale structural similarity index method to filter images before crowdsourcing them to ensure quality and reduce the overall crowdsourcing costs.

Aggregate workers responses: In a weight-judging competition that took place in a fair, West of England in 1907, (GALTON, 1907) accurately estimated the weight of an Ox by taking the median of 787 guesses from 800 attendees. In crowdsourcing, this demonstrates the Wisdom of Crowds, whereby aggregating multiple workers responses will likely yield accurate final answers (Surowiecki, 2005). I say likely, since in crowdsourcing there are many other factors, described in this chapter that influence the quality of the final aggregated answers. While Galton work suggests aggregating more workers responses lead to more accurate outputs, this comes at a higher cost to requesters. It is, thus, essential that requesters first identify how many worker responses suffice and strategies to identify responses reliability to get accurate and robust output. For example, (Snow et al., 2008) demonstrated how, an average of four non-expert workers can produce (aggregated) labels of high performance equal to that of an expert annotators. In chapter 3 of this thesis, I demonstrate how three workers judging speakers' intelligibility suffice to produce quality judgements equivalent to a Speech and Language therapist. While in chapter 4, only one worker response suffices to accurately identify the centre of an object in mobile captured images. Researchers also investigated machine learning (ML) techniques to increase aggregation accuracy. In crowdsourcing multiple-choice tasks, (Aydin et al., 2014) developed ML techniques to weigh workers reliability, where more reliable workers have higher weight. Workers are considered more reliable, the more accurate (compared to ground truth) answers they provide and the more confident (self-rated after each answer) they are. Answers are then weighed by the weight of the workers who provided them, and the highest weight answer is accepted. This strategy improves the accuracy of selecting the right answers by 15% when compared to the "Majority Decision" (Kuncheva et al., 2003), in which the highest voted answer is accepted.

Filter workers responses: Assuring high quality responses with fewer workers input is essential to keep crowdsourcing costs low. This is one of the major challenges in crowdsourcing due to workers diversity (Quinn & Bederson, 2011). Researchers have studied various methods to improve the quality of workers responses. For example, (Dow et al., 2012) demonstrated how providing workers with direct external feedback to revise their responses produce better output. But this model, is very time consuming and costly if the feedback is provided by experts. Whereas, (Hansen et al., 2013) compared the effectiveness (accuracy) and efficiency (time spent) of two strategies to obtain quality responses: Arbitration and Peer review. Following the arbitration strategy, if different workers do not agree on any possible answer, disagreements go to an arbitrator (for additional responses) — this strategy, however, is limited to closed- ended questions. Whereas following peer review strategy, the responses of one worker is reviewed and revised by another. Their findings suggest, although peer review strategy is more efficient, it is not as effective as arbitration. Other researchers presented a time-series prediction model that analyses workers behavioural patterns while performing their tasks in order to predict the quality of their responses and act upon (Jung et al., 2014).

2.3.2 Users

Abundant research investigated various aspects of the capability of crowd workers. Some explored workers demographic profile (e.g., location, education) to recognise workers skillsets. Studies on AMT reported 90% of workers are based in two countries, the USA (76%) and India (16%) (Difallah et al., 2018; Hara et al., 2019) 88.3% of them had some college education including bachelor degree (Mao et al., 2017). While nearly half of the workforce on Figure Eight are mostly from the US, then Venezuela, Great Britain, India and Canada respectively (Jain et al., 2017). Another emerging crowdsourcing platform that targets primarily research tasks, *Prolific*[2], publishes its workforce demographics online[3], including country of birth and residence as well as education and employment. At the time of writing this thesis, almost 50% of Prolific workers are from the UK and 30% from the US, with 32% of workers hold university degree and almost 30% have attended some college education. Further studies focused on influential factors on workers ability to produce quality results, such as the psychological behaviour, satisfaction, and motivation. Mcinnis et al. suggest the anonymity between requesters and workers creates a level of mistrust, that result in dehumanising the work

[2] https://www.prolific.co/

[3] https://www.prolific.co/demographics/

relationship (Mcinnis et al., 2016). As a result, some workers may feel demotivated (Marlow & Dabbish, 2014), while requesters may use it as an excuse to overlook workers genuine efforts and pay less than what workers deserve (Felstinerf, 2011; Bederson & Quinn, 2011), either way will have a negative impact on output quality.

On the other hand, (Kazai et al., 2011) observed workers behaviour while completing labelling tasks (label pages of digital book) by measuring their completion time, percentage of useful labels and accuracy. Kazai et al. then suggested a worker behaviour taxonomy of five levels. First, *Diligent* workers who take their time (indicated by longer average completion time) to provide highly accurate labels with a high ratio of usefulness. Then *Competent* workers who are very efficient and effective as they complete their tasks quickly with no compromise on accuracy and label usefulness. Third, *Sloppy* workers who are more reward-focused and tend to complete their tasks as quick as possible with little regard to quality. Although *Sloppy* workers may provide low accuracy labels, they may still obtain high fraction of useful labels. While the level before last, *Incompetent*, represents workers who take their time completing their tasks but yield low accuracy, possibly due to the lack of relevant skills or task misunderstanding. Yet, *Incompetent* workers may obtain many useful labels. The lowest level in Kazai et al. taxonomy is *Spammers*, they care not about quality and only concerned about the reward. Although *Spammers* may obtain only some useful labels, their accuracy is very low.

These findings imply correlation between workers' traits and the outcome of various task design. While the accuracy of solving labelling tasks corresponds to workers' behaviour group, the study suggests no correlation between the average time spent to complete the tasks and the accuracy for *Sloppy, Incompetent* and *Spammer* workers. Since this study is not conclusive, an extended research is required to develop a more generic model that includes various worker classification based on their personal trait and according to task type and design.

Understanding and respecting workers' characteristic and psychology is a key for effective crowdsourcing task design to produce quality outputs (Alonso, 2013). (Deng et al., 2016) explored workers experiences in relation to nine mutual values: *access* (equal access to work opportunity); *autonomy* (free to choose their work); *fairness* (equal treatment to workers and requesters); *transparency* (certified and clear work process to workers and requesters); *communication* (facilitate direct communication between workers and requesters); *security* (against scam requesters); *accountability* (taking requesters accountable for their unethical actions); *making an impact* (being recognised by others for your good work); and *dignity* (respecting workers). In their study, 210 crowd workers were recruited from AMT and asked to complete a survey about their interaction experiences with the platform as well as requesters. As a result, Deng et al. concluded the nine values in four crowdsourcing structure: *compensation*, *governance, technology,* and *tasks*, and brought together a set of guidelines for workers, requesters, and crowdsourcing platform developers to govern the workflow and improve the service.

Moreover, crowd workers expectations and performance are significantly affected by their work experiences (i.e. rejected payments, incomplete tasks) prior to completing any upcoming task. (Mcinnis et al., 2016) recognised how recently rejected workers tend to be more cautious, avoiding similar tasks and only taking tasks from requesters with high reputation or who previously rewarded them for their effort. As this may often ensure a reward, yet it limits workers from gaining new experiences from new tasks, and restricts requesters access to a limited pool of available workers. Therefore, it's significant for requesters to understand workers, treat them justly and compensate them for their genuine effort to promote their tasks between wider pool of workers and attract more quality workers.

2.3.3 Crowdsourcing workflow

Although crowdsourcing platforms contain tasks of various types that serve various purposes, the approach's workflow is identical. It begins by identifying the problem; defining the requirements; designing the crowd job and tasks; launching them online via a platform; and finally receiving crowd responses and rewarding successful workers. However, the process varies from three perspectives, the worker, the requester and the crowd job (collection of tasks).

The worker is a registered user on a crowdsourcing platform with one main role, which is to complete tasks requested in an open call fashion. Since all crowd workers are invited to complete jobs via an open call, they will choose any task that suits their skills or interest, or level of reward from a list of jobs available on their platform. However, workers access to certain tasks may be restricted due to tasks constrains, such as minimum ranking or required skills (e.g., Foreign language) (Brabham, 2008).

Given the sheer amount of crowd jobs, some platforms have adopted recommendation algorithms. These algorithms rely on particular factors, such as the overall worker's performance and job acceptance ratio to favour some workers for specific tasks. For example, a worker performing high in a task may be offered more similar tasks than those doing the same task but performing less (Schnitzer et al., 2015; Yuen et al., 2012; Geiger & Schader, 2014). As usual, the worker is still free to choose from the list of available jobs, the recommended ones though. Further studies may be necessary to investigate the role of such recommendation algorithms on limiting workers expectations and restricting their experiences, as well as the impact on job completion time for less recommended tasks. Nevertheless, workers are free to give up on completing any task at any time, and may still receive partial payment, or complete it and submit all answers for the full reward. Finally, the worker will receive a feedback indicating whether their responses passed the requester quality and standard check, and whether they will receive the promised rewards.

The requester, on the other hand, represents a business, an academic institution or an individual with a problem that requires human input to be solved. It is highly important that requesters well understand their problem in order to create the crowd job. They shall collect enough details about the problem, identify the requirements, carefully set up the constraints and define the outcome of their job. For large or complex tasks, requesters are expected to break them into smaller as well as simpler tasks to attract more workers and enable them to work quickly and efficiently. The requester also determines the time limit to complete each task as well as the reward's type (e.g., monetary, social recognition) and value. On receiving workers' responses, the requester may judge spontaneously via a pre-set quality measures (e.g., injected answer-known questions) or manually after analysing received responses to whether reward a worker for their accepted responses.

The crowd job this is a collection of (micro) tasks for workers to contribute and solve a larger problem. Each job goes through three sequential phases, *pre-running*, *running*, and *response aggregation* (Luz et al., 2015). In the *pre-running* course requester may select a predefined job template provided by the crowdsourcing platform, or design their own. During this phase, the crowd job is set up with data and input type, and the quality measures are defined to ensure efficient results and eliminate spammers or low performing workers. Large or complex jobs better decomposed into simpler and shorter tasks to collect more judgements quickly. For example, video coding is very time consuming and exhausting. Such job can be decomposed into tasks comprise of short clips and crowdsourcing each clip as a small independent task to workers (Walter S Lasecki, Gordon, et al., 2014). Enabling many crowd workers to contribute into coding the entire duration of the video.

Whereas in the *running* phase, crowd jobs are made available online via a crowdsourcing platform. As previously mentioned, the tasks within each job will be executed following one of two strategies, sequentially or simultaneously depending on the output nature of the crowd jobs. Tasks may be executed sequentially one after another, when a task depends on the outcome of another one. In this strategy, the outcome of a task becomes the input to the one that follows. To the contrary, tasks may need to run simultaneously together when they are fully independent with the outcome of one task has no effect on the others. If necessary, the crowd job can be paused and modified to meet requester's new requirements.

Depending on the nature of the crowd job, some tasks require one worker to complete as a whole, while others require two or more. For example, jobs of creative nature like designing a badge using crowdsourcing methods, is a way to find the best one design. Although many workers may contribute into this creative task, only one design will be accepted and only the winner will be rewarded as promised by the requester (Brabham, 2010) as in www.threadless.com[4].

In the finale phase after crowd workers complete all responses, the requester receives all accepted responses and finalise the full job. Depending on the design of the crowd job and its tasks, all responses will then be collated and go through an aggregation process to produce the overall outcome of the full job. Further quality measures check may be necessary after receiving the results to account for outliers and malicious workers who may have passed the initial quality measures during the *running* phase.

2.3.4 Tasks: the unit of work

Problems that requires crowdsourcing vary in size and complexity (J. Cheng et al., 2015) and to aim for quality results requesters may decompose the problem into smaller tasks. Each task can be crowdsourced separately as a complete unit. The result of all tasks is then aggregated to deliver the final result of the job (Chittilappilly et al., 2016). However, some tasks are small or simple and cannot be decomposed further.

To identify tasks complexity, (Nakatsu et al., 2014) conducted an extensive review of

[4] www.threadless.com is an online community of artists and an e-commerce website. Their products are designed and chosen by their online crowdsourcing community.

crowdsourcing task complexity, and concluded the first task-based taxonomy. They have conceptualised the task complexity into three dimensions based on task characteristics. First is the Task Structure, which represent how well a task and the required contribution are defined. A task is well-structured if it defines the required solution, for instance, transcribe a one-minute audio clip. While, a task is unstructured if the desired solution cannot be defined, often the case in creative tasks, for instance, write a story as in Ensemble study (Kim et al., 2014). Second dimension is the Tasks Interdependence. It consists of Independent tasks that can be completed separately by one worker with little or no interaction with others, and Interdependent tasks that require collaboration of multiple workers or aggregation of previously completed tasks to be resolved. Finally, the Task Commitment, which is presented by the degree of dedication required to resolve a task. Low-commitment tasks are straightforward and can be resolved easily and quickly, such as object labelling or allocation. To the contrary, high-commitment tasks are time consuming and expensive to perform. Although requesters are often aware of these dimensions, noted (Nakatsu et al., 2014), they lack the understanding and knowledge of what to do and how to do it to account for the task complexity. The two case studies presented in this PhD research involve low commitment tasks (e.g., allocate the centre of eye pupil) and high commitment tasks (e.g., transcribe a recorded speech). All designed tasks are interdependent as it is essential to harness multiple responses and obtain average acceptable answer, besides, I employ a quality method where a task quality is measured by precedent tasks. Since the solution of tasks in the two case studies is well-defined, all tasks are well-structured with clear definition and goals.

Moreover, (Gadiraju et al., 2014) explored the most popular crowdsourcing task types based on a study of a thousand workers online, and proposed a two-level classification model. One based on what is required to achieve, and the other based on the workflow or the method to achieve the goals of the tasks. They identified six popular task types based on their goals, *information finding*; *verification and validation; interpretation and analysis; content creation; surveys; and content access*. For each type Gadiraju et al. also identified various methods to resolve the task. For instance, asking the crowd to code a video clip is a *Media Transcription* task, which can be classified correctly as either a *content creation* (since a worker response form a new material) or *interpretation and analysis* (as a worker utilise his/her analytical skills to code a scene). Refer to (Gadiraju et al., 2014) for the complete taxonomy.

Whereas (Luz et al., 2015) proposed an alternative taxonomy of four levels based on the nature of the tasks. First, *Partition* tasks, which collectively resemble one complex task. The

Aggregation tasks, of which responses are aggregated and used in following tasks or jobs. The *Qualification* tasks used to qualify or disqualify crowd workers based on their responses to these tasks. Qualified workers are then recruited to complete *Grading* tasks where they will assess the outcome of the qualification tasks. In this PhD research, I instruct each worker to assess their own *Grading* tasks in order to improve their input and guarantee a reward.

Similarly, (Yuen et al., 2011) provided another classification based on the nature of the tasks. *Named Entity Annotation* tasks where crowd workers are asked to recognise and label or annotate an object, like a bird or car; *Geometric Reasoning,* where workers are asked to identify and analyse shapes or other visual models; *Opinions and Common-sense,* where workers are asked to give their opinion or rationale regarding a particular area; *Relevance Evaluation,* where workers perform partial evaluation task; *Spam Identifications,* where crowd workers help detecting and eliminating spams out of specific content; and finally *Natural Language Annotation,* simple tasks for human intelligence, but very complex for automated approaches, for example (Callison-Burch & Dredze, 2010) harnessed the crowd to create a pool of annotations for speech and language applications.

With that being said, good task design and guidelines for one task type is not necessarily applicable to another. A possible explanation is that the various tasks variables (e.g., question and media types, duration) as well as task types and natures determine the design process and elements, and impact task performance. As such, researchers often investigate crowd task design for one or two types. For instance, (Marcus et al., 2012) studied selectivity estimation (estimating number of items in a dataset) task design. In their study, crowd workers were instructed to complete review images and respond in two methods, labelling and counting. In the labelling method, workers were required to review an image and select one of the given labels, such as car, bus, lorry. While, in the counting method workers were asked to review a set of images and estimate the number of vehicles with specific properties (e.g., type, colour) in the presented set of images. Both methods were ran using different task designs (e.g., answer type, duration, number of displayed images) and the overall outcome suggest that the counting method is more effective than the labelling method and result in shorter completion time. To the contrary, their results of their study on text coding shows that labelling samples outperform counting in text coding tasks (e.g., Labelling: Does this tweet represent “a query” or “news”. Counting: How many query tweets are there?). They argue that this is so as people are better in processing images than reading text. These results also indicate that the type of questions, open-ended or closed-ended in a task influence workers performance too. While it has been reported that closed-ended questions (with predefined answers) may lead to a high work efficiency and improved accuracy (Jain et al., 2017), (Gadiraju, Demartini, et al., 2015) and (Eickhoff & de Vries, 2013) identified the vulnerability of such tasks. The latter two studies reported how spammers, malicious workers and bots (Difallah et al., 2012) take advantage of multiple choice questions to quickly answer the tasks and gain rewards.

On the other hand, using open-ended questions increases motivation (Moussawi & Koufaris, 2013), leads to more innovative solutions and lessens cheating too (Eickhoff & de Vries, 2013). Furthermore, (Alonso, 2013) conducted a study to identify key characteristics that may influence the crowd results in any type of tasks. First is the quality of the question, which should be straightforward and avoid ambiguity. Alonso suggests for labelling tasks to use numeric scale instead of labels to avoid misunderstanding as the result of cultural differences. And for multiple choice questions, Alonso stresses on giving no more than six to seven possible answers to not increase workers' cognitive load and task completion time.

Although not new to the HCI domain, (Morris et al., 2012) was the first to investigate the Priming effect on crowd workers performance. By exposing crowd workers to a given stimuli (e.g., picture of a laughing boy improves responses quality), that activates a mental pathway and enhances workers' ability to process subsequent tasks related to the priming stimuli. Although priming effect is temporarily and last for a short period, so does crowd tasks, they are normally short and circumscribed. For better performance, their findings suggest crowd training is still necessary, and that workers should be exposed unconsciously to the priming stimuli. Likewise, (Harrison et al., 2013) explored emotional priming effects (temporary emotional changes) on effective crowdsourcing of cognitive tasks. Each recruited worker was instructed to read a given positive or negative story before they are asked to complete the visual task. In the visual tasks' participants were instructed to accurately compare in size two displayed items. Their study found crowd workers performance improve not only by stimulating positive emotions, but negative emotions too (like increase caution). Furthermore, Harrison et al. note other significant priming stimuli that are out of requesters control and may influence workers performance, such as news read elsewhere or interactions with people around them. Although these studies found that priming enhances crowd performance, yet research on how to integrate priming into crowdsourcing is under explored.

2.3.5 Other task's design variables

Researchers have explored the influence of other task's design variables on the efficiency and effectiveness of the crowd, such as task's duration, graphical user interface, training methods, and the sequential order of tasks.

- **Task's duration variable** (Hoßfeld et al., 2014) noted crowd workers are less dedicated to the task than in-lap participants and attributed this to the anonymity nature of crowdsourcing. As such, task duration should be kept to minimum and correlate positively with the rewards to attract committed workers. Thus, it is essential in crowdsourcing to decompose complex and large tasks into smaller tasks for quick and low-cost responses and high performance (J. Cheng et al., 2015; Kittur et al., 2011). Tasks should be decomposed enough until each task demonstrates low level of complexity. Tasks are better constantly decomposed until it cannot be divided further and that workers are able to apply different methods to execute such tasks (Doroudi et al., 2016). Whereas crowdsourcing complex tasks often requires high expertise in the task's field, like in writing an academic article, as it requires substantial amount of time and increases cognitive load. Yet, various researchers studied different methods to decompose and quickly crowdsource such complex tasks, which result in faster completion time and higher quality (Nebeling et al., 2016; Kittur et al., 2011). The length of tasks and its influence on performance was investigated in various literatures. (J. Cheng et al., 2015) reported that although crowdsourcing very granulated tasks results in higher overall job completion time, it yields higher quality contribution. Furthermore, (Allahbakhsh et al., 2013) noted crowdsourcing complex large tasks also leads to higher overall job completion time, but this is likely to be a result of fewer people being interested or qualified to execute it.

Having said that, further work addressed the potential to efficiently crowdsource long tasks without further decomposition (Nebeling et al., 2016; Dai et al., 2015). (Dai et al., 2015) proposed a method to divert worker's attention while performing long tasks by introducing short duration entertainment, they named it "micro-diversion". The aim of this intentional interruption is for workers to maintain their attention increase the level of engagement. In their study, the recruited workers were instructed to complete three long tasks, *image classification*, *articles quality rating*, *entity merging* (e.g., one entity might be a TV presenter and another a songwriter, both with the same name, are they the same one entity?). Each kind of tasks contains *no diversion* or one of two diversion types, *narrative webcomic* or a *dice game*. The results indicate that micro-diversion can increases workers' level of engagement and encourage them to remain focused. Not all diversions, when applied to the same task correspond to the same results though. The findings demonstrate that the narrative webcomic diversion increases crowd engagement for entity merging and articles rating tasks. While the dice game diversion proved to be also effective in articles rating tasks, it could not retain workers in the entity merging tasks. This is possibly due to the nature of the selected game being too different than the given task. On the other hand, micro-diversion in the beginning of the image labelling tasks corresponded negatively with the level of engagement. The proposed method seems to have discouraged workers from completing the task, possibly because workers favour such tasks for their ease of completion and quick reward. So, an interruption is not welcomed in labelling tasks.

However, it is recommended before crowdsourcing the full dataset, to crowdsource a subset using various strategies in order to find the ultimate one to apply to the full dataset (Brambilla et al., 2015).

- Graphical user interface (GUI) This is the key entry point for crowd workers to engage and complete a task, since it is the visual element that workers interact with to resolve a task. It should be self-explanatory and clearly define the tasks requirements and objectives, and often the way to complete it. Researchers found a positive correlation between task design and crowd outcome. A user friendly task's GUI (Allahbakhsh et al., 2013; Finnerty & Kucherbaev, 2013) and adequately detailed instructions with clear examples (Jain et al., 2017) enhance workers' efficiency and increase their accuracy. (Finnerty & Kucherbaev, 2013) reported on how crowdsourcing the same content with two different GUI designed tasks, one simple (white background and simple layout), and another complex (patterned background and unstructured layout) returns different results. The simple GUI designed tasks yields better overall outcome than its complex GUI designed counterpart. Furthermore, simple elements like highlighting the controls that capture crowd responses in different colour results in higher performance and lower task completion time (Sampath et al., 2014). It is also important to note that the quality of crowdsourced content is essential for the job success. (Kim et al., 2015) noted how improved images quality and the lightings in them result in better colour perception and accurate ratings, when asked workers to review products colour of online images with the products' real colour. This all confirm the task design guidelines that (Alonso, 2013) proposed to write simple, clear instructions with relevant examples, and highlight what is required from workers and what they will get. Despite the impact of the length of instructions on the quality of the crowd outcome, workers prefer tasks with concise instructions and guidelines (Wu & Quinn, 2017). Thus, requesters must find the balance between informative and sufficient easy to read and comprehend instructions and guidelines to complete tasks successfully.

Others studied why tasks may lead to low performance and unfair rejection. (Mcinnis et al., 2016) found that poorly GUI designed tasks, ambiguous instructions, technical issues, and requesters lack of expertise in task design result in workers' exclusion, lower accuracy and can lessen participation. As technical issues are sometimes out of requesters control when occur at runtime, Mcinnis et al. recommend providing workers with a method to report on broken tasks. Besides, (Gadiraju, Yang, et al., 2017) reviewed the definition of unclear tasks in a survey of 100 crowd workers. The survey first reveals that key factors to unclear tasks are attributed to the presentation of the instructions, lack of relevant examples and poor writing style. Second, the results helped the researchers to deliver an automated model that predicts and measures task clarity, which can guide requesters in designing clearer tasks.

- Training methods vary between tasks and are usually selected based on requesters' choices. However, researchers investigated the impact of various training methods on the quality of crowdsourcing. (Doroudi et al., 2016) tested four training methods and compared them together, along with a no-training method too. While the no-training (*Control*) method provides no training, only the usual instructions and guidelines, the other four provide training tasks without explicitly informing recruited workers. In the first method, *Solution*, workers are presented with extra tasks (of the same nature) to resolve but will not impact their overall score. Second, *Gold Standard* method, for which workers are presented with extra tasks to resolve. Following the submission of workers responses, they are presented with the correct answers along with experts' examples. Third is the *Example* method, where workers are given one or more examples to review but cannot execute any task until the initial training duration (e.g., 20 seconds) elapses. This is so workers are given enough time to review and comprehend the given examples. The last method is *Validation*, for which one or more previously answered tasks are randomly selected and presented to the worker to validate.

The findings proved both *control* and *example* methods well improved retention rate in compare to the three other methods. A likely explanation is that in *control* method there is no time spent on additional training, and *example* method only requires workers to wait for a short period to elapse. Workers as previously mentioned favour quick and short tasks, so they are more likely to give up on the task the longer they stay in the task. The results also suggest increase in workers commitment level as almost 50% of workers who completed at least one training task completed all their tasks. And for accuracy, all training methods outperformed the *control* no-training method, with *Example* method outperforms the rest. This is also evident in the findings from (Jain et al., 2017; Mitra et al., 2015; Wu & Quinn, 2017) that highlight the importance of providing relevant examples and its positive impact on accuracy, performance as well as inter-rater agreement. While *Validation* method proved most effective for subjective tasks (Zhu et al., 2014).

In this research, I focus on utilising some of these methods to train recruited workers. I also use *validation* to improve workers responses, but instead of validating others' responses workers will validate their own. And I will also utilise *gold standards* as one of the quality control methods required.

- The sequential order of tasks and their execution represents the order in which tasks, questions and data are given to crowd workers to resolve. The sequential order of tasks may be used to develop workers relevant skills and retain their productivity by gradually increasing tasks difficulty as they perform more tasks. Various studies explored how the sequence of tasks influence crowd workers performance. For example, (Cai et al., 2016) found a worker's tasks completion time is reduced when completing tasks that were preceded by other tasks of the same content and equivalent or lower level of complexity. This is likely because workers become more familiar with these tasks and best ways to complete them, and gradually gain the skills needed to do more complex ones of the same content. To the contrary, the researchers reported that workers become slower in completing tasks of complexity level lower than the preceded ones. A possible explanation is that starting with higher complexity tasks adds extra cognitive load on workers, which exhaust their mental resources. Cai et al. also found that workers cognitive load is minimised when completing lower complexity tasks before performing more complex ones of the same content. However, their findings suggest no correlation between the sequential order of tasks and the quality of results.

Likewise, (Shao et al., 2019) study confirms that performing a chain of similar tasks improve workers efficiencies. Furthermore, (Lasecki et al., 2015) reported that switching context (tasks preceded with tasks of different nature and content) negatively impact workers speed to complete a task. In addition to influencing workers' efficiency, the sequential order of tasks can impact the inter-rater agreement between workers. In two studies (Damessie et al., 2016, 2018) explored the effect of document presentation order on inter-rater agreement level between crowd workers. In their first study, (Damessie et al., 2016) recruited crowd workers to assess the relevancy of documents of different topics using a four-point scale, with documents presented in two orders. In the first ordering technique, documents are presented based on the relevancy level, from the highest relevant to the least (*decreasing-relevance order*). Whereas in the second ordering technique, documents are presented based on the TREC assigned identifier (special identifier based on various relevancy factors). The findings suggest both ordering techniques increase agreement level between workers responses, with TREC assigned identifier ordering technique yield better results than the decreasing-relevance order. The authors suggest the decreasing-relevance order method may have influenced crowd workers judgement to underestimate the relevancy of documents preceded by higher relevance documents. The second study (Damessie et al., 2018) added one more ordering technique, named *Interleaved Likelihood of Relevance* in which documents are ordered carefully interleaving most relevant with least relevant documents before presenting them to the crowd workers. The authors reported higher inter-rater agreement between workers than that found in their 2016 study, along with a substantial agreement between workers and expert (TREC) judgements.

On the other hand, (André et al., 2014) explored the effect of workers executing the same task simultaneously or sequentially on output quality. They recruited crowd workers and divided them onto two coordination strategies to complete complex tasks. The first group work together simultaneously, while the other work sequentially one after another. The findings were controversial in crowdsourcing domain, as they suggest sequential strategies to be more effective than working together simultaneously. This, the authors claim, is partially due to social processes (e.g., territoriality). However, they also suggest given specific roles to workers working simultaneously may increase their efficiency too.

The following two chapters are built upon the literature to achieve the aim of this research and answer its questions. Chapter 3 introduces crowdsourcing solutions for analysing speech audio data collected in the wild. The aim is to demonstrate the power of the crowd and their capability to produce quality responses equivalent to that of an expert when working with real world data. Chapter 3 also demonstrates the value of crowd-based feedback to people with speech difficulties. Likewise, chapter 4 introduces crowdsourcing methods that deliver highly accurate crowd responses when localising the centre of a target, and demonstrates the value of such work for crowdsourcing and eye tracking communities. In addition, chapter 4 evaluates crowdsourcing self-versification methods and their influence on quality and costs.

3 Designing for Quality Crowd Speech Assessment to Support the Self-Monitoring and Management of Speech and Voice Issues

In this chapter I present Speeching, a mobile application (app) that supports the self-monitoring and self-management of speech and voice issues for people with Parkinson’s (PwP) leveraging crowdsourcing to assess and feedback on users’ speech data. The app enables PwP participants to audio record voice tasks and post for crowd feedback. Crowd workers, who are not familiar with the PwP’s voice patterns, then assess and rate the voice tasks. The PwP user then receives feedback via the Speeching app based on the crowd workers collated judgements, illustrating how their speech was perceived by novice listeners unfamiliar with their voice pattern. This allows the PwP to better understand their progress as they practice speech tasks at their convenience. The study was conducted in two phases. The first to assess feasibility of novice listeners (crowd workers) to judge speech and voice that are comparable to those of experts. Then to conduct a trial deployment to evaluate the provision of feedback through the Speeching app and its value for PwP participants. The study highlights how Speeching, and similar applications, can provide users with new opportunities for self-monitoring health and wellbeing. Digital applications like Speeching can improve the means by which participants without regular access to clinical assessment service can practice and receive feedback to better self-manage therapeutic interventions in speech and voice training tasks.

3.1 Introduction

Crowdsourcing has emerged as a research tool to collect and analyse large sets of raw data in various domain including the medical domain (Swan et al., 2010; Chunara et al., 2012). As a research tool, the benefits of making connections with participants and gather information has been well-acknowledged. Crowdsourcing may be able to contribute to everyday healthcare, but this arena has not been explored at length. One area this chapter is particularly interested to explore is the value of leveraging the crowd to provide support in personal health. Personal health is important to individual medical participants because concepts like self-care, self-management, personal motivation, and constant monitoring of health conditions and changes can create marked improvements in individual's health (Barlow et al., 2002; Nunes & Fitzpatrick, 2015).

One role that crowdsourcing could have helped in personal healthcare is through the application to speech and language therapy (SLT). SLT is the training, practice, and use of specific skills related to conditions that impact the ability to speak and use one's voice. Acute conditions such as a stroke or traumatic brain injury occur suddenly and alter the way in which a person speaks. Degenerative conditions occur over time, as the vocal capacity of the patient lessens, and are caused by a number of conditions such as Parkinson's disease, motor neuron disease, and dementia. The SLT practitioner delivers series of exercises that require the patient to repetitively practice and build fluency, strength, and voice capability. These exercises are commonly delivered and practiced in clinical setting while the speech and language therapist monitors and maintains information about the patient's progress, gains, changes, and challenges. Repetitive practice cannot always occur in a clinical setting, and often there is SLT practice work that the patient accomplishes at home. However, practicing at home comes with motivational barriers in the self-directed practice of speech (Nijkrake et al., 2007), and treatment may not persist through the long-term after therapy has occurred (Green et al., 2002; Wight & Miller, 2015). Speech and language therapists (SLTs) acknowledge concern that clinical and therapeutic demands extend beyond capacity in both developed and developing countries (Miller, Deane, et al., 2011; McKenzie, 1992). As a result, people with impaired speech may benefit from new approaches in self-directed therapeutic practices.

In response I developed Speeching, a crowd-based digital analysis solution that provides users with direct and meaningful feedback. The aim is to facilitate self-management-and-care, and motivate individuals with impaired speech to complete SLT exercises at their convenience outside of the clinic. The study focuses on persons with Parkinson's disease (PwP), who are likely to experience speech difficulties as a result of neuromuscular degeneration (Ho et al., 1999; Miller et al., 2007). Speeching system comprised of a smartphone application that allows participants to audio record their self-practice of a series of speech tasks and to upload these to a remote server. The recordings are then judged by crowd workers based on ease of listening, and speech pace, pitch variability and volume. The responses from the crowd workers are then collated and delivered to the participants via the app, Speeching, and visualised to provide therapeutic directions to support off-clinic practice of SLT exercises. This chapter first demonstrates the feasibility of using crowd workers to judge recorded speech compared to expert judgements. Based on the outcome of the feasibility study, the crowdsourcing solution, Speeching, was developed and deployed in a real-world pilot study to establish its acceptability among Parkinson's community.

This study highlights the capability of crowdsourcing to offer and support new form of self-management-and-care practices. It provides answers to the research questions through a practical deployment of a novel crowdsourcing solution to analyse and feedback on speech audio data. The solution was explored in two iterations, feasibility and deployment, to identify best ways to design the crowd task and achieve the desired quality standard. The final iteration implements clear tasks design that matches with SLT's speech and language assessments' tasks. The aim is to establish the capability of crowd workers to generate quality assessment equal to experts, and provide a valuable and meaningful feedback to those being assessed. This research provides four key contributions to the field of HCI in addition to a fully functioning crowdsourcing solution and a large in-the-wild crowd-annotated speech audio dataset. The first contribution is the demonstrative feasibility of crowdsourcing to produce quality judgements of impaired speech equivalent to SLT experts' judgements. The second is the exemplification of real-world crowdsourcing application as a method that has the capacity to present meaningful data from crowd workers directly back to patients, and the benefits and challenges that occur within this chosen method. Third, this study provides an enhanced understanding of how participants with Parkinson's are influenced by the crowdsourced ratings, and the value they placed in Speeching that promotes the self-care practice of therapeutic tasks. Lastly, this study offers insights for future researchers who seek to further explore the application of crowdsourcing in personal health.

3.2 BACKGROUND

3.2.1 Crowdsourcing Health

Crowdsourcing research in healthcare has largely focused on the collection of raw data. Examples include studies to understand the capacity of online health communities to act as representatives of wider populations (Bove et al., 2013); exploit the personal data collected by health communities about themselves, to explore preventative medicine (Swan, 2012; Swan et al., 2010); explore how to harness online communities to provide new sources of patient data for research (Patientslikeme, 2015); and to investigate how can online communities take a supportive role among specific patient group (Wicks et al., 2012). Crowdsourcing has also been used to facilitate analysis of patient data, and often involved crowd of experts. For example, Crowdmed[5] is an online platform that offers a pool of medical experts to solve medical conditions posted by online patients, whereas (Xiang et al., 2014) leveraged crowdsourcing techniques for general practitioners to diagnose patients with illnesses that require multiple diagnoses consensus by different doctors. On the other hand, non-expert crowd workers have also been harnessed to analysis large scale clinical data. Some non-expert crowdsourced examples include the use of crowd workers to identify malarial parasites in images of blood samples (Chunara et al., 2012), identify genome protein structures (Cooper et al., 2010), and classify colonic polyps within radiography scans (Nguyen et al., 2012). Crowdsourcing therefore has been effectively used in medical research, medical diagnosis, and medical imaging.

Crowdsourcing has been used beyond the healthcare context in interactive, user-supported systems and human powered assistive technologies that are influential in modern works (Bigham et al., 2011). Examples include VizWiz a smartphone application that provides near real-time feedback on visual information to blind people (Bigham, Jayant, et al., 2010; Burton et al., 2012), and the ASL-STEM Forum an online portal for contributing sign language describing scientific terminology for deaf or hard of hearing people (Cavender et al., 2010).

[5] Crowdmed https://www.crowdmed.com/

Both VizWiz and ASL-STEM are samples of human powered assistive technology that leverages crowdsourcing to support participation and motivation of persons with difficulties. However, a gap can be identified in the current literature, where there has been relatively little work that investigated the use of non-expert crowd workers to support self-management of off-clinic healthcare practices, exercises, and therapeutic tasks. This study explores the gap in research by leveraging crowd workers to provide feedback ratings to support, promote, and motivate PwP's self-monitoring and personal healthcare management.

3.2.2 Crowdsourcing for Speech Data

Researchers have studied the application of crowdsourcing to overcome speech analysis problems in the collection (McGraw et al., 2009) and transcription of speech data (Marge et al., 2010; Audhkhasi et al., 2011; Parent & Eskenazi, 2011; Wolters et al., 2010). Others have examined crowdsourcing potential to advance speech recognition systems, like PodCastle that offers full text speech search and encourages users to correct recognition errors to train the algorithm (Goto & Ogata, 2011). Moreover, researchers have utilised crowdsourcing to measure the quality of speech samples. Parent et al. highlighted in their review (Parent & Eskenazi, 2011) of 29 papers the value of employing reductive measures of intelligibility, where crowd workers were recruited on AMT to transcribe and classify short utterances on speech samples collected by users of a transport information system. Whereas Marge et al. (Marge et al., 2010) evaluated the reliability of crowd workers in transcribing spontaneous speech samples. Their findings indicate that crowd workers transcription accuracy approached that of an expert, and that shorter segments of speech were more likely to have faster turnaround times and higher rates of transcription accuracy (Marge et al., 2010). Other studies have evaluated crowdsourcing techniques to rate the perceptual aspects of speech. Evanini et al. (Evanini & Zechner, 2011) studied the use of crowdsourcing and examined the viability of utilising crowdsourcing for annotating prosodic stress and boundary tones on a corpus of spontaneous speech. The results show high level of agreement between crowd workers when compared to experts (Evanini & Zechner, 2011). However, their study was conducted with 11 annotators, from an outsourcing company, who were carefully selected and had obtained university-level education and are highly proficient of English language. The annotators also received training prior to performing any task. While their research can be considered crowdsourcing, it breaks one of the online crowdsourcing fundamentals, and that is the random selection of workers. To the contrary, the crowdsourcing solution presented in this chapter recruits random crowd workers regardless of their education, and does not offer workers any training. The solution primarily relies on the quality of the task's design to achieve crowd responses equal to that of an expert.

Such research studies provide a range of samples of crowdsourcing for speech data, and have highlighted the methodological considerations in this area of research. Still, there are number of difficulties to consider when assessing impaired speech for clinical population. As such, it is essential to understand current SLT clinical literature and the methods and practices used to measure speech and voice data of PwP in order to address associated difficulties.

- Parkinson's Speech

Approximately 90% of PwP will experience speech and voice degeneration through the progression of the disease (Ho et al., 1999). Changes that occur in speech and voice include volume reduction, prosody (stress and intonation patterns), level of loudness (monoloudness), and variation in pitch (monopitch). Besides, PwP may experience a hoarse, rough, breathy, or trembling speaking voice if their perceptual vocal quality becomes impaired (J. Holmes et al., 2000; Tjaden, 2008). For the PwP, such characteristics can cause feelings of lowered confidence, embarrassment, and increased difficulty speaking with strangers. As a result PwP may avoid social situations, which indicates the importance of speech therapy for PwP to retain social interactions, confidence, and self-esteem (Miller et al., 2007, 2008, 2006).

Qualified SLTs diagnose, measure, assess, and plan therapeutic interventions for PwPs with voice difficulties. A clinical interview with an SLT involves speech sample collections from the client, which undergo formal and/or informal assessment. However, one acknowledged issue with the clinical assessment is that SLTs are very experienced and familiar with impaired speech patterns, and this familiarity can cause a predisposition to score higher during the assessment (Miller, 2013). Therefore, best practice guidelines recommend that SLTs use naïve listeners to create a representative rating for the SLT to use comparatively. In the clinical setting it is, however, difficult to use naïve listeners due to time, resources, and clinical constraints (Ziegler & Zierdt, 2008). Furthermore, PwPs limited access to clinical services, and time restraints, means many PwPs will not benefit from such services the first instance (Miller, Noble, et al., 2011).

- Measuring intelligibility

The challenges of speech intelligibility testing are in the access to clinical facilities, the predisposition of familiarity, and the reasonability of double assessments as recommended by best practice guidelines (Ziegler & Zierdt, 2008; Miller, Noble, et al., 2011; Miller, 2013). Researchers responded to these challenges by studying how online digital platforms could remotely conduct speech intelligibility testing, which would allow for reducing familiarity and increasing both accessibility and assessment capacity. The Munich Intelligibility Profile (MVP) is an online system that provides SLTs with remote access to intelligibility assessments for dysarthric[6] speech (Ziegler & Zierdt, 2008). In MVP people with impaired speech are examined online by trained listeners, which later evaluated by SLTs panel. Despite the success the MVP has achieved in decreasing individual deviation from the mean with increased number of listeners, and offering accessible online platform to obtain standard intelligibility measures from large dataset, it has created a level of external control. First, the analysed speech samples were collected in clinical settings and listeners responses were reviewed by a panel or expert SLTs. Besides, moderators coordinated the assignment of speech samples to listeners, then collated and appraised listeners' responses (Ziegler & Zierdt, 2008).

Crowdsourced workforces, or crowd workers, create an abundant, affordable workforce of listeners accessible through online platforms (e.g., Amazon Mechanical Turk) in the context of speech intelligibility testing. To date, there is not significant and established work that has previously examined the potential for crowd workers to provide speech assessment into a program of speech therapy. Although, more research have examined and utilised crowdsourcing platforms to provide diagnostic speech ratings. For example, untrained listeners crowdsourced through AMT were instructed to classify speech samples of children with articulation difficulties as correct or incorrect (McAllister Byun et al., 2015). The classifications from untrained listeners were compared to the judgements of experienced listeners, and found that there was an extremely high (0.98) agreement between non-experienced and experienced listeners (McAllister Byun et al., 2015). This highlights the potential for crowdsourcing to have a role in SLT practice, and for researchers to examine how crowdsourcing can be used in measures of intelligibility beyond the binary approach.

[6] Dysarthria is a motor speech disorder characterized by unclear articulation of words. Words will be linguistically normal unless an additional underlying impairment is present. PwP experience hypokinetic dysarthria characterized by reduced volume, abnormal speaking rates and harsh or breathy vocal quality [16].

The Speeching case study addresses these gaps by exploring novel methods to collate real world speech data, and examining the potential for leveraging the crowd to provide feedback on impaired speech. The study occurs in two phases. Phase one demonstrates the feasibility of anonymous crowdsourced workers to rate impaired speech. The second phase involved in-the-wild deployment of the Speeching app to collect speech and voice samples from PwP and provide them with meaningful feedback, unsupervised in their home environment at their convenience.

3.3 Phase 1: Testing Speeching Feasibility

The first phase aimed to explore the development of crowdsourcing tasks which might elicit ratings of Parkinson’s speech equivalent to expert ratings. Degeneration of speech and voice due to Parkinson’s has specific elements of impairment often selected for investigation like rate, pitch, and volume, which are the most common variables in degenerated speech due to Parkinson’s (J. Holmes et al., 2000).

3.3.1 Selecting the sample dataset

In this feasibility phase, a sample of twelve speakers were selected from a previously compiled data set of 125 persons with Parkinson’s collected in a controlled lab environment (Miller et al., 2007). The data set was solicited from an SLT expert in Parkinson’s speech, who reviewed all 125 samples in order to select the representative sample. The sample was made up of three groups based in speakers’ intelligibility problems: Mild, Moderate and Severe, with two male and two female speakers in each group. The audio recordings of the selected speakers composed of equal samples of ten single words (unconnected speech), and nine sentences (connected speech) from the Grandfather Passage (Darley et al., 1975).

3.3.2 Designing the micro-tasks

The Speeching tasks were designed together with an SLT expert in Parkinson’s speech to simulate a standard SLT assessment. In such assessment SLTs listen to a range of single words, sentences, and longer samples of speech, from the PwP while reading a short text, describing a situation, and engaging in open ended discussions about a topic. It is a common practice that SLT will make an initial recording of the PwP’s pre-therapy voice before conducting a clinical assessment and diagnosis. The SLT then assess the volume, rate, and vocal qualities of the

PwP's speaking patterns and voice to identify vital difficulties being experienced. The SLT uses a range of standardized assessments to objectively measure the PwP's speaking capabilities, alongside nonstandard methods relative to the expertise of the SLT.

The Speeching crowd tasks were designed to judge two categories of speech samples. The first was unconnected speech or single words, which is a range of single random words that are not related to one another. Unconnected speech provides a measurement of intelligibility in isolation by removing context and flow that may add to the listener's ability to hear and relate the words together in an intelligible message. Unconnected speech category was selected since it is widely used in SLT assessments, and apportions for a finer analysis of the specific sound contrasts a speaker is having difficulty with, providing direction for therapeutic input. Although this was not fully explored in my work, I wanted to include this task for crowd analysis to scope wider, future potential for the system. So, for the unconnected speech crowd tasks I instructed crowd workers to listen to one word recording and select the target word from ten similar words (e.g., sheep, keep, heap). The ten single words in the word recognition tasks of each assessment were part of an assessment conducted by (Miller et al., 2007) designed to target specific sound contrasts.

The second category of speech samples was the connected speech. Connected speech is comprised of sentences that allow for an analysis of acuity and flow. To rate samples of this category, two types of measurements were applied: Ease of Listening (EOL) and perceptual measures. EOL measurements were subjective based on the crowd workers' effort to understand the PwP speaker. The measurement used a five-point Likert scale, which has been previously used with novice listeners unfamiliar with dysarthric speech and was found to have a strong correlation to intelligibility scores (Landa et al., 2014; Miller et al., 2007). Whereas in perceptual measurements of speech, the listener's perception of rate, pitch, variance, and volume required more complex appraisal. To provide a more multifaceted approach than Likert scale, a continuous scaling system was applied to improve sensitive accuracy of responses (Côté, 2011; Miller, 2013). It has been recommended by (Miller, 2013) to apply Direct Magnitude Estimation (DME) for perceptual intelligibility measures. In the DME perceptual scaling an anchor or baseline exemplar of impaired speech is played to the listener so to allow for an estimation of the magnitude of variance in the connected speech tasks (Gary & S., 2002; Miller, 2013).

Since the recruited crowd workers were naïve listeners in disordered speech, they were expected to exhibit variability in rating volume, pitch variance and rate. To mitigate this possibility the study applied a continuous scale of 0-100 that allow for large range of variability between listeners without impacting the sensitivity of ratings that may have been observed in a discrete scale. The baseline exemplar sample was chosen by the SLT expert who selected one male and one female speaker from the larger dataset of 125 speakers. Each selected speaker was representative of a moderate speech impairment of pitch, rate, and volume variance. These baseline speakers were not amongst the twelve speakers who had been selected for this case study for analysis. Baseline exemplar samples were gender matched to the participants from the selected representative sample. Crowd workers were instructed to rate the speech, out of 100, for volume, rate and pitch variance using the baseline exemplar as a reference point for a score of 50.

3.3.3 Participants

In the feasibility study all crowd workers were UK based and recruited from AMT (33 workers in total) to complete the rating tasks of the entire dataset of 282 speech samples. To control the quality of crowd responses, two SLT experts in Parkinson’s speech rated the entire same dataset, which was then used as the crowd tasks’ gold standard. Crowd tasks were randomly and automatically assigned to recruited crowd workers. Each speech sample was crowdsourced with a minimum of three judgements, whilst ensuring a listener cannot rate the same sample twice. The samples with three different ratings were indexed and the task randomized again until the data set was fully crowd rated. The listeners were only required to complete 70 tasks (~25%) to receive a monetary compensation for their time. The compensation was equivalent to the UK minimum wage (at the time of this study) and based on the average time to complete each task I estimated prior launching the study.

3.3.4 Phase 1 Analysis

This study examined the correlation between crowd workers and experts on recognising unconnected speech tasks, and measuring pitch, rate, and volume, as well as judging the ease of listening tasks. Each data sample was judged by a different set of crowd workers following a systematic task distribution. Considering that our observations were based on independent data samples, Spearman’s Rho was selected. Following Spearman’s Rho test highlights potential differences between distinct groups and measures the strength of association between the crowd workers and the experts. The success rate of recognising single word tasks across

	Measure: Volume			**Measure: Pitch**			**Measure: Rate**		
	Median (IQR; Q1, Q3)	Range of scores	Spearman's r (p)	Median (IQR; Q1, Q3)	Range of scores	Spearman's r (p)	Median (IQR; Q1, Q3)	Range of scores	Spearman's r (p)
Expert	98 (IQR=23; 90, 113)	60-120	—	100 (IQR=10; 90, 100)	50-115	—	75 (IQR=40; 60, 100)	40-205	—
AMT	100 (IQR=35; 85, 120)	50-123	0.16 (p=0.57)	100 (IQR=20; 80, 100)	20-140	0.81 (p=< 0.01)	85 (IQR=50; 50, 100)	20-180	0.71 (p=<0.01)

Table 3-1 Summary of results for phase 1 study on the measures of volume, pitch variability and rate

each speaker was also computed by converting judgements onto a binary correct / incorrect score and aggregate them into a total % of words correct score. Whereas the median score out of 100 was taken from each group of three crowd workers who had analysed the speech sample to measure pitch, rate and volume and compared to the median score of the experts (median was chosen over mean due to the nature of the continuous rating scale, to account for possible outliers in the data).

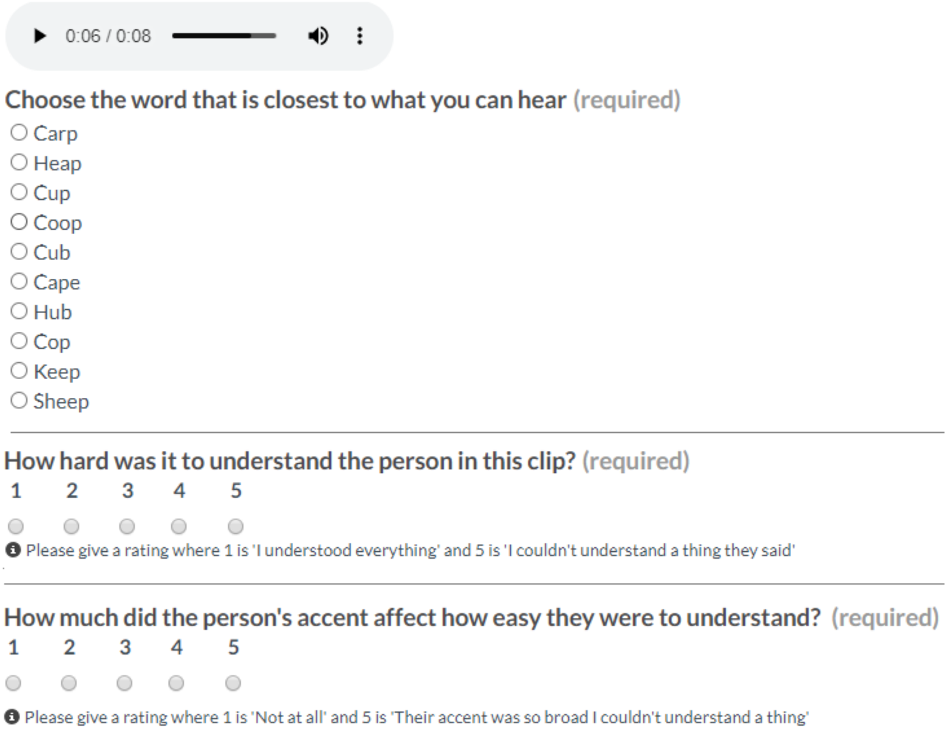
3.3.5 Phase 1 Findings

On the measures of volume, pitch and rate Table 3-1 summarises the results and shows the value of the lower quartile (Q1) and upper quartile (Q3), as well as the interquartile range (IQR) of observed scores. The results show strong correlation between the crowd workers and the experts on judging pitch and rate, implying that crowd workers perceptual measure of speech is comparable to that of an expert. Besides, strong correlation (Evans, 1996) on the unconnected speech word recognition task was observed between the scores of the experts and the crowd workers, which suggests that crowd workers responses matched those of the expert. While a substantial agreement was found for the ease of listening tasks (Landis & Koch, 1977). As such, one can speculate that we can leverage crowdsourcing to recruit anonymous crowd workers who will provide quality ratings comparable to that of an expert in the measurements of speech and voice changes for PwPs speech. The results therefore support the feasibility of our crowdsourcing approach.

However, weak correlation was observed for the measure of volume between the experts and the crowd workers. Although the representative speech data was collected in lab environment, the weak correlation could be the result of inconsistent quality of the recordings between speakers. Inconsistent quality could be associated to external factors like recording environment, equipment, external noise, and other vocal elements. One may speculate that crowd workers rated the quality of a recording instead of the actual speaking volume, indicating a limitation that phase 2 must address if volume measurements is to be included. So, to overcome this limitation, crowd workers should be asked to rate speaking volume of a recording in comparison to another recording of the same user collected in the same way each time. I should also point out that the experts and the crowd workers provided a comparable range of scores for speaking volume (60-120 and 50-123 respectively), which is also smaller than the range of score in the pitch and rate measures. The volume crowd task instructed crowd workers to judge the volume in the sample being scored in comparison to the baseline sample. The volume question stated a score of more than 100 indicates that you think the clip on the right (the sample being scored) exhibits more severe problems in terms of volume (than the baseline), with the reverse being stated for less than 100 (less severe). Although not deliberately, the question failed to define what this study considered as a volume impairment, the question should have stated that a low volume reveals severe problems. Unlike the experts who are experienced in diagnosing volume problems and their impact on the speaker, it is conceivable that the crowd workers were rating on the lower end of the scale to indicate any change in volume. So, while the experts were rating volume impairments, the crowd workers were possibly rating lower for lower volumes. In several volume tasks, crowd workers rated the speaking volume samples between 60 – 85, while the experts' ratings were between 100 – 120 of the same samples. As a result, the volume questions were revised for phase 2 to ensure clarity of what is being asked.

3.4 Phase 2: Implementation of Speeching

The Speeching system consists of several components. A mobile application (app) that enables its users to self-monitor their speech problems. The app prompts its users (in this phase are the participants with Parkinson's) to complete a variety of assessment tasks in order to collect audio samples of their speech. Collected speech samples are then uploaded to the Speeching crowd service to prepare the recordings and set up the crowd 'job' (a collection of independent recordings) before uploading them to CrowdFlower for ratings. CrowdFlower was chosen in preference to AMT in response to the AMT financial restrictions imposed in the UK. Each job consists of two forms of speech samples, unconnected and connected speech, and five crowd workers are recruited to listen and assess the samples. For the unconnected speech samples (n=10), crowd workers are instructed to listen to a single word and choose the word they have heard from a list of 10 options of similar sounding words. While for the connected speech samples (n=3), the workers are instructed to listen to the full recorded sentence and rate its understandability in addition to pitch variability, volume, and rate. On completion of each job, crowd responses are sent back to the Speeching service, where they are collated and their median score is calculated before sending them back to the user via the app, as performance feedback on their speech. The aim of this feedback is to improve users' intelligibility by highlighting the areas of their speech that require practice, allowing them to manage and conduct targeted exercises on their speech via the app.



Choose the word that is closest to what you can hear (required)

○ Carp
○ Heap
○ Cup
○ Coop
○ Cub
○ Cape
○ Hub
○ Cop
○ Keep
○ Sheep

How hard was it to understand the person in this clip? (required)

1	2	3	4	5
○	○	○	○	○

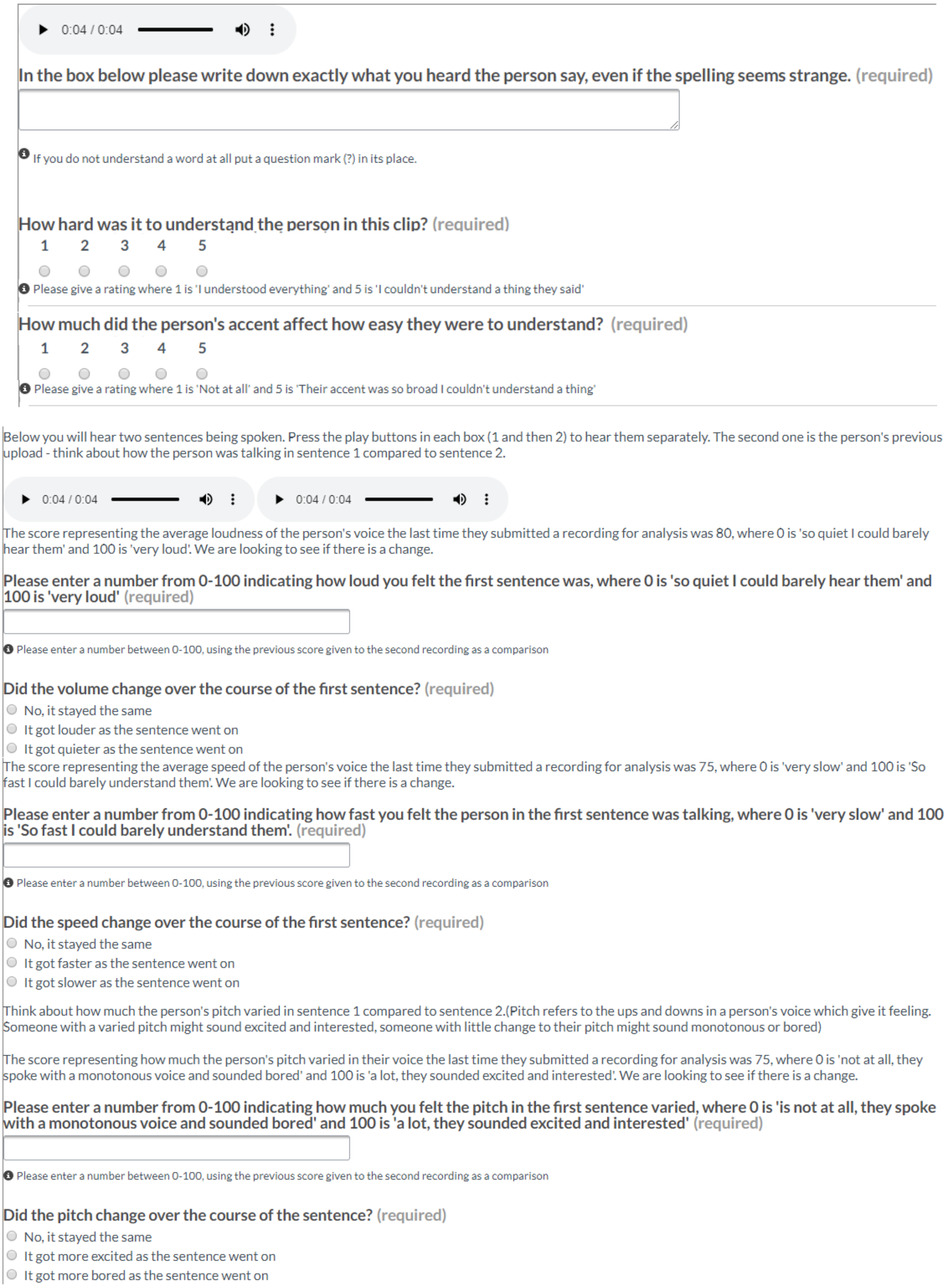
Please give a rating where 1 is 'I understood everything' and 5 is 'I couldn't understand a thing they said'

How much did the person's accent affect how easy they were to understand? (required)

1	2	3	4	5
○	○	○	○	○

Please give a rating where 1 is 'Not at all' and 5 is 'Their accent was so broad I couldn't understand a thing'

Figure 3-1 Single-word (unconnected speech) Crowdsourcing task



In the box below please write down exactly what you heard the person say, even if the spelling seems strange. (required)

If you do not understand a word at all put a question mark (?) in its place.

How hard was it to understand the person in this clip? (required)

1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Please give a rating where 1 is 'I understood everything' and 5 is 'I couldn't understand a thing they said'

How much did the person's accent affect how easy they were to understand? (required)

1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Please give a rating where 1 is 'Not at all' and 5 is 'Their accent was so broad I couldn't understand a thing'

Below you will hear two sentences being spoken. Press the play buttons in each box (1 and then 2) to hear them separately. The second one is the person's previous upload - think about how the person was talking in sentence 1 compared to sentence 2.

The score representing the average loudness of the person's voice the last time they submitted a recording for analysis was 80, where 0 is 'so quiet I could barely hear them' and 100 is 'very loud'. We are looking to see if there is a change.

Please enter a number from 0-100 indicating how loud you felt the first sentence was, where 0 is 'so quiet I could barely hear them' and 100 is 'very loud' (required)

Please enter a number between 0-100, using the previous score given to the second recording as a comparison

Did the volume change over the course of the first sentence? (required)

○ No, it stayed the same
○ It got louder as the sentence went on
○ It got quieter as the sentence went on

The score representing the average speed of the person's voice the last time they submitted a recording for analysis was 75, where 0 is 'very slow' and 100 is 'So fast I could barely understand them'. We are looking to see if there is a change.

Please enter a number from 0-100 indicating how fast you felt the person in the first sentence was talking, where 0 is 'very slow' and 100 is 'So fast I could barely understand them'. (required)

Please enter a number between 0-100, using the previous score given to the second recording as a comparison

Did the speed change over the course of the first sentence? (required)

○ No, it stayed the same
○ It got faster as the sentence went on
○ It got slower as the sentence went on

Think about how much the person's pitch varied in sentence 1 compared to sentence 2.(Pitch refers to the ups and downs in a person's voice which give it feeling. Someone with a varied pitch might sound excited and interested, someone with little change to their pitch might sound monotonous or bored)

The score representing how much the person's pitch varied in their voice the last time they submitted a recording for analysis was 75, where 0 is 'not at all, they spoke with a monotonous voice and sounded bored' and 100 is 'a lot, they sounded excited and interested'. We are looking to see if there is a change.

Please enter a number from 0-100 indicating how much you felt the pitch in the first sentence varied, where 0 is 'is not at all, they spoke with a monotonous voice and sounded bored' and 100 is 'a lot, they sounded excited and interested' (required)

Please enter a number between 0-100, using the previous score given to the second recording as a comparison

Did the pitch change over the course of the sentence? (required)

○ No, it stayed the same
○ It got more excited as the sentence went on
○ It got more bored as the sentence went on

Figure 3-2 Connected speech Crowdsourcing task

3.4.1 The Speeching App

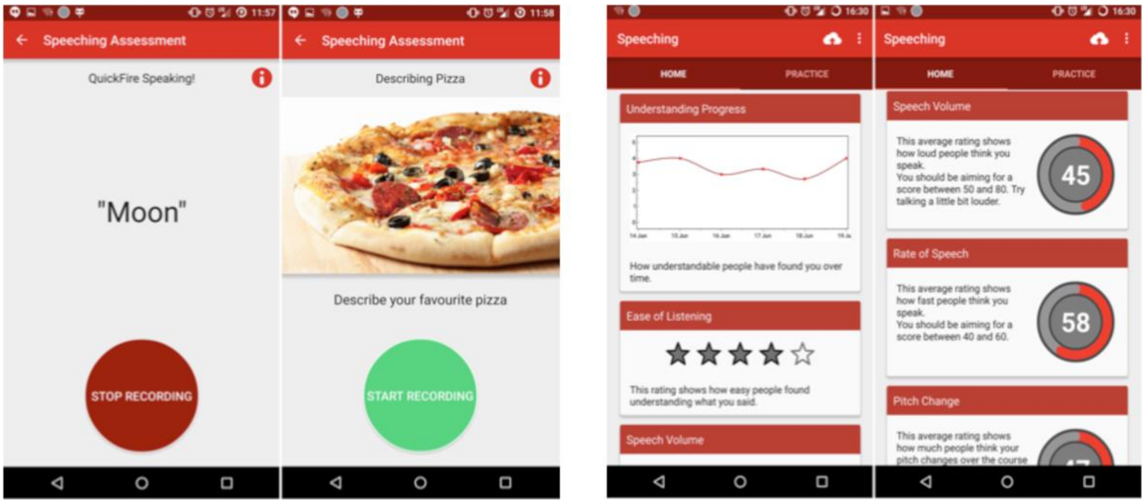
The Speeching app is made up of three areas: the assessment area where users complete several audio-recorded speech stimulation tasks to be assessed by the crowd; the feedback area where users can monitor and review the performance of their crowd assessed speech; and finally, the practice area where users can conduct self-directed exercises targeting speech issues common to Parkinson's.

Assessment area

The assessment area consists of two types of tasks where the users' speech is audio recorded while completing each task, unconnected speech, and connected speech. The unconnected speech task was derived from Miller et al. (Miller et al., 2007), in which a participant read a single word as it is presented on the screen. The unconnected speech task prompts the user to pronounce ten single words (in no specific order), while recording each one individually as they appear on the screen by pressing the provided start/stop button. Whereas the connected speech task prompts the user with a combination of reading and free speech tasks presented as a scenario to provide structure. As users make their way through the scenario, they are instructed to either read a sentence as it appears on the screen or describe an image or a situation (e.g., describe your favourite pizza). In each scenario the user is instructed to read two samples and provide one free speech sample as the instructions appear on the screen, recording each sample individually using the provided start/stop button. For a consistent recording method and quality, the app prompts participants on every assessment recording to hold the phone "one hand's distance away" from their mouth prior to speaking. When completing the assessment, the user is prompted to either discard the overall recorded 13 samples or upload them to the Speeching service for a crowdsourced review.

Practice Area

This area enables users to exercise and complete their daily tasks at their convenience, and is easily accessible via a designated tab in the Speeching app. Tasks completed in this area are also recorded for practice purposes only, and samples cannot be uploaded for crowd review, but can be played back for users to reflect on. This area aimed to improve two of the most common speech problems that PwP's commonly develop, volume and rate (Canter, 1963). Besides, (Fox, 2002) have expressed the benefits of improving volume to overcome other



Describe your favourite pizza

How understandable people have found you over time.

This rating shows how easy people found understanding what you said.

This average rating shows how loud people think you speak.
You should be aiming for a score between 50 and 80. Try talking a little bit louder.

This average rating shows how fast people think you speak.
You should be aiming for a score between 40 and 60.

This average rating shows how much people think your pitch changes over the course

Figure 3-3 Screenshots for the app; a) Speeching assessment (left) and b) feedback screen (right)

speech and voice issues associated with Pitch variance like intonation. The benefits of carrying out the two types of tasks was thoroughly explained in a video tutorial made by an expert with SLT and Parkinson’s speech and voice issues. The video also explains how to carry out each exercise. For volume exercises the user is required to set a target by taking a short audio recording (i.e., counting to 10) of their loudest speaking voice. The user is then presented with a passage to read while attempting to maintain their volume to an equivalent or higher level than the recorded target. As they read the displayed passage the app measures the intensity of their sound and visualise it on the screen, advising the user when their volume is maintained (green light is shown) or below their target level (red light is shown).

Whereas the second exercise targets improving speech rate, in which users begin by following an auditory metronome and reading aloud one word per beat, encouraging them to develop a habit of slowing their speech down (e.g., WHAT-TIME-WILL-THE-BUS-BE-COME-ING). The metronome rate is adjustable to meet user’s preference and skill level. In its further stages this exercise also enables users to utilise the provided metronome in a more naturalistic way, improving their natural intonation and stress patterns that are so common in everyday speech. In this case the important words are read aloud on the beat to add a natural stress pattern (e.g., what **TIME** will the **BUS** be **COMING**).

Feedback area

In the Speeching app users can track the progress of their speech and voice via the designated feedback screen. Users are presented with statistical score graphs showing their progress in ease of listening (EOL) tasks over time, in addition to their most recent scores (the median) of EOL, pitch, rate and volume tasks (Figure 3-3). To enable users make sense of their scores and improve upon, the app assigns 'goals' to the presented scores. The app prompts its users to aim for a score between 50 – 90 in both volume and pitch tasks, and 40 – 60 in rate tasks. These measures were chosen with the SLT expert help to explore the influence of such in-app's goals recommendations might have on the participants. As such, for participants scoring lower or higher than the threshold range, messages were added beside their scores to suggest how they should modify their speech. For example, when a user receives volume score of 45 the following message is displayed besides the score "This average rating shows how load people think you speak. You should be aiming for a score between 50 – 80. Try talking a little bit louder").

3.4.2 Integration with Crowdsourcing Services

The Speeching solution was integrated into an online crowdsourcing platform (Crowdflower) via the Speeching server. The server was developed in Microsoft C# and provides a web service API that orchestrates the work between the Speeching app and Crowdflower. Once a user uploads their assessment to the server, the server creates a crowd assessment job and posts it to Crowdflower. Each job requires five crowd workers to assess all samples, and their final responses are then aggregated on the server, where the median score (to account for outliers) is also calculated. The median score is then accessed from the Speeching app as the user feedback. While asking for more ratings than five may achieve more accurate responses, it certainly will increase the cost and potentially make it unaffordable for the general public to use.

3.4.3 Micro-task design

Tasks were carried over from the evaluation phase with minimal adjustments. Unconnected speech was designed as a multiple-choice task, requesting crowd workers to select one word out of 10 similar words (e.g., coop, cup, cape, cope) that they thought they heard. Whereas in the connected speech task I reused the ease of listening (EOL) rating as in the evaluation phase, and improved on the measures of pitch variance, rate, and volume. To measure pitch, volume and rate crowd workers are provided with a comparative sample (baseline) to use in their ratings. Unlike the evaluation phase which used random baseline samples, in this phase the baseline sample is the user’s own speech, to indicate progress level in one’s speech. The initial baseline sample is created after crowd workers rate the user’s first speech sample, for volume, rate, and pitch variance on a scale of 0 -100. Besides, to eliminate any confusion around the volume task the volume question was modified to read “enter a number from 0-100 indicating how loud you felt the sentence was, where 0 is 'so quiet I could barely hear them' and 100 is 'very loud'”. In subsequent tasks, crowd workers were presented with the previously rated sample (as baseline), and the median rating that this sample was given for each measure by the previously recruited crowd workers who rated it in order to rate the new sample in comparison. This allowed for quality control within our own analysis, since crowd workers were given a baseline of what a speech sample (rated with a score of 60, for example) sounded like. This design aimed to promote comparable scoring among crowd workers and ensure users obtained scores that were relative to their previous submission.

Name	Age	Years since diagnosis	Speech severity[1]	Main issues[2]	Uploads	Mean range of pitch (SD[3])	Mean range of rate (SD)	Mean range of volume (SD)	Mean EOL (SD) [2]
Aaron	69	10	Moderate	Rate and volume	5	43.4 (18.8)	55.8 (23.8)	50.8 (21.3)	3.0 (1.3)
Damian	52	9	Severe	Slurring, rate and volume	24	27.8 (13.4)	40.7 (20.4)	35.4 (17.4)	2.5 (1.2)
Neil	61	21	Moderate	Breathy quality and volume	2	36.3 (17.0)	40.3 (19.0)	37.5 (17.4)	2.0 (1.0)
Jill	70	5	Mild	Slurring and volume	18	37.0 (16.4)	37.7 (16.8)	39.7 (17.7)	2.4 (1.1)
Jerry	74	8	Severe	Slurring, volume, rate and pitch	39	41.6 (20.8)	51.1 (25.8)	44.7 (21.4)	2.2 (1.0)
Robert	61	11	Moderate	Volume	31	43.4 (18.8)	44.5 (19.0)	50.6 (21.6)	2.8 (1.3)

[1] Participants perception of speech severity

[2] Main issues as reported by participants

[3] Standard deviation.

[4] on a scale 1 – 5 with 1 being most severe

Table 3-2 Speeching participants information and phase 2 quantitative results

3.5 Real World Deployment with PwP's

The Speeching app was deployed in a real-world scenario that aims to enable PwPs monitor and improve their speech and voice. The purpose of this deployment was to evaluate the proposed crowdsourcing approach on people with speech difficulties, specifically PwPs, who could receive and react to the crowd evaluation. For this deployment 6 people with Parkinson's were recruited through Parkinson's UK local groups following a presentation describing the research objectives. Participants of any age or stage of Parkinson's were considered for the study, so long as they reported difficulties with their speech—See Table 3-2 for a profile of each participant and their reported speech issues. Each recruited participant was visited by a member of the research team in his/her home and received a smartphone set up with the Speeching app. Participants received instructions on how to use the app along with a user manual demonstrating step by step how to conduct both assessment and practice sessions. During this initial visit, the researcher also instructed participants to carry out an assessment task to collect a baseline measure of their speech, and to discuss the usability of the app. Participants were made aware that the app does not permit them to retake individual assessment items, but it is their choice whether or not to post their audio-recorded completed session for crowd rating—this was suggested by the SLT expert, since conventional speech therapy techniques often do not allow retries. During this visit, the researcher also explained the feedback process and the time it takes to receive a feedback on posted assessment sessions. The researcher then showed participants how to use the app to practice and described to them the different practice types the app provides.

All participants were given a week to trial the app, and were advised to carry out as many practices and assessments as they wished at their convenience. However, the researcher urged that on at least one day the participants used the practice area and completed one other assessment before the end of the deployment. Furthermore, the participants were informed that completed assessment could be posted any time for crowd ratings during the deployment phase and that they should receive feedback within an hour of posting it. Later, midway point of the deployment, participants were contacted again via telephone to discuss the usability of the Speeching app and to help with any app related issue they might be facing.

On the completion of the deployment phase each participant took part in a semi-structured interview about their experiences with the Speeching app. Interviews lasted on average for 30 minutes with the shortest being 19 minutes and 45 minutes for the longest, and involved open questions on topics related to the app's usability over the deployment week and the impact of the provision of feedback on their speech. Participants described the app's features they liked and disliked and reported the frequency and ease of use. And gave their opinion on anonymous rating and whether or not the provision of feedback was useful and motivation for change. Interviews were audio recorded and were transcribed verbatim for later analysis.

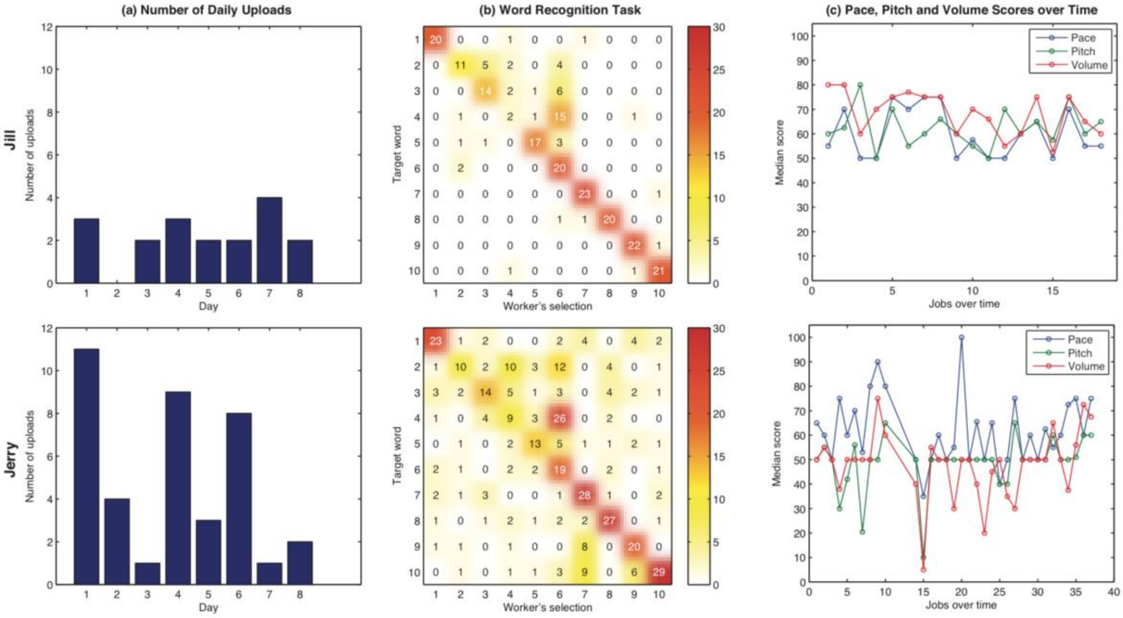


Figure 3-4 Comparative descriptive results for Jill (top) and Jerry (bottom)

On the other hand, I have collected quantitative data during the deployment phase. The data included the number of assessments and speech samples posted for crowd ratings each day of the deployment, and the provided crowd responses for each of the rated measure. For the entire duration of the deployment 122 crowd jobs were created for anonymous ratings. That yield 6,306 completed ratings by the recruited crowd workers, comprising scores for volume, rate, pitch variance, ease of listening and single word recognition.

Table 3-2 shows a full breakdown of participants' level of engagement during the 7 days deployment. Their engagement varied in the frequency they used the app, with crowdsourced assessments ranging from 2 – 39 over the whole deployment period. Since the range of speech issues and severity varied across participants, their data was investigated independently.

3.5.1 Phase 2 Quantitative Analysis

Figure 3-4 illustrates descriptive data comparing Jerry, who claimed to have severe difficulties in multiple speech elements, to Jill, who assumed to have mild volume and voice clarity issues. Figure 3-4 (a) presents the number of daily uploads each participant provided over the course of the deployment; Figure 3-4 (b) shows confusion matrices detailing the number of times that single words were recognized as either the correct target, or another word entirely (from 1-10 the words were, cape, carp, coop, cop, cub, cup, heap, keep sheep); Figure 3-4 (c) shows the median scores for rate, pitch and volume presented to the participants following each upload in the deployment. This stage of the analysis extended the findings from Phase 1 by exploring two more questions. First, how to utilise the single word (unconnected speech) recognition task to inform therapy objectives? For which I calculated the confusion matrix (see Figure 3-4 b) for each of the 2 participants to visualise the crowd performance in recognising single words. The matrix highlights the types of errors that were being made by speakers, as determined by the crowd worker's selections. Second question. what impact does the provision of feedback about perceptual speech measures have on facilitating convenient out of clinic practice of speech? For which, the speakers' I thoroughly analysed participants' scores over the 7 days deployment and the extent to which the crowd workers were giving similar ratings for each measure.

Consequently, the received 5 crowd ratings of each analysed speech sample were aggregated and the mean as well as the mean standard deviation (SD) over each sample was computed. Figure 3-4 illustrates the mean range and mean standard deviation for Jill and Jerry. The mean and mean SD method was adopted since each sample measure received five ratings from five different crowd workers. Besides, the feasibility study at Phase 1 had already established the capability of crowd workers to provide comparable ratings to SLT experts in Parkinson's speech, in the measure of ease of listening, rate and pitch variance. Yet, phase 1 also highlighted issues crowd workers faced in rating volume. As such, it was necessary to examine the proposed "in the wild" recording method (hold the phone "one hand's distance away" from the speakers mouth) in the deployment study, and whether the modified 'volume' question would improve ratings. Therefore, an SLT expert in Parkinson's was consulted to listen to and rate volume on 28 speech samples of the entire collected data set, following the same rating procedure as the crowd workers, before computing the correlation coefficient between expert's ratings and crowd workers'. The subset given to the expert was randomly selected and includes five speech samples from each speaker, of which two samples were excluded from the analysis since they contained no speech.

3.5.2 Quantitative findings

The findings suggest that the proposed crowdsourcing single words recognition task effectively identifies severe intelligibility problems, as in Damien and Jerry, when compared to speakers with milder speech difficulties. Such method helps to identify a variety of participant performance and opens new opportunity for future work aimed to deliver therapeutic intervention to support off-clinic practice of SLT exercises. For instance, a lot of Jill's misrecognised words were down to the misinterpretation of vowel contrast (e.g., cup perceived as cop 15 times) probably due to her accent. In Jerry's case, also, many more crowd workers identified words that are not the same as the target. His confusion matrix shows the bulk of errors were caused, not only by the artefact of his accent, but primarily by the difficulty to articulate word initial sound contrast (e.g., coop perceived as hub four times, or sheep to heap nine time), indicating more sever intelligibility issues. Jerry spoke at a very fast rate, and often ran out of breath, forcing him to speak quietly and to wrongly position his articulators (e.g., tongue, lips), which caused a slurred speech and imprecise consonant production. This opens the space for future improvements to solutions like Speeching, where tasks selection is automated to target the repeated practice of word initial sounds, with the aim to improve intelligibility with no clinical intervention.

On the other hand, the analysis of the perceptual measures of pitch variance, rate and volume showed inconsistent ratings within the measures that participants perceived as their primary speech impairment (see Table 3-2). For instance, Aaron, who reported rate and volume as his primary issues received wider range of ratings in his volume and rate samples, likewise, Robert received the widest range of ratings in his volume. This suggests that inexperienced listeners (crowd workers) had more difficulty rating speech impairment with increasing severity. Such a problem was reported in previous studies like (Landa et al., 2014). who realised that listeners have more difficulty to agree on speech ratings with increasing severity. To further scope this space, it is necessary for the like of this research in the future to identify methods to train crowd workers to quantify more severe problems with the impaired speech. Future work might potentially consider exploring the measure being studied in isolation. To give an idea, raters could be asked to listen to a speech sample recording and only rate the volume in relation to a standardised tone (beep sounding at 70 dB) which they increase or decrease to match the volume in the speech sample. Other methods may explore the potential for listeners to rate pitch variance in speech sample by drawing a line as they listen to the speech adding peaks and troughs as pitch increases and decreases. Finally, to evaluate our solution in response to the issues found from Phase 1 in rating volume, a Pearson's Correlation Coefficient was computed to investigate the correlation between the experts and untrained crowd workers on the measure of volume. The results indicated a moderate, almost high, positive correlation of r=0.57 (Evans, 1996).

Engagement and Cost Analysis

Over the 7 days deployment participants submitted a total of 122 speech assessments for crowd analysis, and 86 crowd workers were recruited to rate them, with an average of 8.9 jobs per worker. Each job took in average 59 minutes to complete from submission of the tasks by participants, to crowdsourcing 5 unique rates, and ending with the provision of feedback. Furthermore, the cost to complete each job was an average of $2.10, that is $0.42 per rate paid to every worker successfully completed the tasks. In total the full deployment of 122 submissions costed only $256.20, that is equivalent to approximately two visits to a specialist in SLT in the UK, who is paid an average of $110 per hour (not accounting for travel time and costs).

3.5.3 Qualitative Findings

Inductive thematic analysis was applied on the qualitative interview data about participants' experiences with the Speeching app following methods recommended by (Braun & Clarke, 2006) to identify themes across the data set. Three major themes were identified after the data was coded at the sentence to paragraph level; appreciation of the anonymity in particular anonymous raters; feedback and self-understanding; and issues with exercising and tasks. These three themes are discussed here in details:

Appreciation of the anonymous crowd

Participants responded positively about utilising crowdsourcing methods to analyse their speech and provide meaningful feedback. There was a common agreement between the participants that people within their social networks are often not good markers of their ability. Damien highlighted how feedback received from related people can be biased in comparison to that received from crowdsourcing methods, saying: *"It was interesting to see how people rate you, because people don't usually tell you what they think"*. The Speeching app was highly appreciated for its capability to deliver how others perceives one's speech, without necessarily having to ask relatives and friends for feedback. Besides, Robert repeated this sentiment: *“sometimes I just talk to people and they just look at me”*. He described how asking others opinion about his speech can cause embarrassment and discomfort, unlike the anonymity of the crowdsourcing method: *“if you're face to face with a person, it can be embarrassing, if they're saying that your speech needs to be improved, it's like, “Yes, okay.” If it is a machine that you know is via a person, I think that's quite nice. There's some kind of validation to it…I know some human is marking the progress.”*. Moreover, Robert found crowd ratings very encouraging to practice more and improve his speech *“it's quite a boost to you in terms of how they understand you, and trying to achieve a better rating.”*

Feedback and self-understanding

The feedback feature was appraised by the participants, who found in it a means to be more conscious about their speech and achieve improvements. Roberts found the crowdsourced feedback incentive to practice more and do better *“I kept wanting to get to 5 [in EOL]. And then speech volume, I wanted to increase that one, as well.”*. Besides, he was also happy about the agility of the app and the time it takes to receive his feedback *“getting it within, say, half an hour, an hour, is good…being so instantaneous”*. Likewise, Damien also used the feedback to challenge himself and improve his scores. His wife explained how he would work harder to accomplish better crowd ratings if he was rated lower in the last assessment *“When he did one and he got the assessment and it was low he would do it straight again to see if he could up it”*. While Aaron, could not use the app frequently during his deployment, this was due to limited internet connectivity available to him. However, he managed to use the app a few times and found the feedback received from the crowd insightful, particularly on the pace of his speech *“I was a bit surprised at the scores of speed...I think that is reflective on my speech at the moment because I speak very quickly”*. And like the other participants, Aaron found the Speeching app useful to help people with speech disorder to self-reflect on their speech *“this tells me that I can improve if I’m willing to change…Being reflective is enough for me”*.

Furthermore, some of the participants enjoyed the listen back function within the practice area. Jill, Neil and Jerry realised they can self-monitor their speech and reflect upon by listening back to their own recordings after completing a practice session. Jill, who was the most passionate about the listen back function, found the function very helpful for practicing and tuning her speech *“it does help you to realize that you're not speaking properly, and for certain words there's no clarity in them, for other people, you know?”*. This function helped Jill to determine words and phrases that were affecting her speech, and enabled her to target certain elements of her speech in order to improve it. Whereas Neil described how the listen back function helped him to comprehend how he sounded to other people *"I thought I was disturbing the house by shouting, I played my voice back and it sounds like I'm whispering"* and encouraged him to improve it, particularly the volume of his speech. Such feedback is particularly important, for it is very common between PwPs to wrongly perceive the loudness of their speech (Ramig et al., 2001), so providing tools to enable PwPs to understand how their voice sounds is particularly promising.

Problems with practicing and tasks

Participants discussed in depth the metronomic pacing (for speech rate) and volume monitoring tasks provided in the practice area of the Speeching app. They identified several shortcomings, in particular with the speech rate exercise: *"He was going faster…He's way ahead of what the beeps* [from the auditory metronome] *were."* (Damien's wife). Likewise, Robert and Neil found it difficult to follow the auditory metronome: *"I didn't like the pacing... I understand it theoretically, but I can't do it practically"* (Robert). Robert also explained how the position of the dB level monitor at the bottom of the app screen made it difficult to monitor the loudness of his speech during the exercise *"The text is here, and the green light's there. So you've got to try and concentrate."*. Participants also described how having control over the content of the practice tasks could increase motivation and enrich the app's engagement level. For example, Aaron wanted to add his material to read, whereas Damien reported that some scenarios were irrelevant to him: *"I wouldn't get on the bus"*. Furthermore, Jerry described the scenarios as just too simple: *"it asks you stupid questions"*. To the contrary, Robert and Jill enjoyed the scenarios considering their day-to-day activities *"they're all interactions you use every day…I go to the paper shop… I say, "Good morning, how are you?" So, it's a set routine"* (Robert), yet both participants noted that richer and more diverse content would be appreciated.

3.6 Discussion

3.6.1 Crowdsourcing the Analysis of Impaired Speech

The feasibility study established that anonymous raters recruited via online crowdsourcing platforms are capable to produce comparable ratings on impaired speech to that of an expert, depending on the quality of the crowd tasks design. For instance, the feasibility study revealed some issues related to the volume task that affected crowd judgements, like the ambiguity of its question and audio quality, which I addressed in phase 2 deployment. The deployment phase improved the volume question and introduced a more consistent way of recording participants speech in natural settings. Moreover, the Speeching app demonstrated the potential and benefits of using crowdsourcing methods for diagnosing elements of vocal impairments. Future work of this kind might explore the potential of offering specialised low-cost and abundant task force to support expert diagnosis for vocal issues. Besides, using binary selection task like the one used in (McAllister Byun et al., 2015) along with more crowd training may well reveal speech and voice issues from data collected in the wild. While other work explored the diagnostic potential of utilising automatic voice analysis. For example, Arora et al (Arora et al., 2015) used voice-based system to diagnose Parkinson’s disease. Likewise, (Zhang, 2017) proposed an affordable machine learning voice-based solution to early prediction of Parkinson’s disease. However, automatic methods, including Zhang’s still lack accuracy and require lots of annotated data to train on and improve. As such, incorporating crowdsourcing methods like Speeching into automatic methods greatly enhances automated algorithms and provides access better to off-clinic SLT level feedback, without the need for SLT resources. Such digital solutions could fill indispensable therapeutic gap, reaching wider vulnerable audience with no access to clinics, especially so when 90% of PwPs experience vocal problems, but less than half of them is thought to have access to SLT (Miller, Noble, et al., 2011).

3.6.2 Trust and appreciation of the crowd

Participants enjoyed and appreciated the implementation of crowdsourcing methods, in particular the genuine human (workers) ratings. They leveraged the provisioned feedback from the crowd to understand how they are heard by others and monitor their progress over the time. This, in particular, is a benefit of the Speeching system, it enabled participants to self-reflect and self-monitor their speech, and supported them to engage more in self-management practices. However, future larger-scale work is recommended to clarify the potential of using the Speeching system as a motivator, and to explore participants reaction and level of engagement if their scores increasingly worsened. Although, degeneration in ability is nearly inevitable concern for PwP, using solutions such as Speeching could be beneficial for users who are determined in their rehabilitation efforts.

Participants also appreciated the level of privacy this app provides. It keeps their identity hidden and provides anonymous crowd ratings to their speech, eliminating any embarrassment surrounding their speech and what others may think of it. Moreover, participants expressed the benefits of obtaining anonymous speech ratings, and found it more reliable than that of a friend or relative (who remain polite) or experts who are trained in listening to impaired speech. The latter is the reason behind choosing the crowdsourcing method to obtain impaired speech measures from non-expert crowd workers, since the ‘familiarity effect’ has been widely evident (Ziegler & Zierdt, 2008; Miller, 2013; Landa et al., 2014). To the contrary though, Aaron feels his close network of relatives and friends would give more truthful ratings as they would be “hard” on him. With that in mind, Aaron’s opinion encourages future work to explore a person’s social capital as raters and motivators, to support the sustainability of the system within healthcare (Morrow & Scorgie-Porter, 2017).

On the other hand, however, there are number of limitations in the Speeching system that shall be addressed in future work. Key limitations encompassing advanced privacy and security concerns around the anonymity of the crowd and their access to personal data submitted by PwPs. These concerns were identified by Lasecki et al. (Walter S. Lasecki, Teevan, & Kamar, 2014) who found that crowdsourcing systems are vulnerable to unauthorised data mining and malicious manipulation. They propose a crowd task design that identifies and leverages reliable workers, to find and alert job requesters to data which might be subject to malicious attack. That is a key concern that requires more attention in future work. One potential solution might be to trade-off anonymous crowd in favour of connected individuals, leveraging charities and support groups as well as participants’ close social network. Still, the reliability and the quality of the ratings obtained from connected individuals, and the added benefits and ethical financial implications (Dolmaya, 2011) deserve further investigation.

3.7 Conclusion

The work I report on here acts as a first step for understanding the ways in which a crowd of non-experts might provide useful and timely feedback to support personal care around speech. Through the development and evaluation of Speeching I have highlighted the validity of using a crowd as lay listeners and raters of Parkinson’s speech, as well as the potential utility and acceptability of the system to people with Parkinson’s. Future work is needed to evaluate the system with a larger group of individuals with a wide range of speech difficulties. Furthermore, longer trials will enable us to study whether the gains and new practices experienced during these trials are sustained over extended periods of time.

4 Designing for Quality Eye Tracking Data in Real-World Settings

For decades, Eye tracking has held the promise as the ultimate human computer interface. It is the method of measuring an individual's eye movement to identify both where a person is looking (*gaze*) and the sequence in which the person's eyes are shifting from one location to another. And it tells us about fixed points of interest in which the person's eyes are relatively stable (*fixation*) for a minimum duration of 100-200 ms (Jacob, R.J.K., and Karn, 2003), as well as the rapid eye movements (*saccade*) from one fixation to another. While eye tracking techniques are diverse, three have emerged as the predominant techniques and are broadly used in research and commercial applications (Majaranta & Bulling, 2014b). Electrooculography EOG eye tracking has been used for ophthalmological studies; it enables researchers to measure relative movements of the eyes with high temporal accuracy using electrodes attached to the skin around the eyes (Bulling & Gellersen, 2010). The two other techniques are video-based and have a lot of properties in common: Videooculography (VOG), and Infrared induced corneal reflection IR-PCR. Both rely on the detection of pupil positions to estimate gaze positions from images typically delivered by off-the-shelf components and video cameras. While VOG provides acceptable accurate point of gaze measurements (e.g., about 4° (Hansen & Pece, 2005)), the IR-PCR, due to the additional IR-induced corneal reflection, provides higher accuracy of up to 0.5° of visual angle (Majaranta & Bulling, 2014a). As such, video-based, and more specifically IR-PCR, eye tracking has emerged as the preferred technique for developing applications like eye-aware or attentive user interfaces, usability studies or gaze-based interaction, and commercially like in marketing research.

Unlike video-based eye tracking, EOG provides lower spatial point of gaze accuracy and its signals are subject to signal noise and artefacts (Bulling & Gellersen, 2010). Resulting in EOG eye trackers to be less suited for real-world settings. On the other hand, lighting conditions mostly and highly impacts current video-based eye tracking, contrary to EOG, making it challenging for video-based techniques to fully function outdoor. Still all current techniques are susceptible to calibration drifting, particularly if recorded in mobile settings.

This chapter focuses on the video-based eye tracking, for its high accuracy when conducted in lab settings, low cost to construct using off-the-shelf components, and less intrusive than EOG eye tracking. In the following CrowdEyes study I investigate the potential of harnessing the crowd to overcome video-based eye tracking technology's major challenges (i.e., sunlight, eye-makeup) when recording in real-world settings, and report on the associated costs and quality. In the following sections I will refer to video-based eye tracking as "eye tracking".

4.1 CrowdEyes: Crowdsourcing for Robust Real-World Mobile Eye Tracking

Like all major technologies eye tracking has faced many challenges in order to meet with industry expectations. Its design has evolved noticeably from stationary, large, and heavy head-mounted devices to small, mobile spectacle-like head mounted devices. But despite the remarkable improvement in the quality of data collection and the technology's accuracy and robustness, thus far, the technology has suffered from a number of drawbacks when it comes to practical use in real-world settings. Common challenges, such as high levels of sunlight, eyewear (e.g., spectacles or contact lenses) and eye make-up, result in visual data noise, causing failure in the automated pupil detection processes. Which in return, undermines their utility as a standard component for mobile computing, design, and evaluation. To work around these challenges, I introduce a crowdsource solution that aims to increase the accuracy and robustness of eye tracking technology—I named it CrowdEyes. I present a pupil localisation task design for crowd workers along with a study that demonstrates the high-level accuracy of crowdsourced pupil localisation in comparison to state-of-the-art pupil detection algorithms. I further demonstrate the convenience of our crowdsourced solution, and introduce analysis pipeline in a fixation-tagging task. This chapter validates the accuracy and robustness of harnessing the crowd as both an alternative and complement to automated pupil detection algorithms, and explores the associated costs and quality of the proposed crowdsourcing approach.

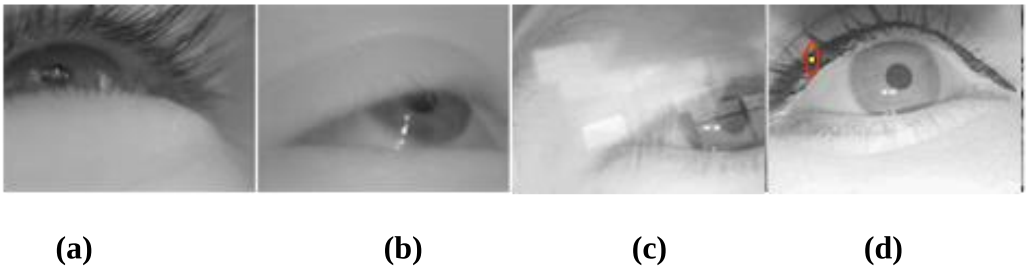


Figure 4-1 Challenging pupil images in real-world scenarios: (a) natural light reflection, (b) droopy eyelids, (c) spectacles and (d) eye make-up. The red circle in (d) indicates false pupil detection when wearing eye make-up.

4.2 Introduction

Video-based eye tracking relies on image processing computer algorithms for the detection of pupil positions to estimate gaze positions from images typically delivered by off-the-shelf video cameras. Harnessing the potential of eye tracking, video-based has been used in a wide range of research, commercial and non-commercial applications; from attentive user interfaces, like in driving or aviation, to detect and warn on the presence of fatigue or high-workload conditions by examining eye movements (You et al., 2012; Singh et al., 2011) (see (Peißl et al., 2018) for in-depth review); to skills assessment (e.g., assessing drivers and cyclists hazard perception skills (Mackenzie & Harris, 2014; Mantuano et al., 2016)); in teaching clinical anatomy (Sánchez-Ferrer et al., 2017) and clinical diagnosis (e.g., in Dementia (Pavisic et al., 2018) and autism (Klin et al., 2002)); as well as wayfinding research (e.g., to enrich direction guidance systems in public infrastructures outdoor (Schrom-Feiertag et al., 2014) and indoor environments (Ohm et al., 2014)). In the field of HCI, eye tracking has also been used to evaluate technologies, such as the usability and safety standards for in-vehicle information system (Purucker et al., 2017) and display screens in public spaces (Dalton et al., 2015).

Despite their diverse forms and considerable potential for applications, most eye tracking studies are conducted in artificial or semi-artificial environments—either in controlled environments (e.g., laboratories) or in virtual reality. Although eye tracking performs comparatively well in controlled environments, the technology fails considerably under real-world settings (Singh & Singh, 2012; Evans et al., 2012; Bengoechea et al., 2012; S. Cheng et al., 2015). Such failure is greatly attributed to highly challenging pupil detection factors, including: i) unrestrained lighting (Evans et al., 2012); ii) pupil occlusion by the eyelid and eyelashes (Evans et al., 2012); iii) eyewear (Fuhl, Tonsen, et al., 2016) (e.g., spectacles or contact lenses); iv) eye make-up (Fuhl, Tonsen, et al., 2016); and v) motion-blur (Li et al., 2006) (e.g., from fast eye movements during saccades). In particular, unrestrained lighting conditions, while travelling and recording outdoor, cause visual-noise and differences in contrast that constraint the effectiveness of automated pupil detection algorithms. And since most video-based eye tracking operates in the infrared light spectrum, in various mixed lighting or outdoor conditions where infrared light (e.g., sunlight) floods the eye camera(s) (Figure 4-1a) the automatic detection of pupil features becomes difficult. Besides, medical conditions like Ptosis (pathologic eyelid drooping) cause the eyelid to partially obstruct the eye pupil (Figure 4-1b), making it difficult to automatically detect pupil features. Likewise, goggles (Figure 4-1c) and eye make-up (e.g., mascara) (Figure 4-1d) result in substantial and varied forms of reflections and generally high amounts of noise, creating dark curvy pupil-like edges, which make it difficult to detect the actual pupil.

This study proposes and investigates an alternative approach that leverages the power of crowd to achieve robust and accurate eye tracking measures. Whereas eye tracking is often restricted to controlled environment due to the challenges facing automated pupil detection methods, our proposed approach, *CrowdEyes*, offers mobile unobtrusive eye tracking with all standard metrics independent from most common automated pupil detection challenges. *CrowdEyes* follows a procedure to localise the pupil position without automated pupil detection algorithms. It begins by converting the captured video of the eye into an image sequence, marking key frames, and crowdsourcing the localisation of the pupil position in these frames. The localised pupil centre positions are then used to generate standard eye tracking metrics (e.g., gaze and fixation positions and durations, saccades) using the according methods from the open-source eye tracking platform *Pupil* (Kassner et al., 2014). Crowd workers can then semantically label generated fixations. *CrowdEyes* is envisaged as a runtime tool operating on mobile wearable (or screen-based) eye trackers. Though, to accommodate for the present technical restraints of mobile devices, our initial proof-of-concept data is processed offline after collection.

My contribution is twofold. For the crowdsourcing research community, I: i) investigate and evaluate the design of crowdsourcing tasks and strategies to affordably improve mobile wearable eye tracking technologies; ii) propose a crowd task quality assurance method that enables workers to evaluate and refine their own entries; and iii) provide experimental evidence that demonstrates that such quality methods also motivate workers to improve their accuracy. Second, this research contributes to mobile wearable eye tracking by: i) working around automated pupil detection challenges in real-world settings; ii) reliably localising pupil centre positions; and iii) providing a tool for the mobile eye tracking research community to generate training datasets for pupil detection algorithms on demand by harnessing the crowd.

4.3 Literature Review

The work presented in this chapter looks over various areas of research, including computer vision-based eye tracking, self-reporting based eye tracking, and crowdsourcing as an alternative and complement to automated detection and recognition systems.

4.3.1 Computer vision-based eye tracking

Over the past decade, there has been a steadily growing amount of research on the design and development of eye tracking technologies. Studies have investigated and evaluated novel pupil detection algorithms (e.g., (Swirski et al., 2012; Fuhl et al., 2015; Valenti & Gevers, 2012; Timm & Barth, 2011)) and new calibration techniques (e.g., (Lee et al., 2013; Sugano & Bulling, 2015; Villanueva & Cabeza, 2008)), seeking robust and accurate commercial eye trackers capable to operate under the aforementioned conditions. But, since the vast majority of the state-of-the-art pupil detection algorithms are based on edge filtering methods (Fuhl, Tonsen, et al., 2016), they are very susceptible to failure under real world settings. Tonsen et al. (Tonsen et al., 2016) investigated five state-of-the-art pupil detection algorithms: *Gradient* (Timm & Barth, 2011), *ExCuSe* (Fuhl et al., 2015), *Isophete* (Valenti & Gevers, 2012), *Swirski* (Swirski et al., 2012), and *Pupil-Labs* (Kassner et al., 2014) and evaluated their pupil detection success rate using the large and challenging real-world *Labelled Pupil in the Wild* (LPW) dataset (Tonsen et al., 2016) of 130,856 eye video frames from 22 participants.

The outcome clearly shows that, despite advances in general pupil detection accuracy, the algorithms still yield inadequate pupil detection rates under real-world settings. In particular eye make-up was found to be a key issue for pupil detection, with 60% of the data for participants wearing eye make-up yields no detection—this is due to eye make-up creating dark curvy eyelashes or spots around the eye region that confuses such algorithms. As a result, eye tracking metrics are considerably affected by substantial data loss and inaccurate pupil detection (Holmqvist et al., 2012). Whereas inaccurate pupil detection reduces dwell time (gaze time spent in the same area of interest), failure to detect the pupil decreases the number of fixations and increases fixation duration (Holmqvist et al., 2012). Recently, Fuhl et al. developed a novel pupil detection algorithm named *ElSe* (Fuhl, Santini, et al., 2016), which outperforms five common algorithms (*Starburst* (Winfield & Parkhurst, 2005), *Pupil-Labs*, *ExCuSe*, *Swirski*, and *Set* (Javadi et al., 2015)) in an evaluation study (Fuhl, Tonsen, et al., 2016) that leveraged a large-scale multiple dataset of previously annotated images (from (Tonsen et al., 2016), (Fuhl, Santini, et al., 2016), (Swirski et al., 2012), and (Fuhl et al., 2015)). However, while *ElSe* slightly advances on the performance, it fails to robustly and accurately detect pupil positions in the presence of poor or mixed lighting conditions, reflections, or eye wear (e.g., spectacles, make-up) (see (Fuhl, Tonsen, et al., 2016) for detailed results).

Due to the current limitations of pupil detection algorithms, outdoor studies are often avoided, and participants wearing spectacles, eye make-up, or who display Ptosis are commonly excluded. This leads to significant limitations on applications of eye tracking and constraints in how and where eye tracking can be deployed.

4.3.2 Self-reporting based eye tracking

Studies have investigated alternative methods for determining gaze directions without the use of eye trackers. In a crowdsourcing study, (Rudoy et al., 2012) introduced a self-reporting approach to collect gaze direction data from crowd workers online. Workers were instructed to watch a short video followed by a grid screen with unique codes, which was briefly displayed at the end of each video. Workers were then instructed to type in the code they saw first to determine their last gaze direction. While this method determines the direction of the last gaze, it does not determine the direction of all other gazes over the period of watching the video. Likewise, Cheng et al. (S. Cheng et al., 2015) developed a self-reporting gaze direction approach based on computer pointing devices (e.g., mouse or trackpad). Unlike (Rudoy et al., 2012), Cheng's method determines the initial and last gaze directions as well as all in between gaze directions from crowd workers online. In Cheng's study participants were asked to view an image frame followed by a 9×9 grid image. The workers are expected to remember the sequence in which they shifted their sight (gaze) from one point on the viewed image to another until the grid is displayed. Workers are then required to recall the locations and sequence of their gaze, and click the relevant grid cell.

Despite good results from these self-reporting gaze tracking methods, they still suffer from a number of drawbacks, including: i) intrusiveness and full reliance on participants to self-report; ii) a rise in cognitive load that could influence participant responses; iii) heavily dependent on participants' memory, particularly when recalling all gazes and their sequence; iv) limited applications to on-screen only; and v) incapable to determine other important eye tracking measures (e.g., fixation durations and saccades).

As such, robust, unobtrusive, and pervasive real-world eye tracking methods remain an unresolved challenge. Since self-reporting methods suffer number of substantial drawbacks, and automated pupil detection algorithms are insufficient in real-world settings, our approach proposes a workaround solution by harnessing the crowd.

4.3.3 Crowdsourcing-based image annotation methods

Various literatures have studied how to use crowdsourcing to supplement inadequate automated algorithms. Studies have revealed that object identification is, to date, still a challenging task for automated algorithms, while crowd workers do remarkably well in similar tasks. For example, Su et al. (Su et al., 2012) leveraged crowdsourcing to gather quality image annotations (e.g., drawing a bounding box around each animal in an image) for more than 1 million images. Such a dataset could then be used in machine learning to train automated object recognition algorithms. The outcome of Su et al. study is very promising, the evaluation shows 97.9% of the images were successfully annotated by crowd workers with a very high accuracy of 99.2% (Su et al., 2012). Likewise, Hipp et al. identified cyclists, pedestrians and vehicles in publicly available webcams in two road intersections by harnessing crowd workers to annotate the records (Hipp et al., 2015). They report a high inter class correlation (ICC) between workers, equivalent to the ICC of two trained researchers who completed the same annotation tasks.

Such results demonstrate the potential of using crowdsourcing to localise the eye pupil. However, in contrast to the latter two studies that evaluated target classification in an image or fitting a bounding box around it, our study focuses on accurately localising the centre of the eye pupil ellipse in images in noisy real-world settings (i.e., blurred images, or contain high light reflections). But to assert that crowdsourcing is a good candidate to supplement eye tracking automated pupil detection algorithms, further study is required to address major challenges, such as, the enormous number of images to be crowdsourced and the associated processing time and costs, as well as the high level of accuracy to localise the center of the eye pupil.

I address these challenges by developing frame selection methods to exclude highly similar frames but keep one, designing crowd tasks to localise pupil center and label the targets being gazed upon, and introducing a quality assurance approach based on self-validation and refinement. The solution is evaluated against the Labelled Pupils in the Wild (LPW) dataset and I report measures for localisation accuracy, robustness, processing time and costs.

4.4 Method

I expand recent work that has investigated the utilization of crowdsourcing in object labelling (Russell et al., 2005; Su et al., 2012; Hipp et al., 2015) and recognition (Sinha et al., 2006), and the gathering of (Xu et al., 2015) as well as the self-report on (S. Cheng et al., 2015; Rudoy et al., 2012) eye tracking data. *CrowdEyes* utilises a conventional mobile head-mounted eye tracker, which, unlike Rudoy et al. (Rudoy et al., 2012) and Cheng et al. (S. Cheng et al., 2015), doesn't require participants to self-report (i.e. specify where they gazed) or interfere for data collection (i.e. participants complete some tasks in order to know where they gazed). Our system expands the *Pupil* open-source eye tracking software platform (Kassner et al., 2014) and comprises of two key elements: i) a low cost DIY mobile eye-tracker based on the *Pupil* open-source platform; and ii) a set of crowd tasks to accurately localise the centre of eye pupil as well as calibration target (e.g., marker). CrowdEyes leverages existing commercial crowdsourcing platform CrowdFlower to recruit crowd workers who will complete the system's crowd tasks. I utilised the open-source heterogeneous head-mounted mobile eye tracking LPW dataset (Tonsen et al., 2016) recorded under natural (indoor and outdoor) conditions to evaluate the accuracy and robustness of our approach. The outcomes were then compared to (Tonsen et al., 2016)'s reported measures of five state-of-the-art automated pupil detection algorithms (*Pupil-Labs* (Kassner et al., 2014), *Isophete* (Valenti & Gevers, 2012), ExCuSe (Fuhl et al., 2015), *Gradient* (Timm & Barth, 2011) and *Swirski* (Swirski et al., 2012)). I also establish and demonstrate a novel method to crowdsourcing motivation and quality assurance based on a worker response self-validation and refinement cycle. Furthermore, our study demonstrates the potential for *CrowdEyes* to be extended to include crowd data analysis tasks that are conventionally very laborious and time-consuming; in this case, the annotation and labelling of fixations. Accordingly, our contribution is to establish and demonstrate crowdsourcing-based methods for cost-effective, robust, accurate and extensible approaches to ubiquitous mobile eye tracking.

4.5 Crowdeyes design considerations

The design and architecture of *CrowdEyes* consists of two major elements. The first element is the use of existing off-the-shelf image capture hardware and open-source eye tracking software. The second is a set of crowd task design, which accommodates crowdsourcing platform constraints, data types, as well as crowd response quality to localise captured eye pupil, then annotate gaze and fixation points. This section seeks to highlight the key factors that have influenced the design of *CrowdEyes*.

4.5.1 The eye tracker gadget

To date, commercial mobile eye tracking systems are exclusive to a small fraction of the market, limiting its applications and use. Costs, to only obtain the gadgets, range from US $10,000 to $30,000 (S. Cheng et al., 2015). While on the other hand, it is possible to produce DIY head-mounted eye trackers to run by available open-source eye tracking platforms. Since eye-tracking software are computationally intensive (e.g., memory, CPU), both mobile devices and portable PCs are inadequate to perform the technologies essential operations. At the time of conducting this research, mobile devices lack the support of multiple concurrent camera captures (eye and world), and off-the-shelf portable PCs are inadequately powerful to accommodate all eye tracking requirements (e.g., concurrent camera capture, pupil detection and gaze mapping), a workaround is required to provide robust DIY mobile eye tracking.

4.5.2 Software and real-time performance

In this study, I have leveraged and improved on the well-established and widely used platform, *Pupil* (Kassner et al., 2014), as a common practice to reach wider range of users. Its open-source nature has enabled us to modify how it functions and to introduce new features, such as the proposed crowdsourcing pipeline for localising the centre of eye pupil and calibration targets, and for annotating what a user gazed or fixated on. Since eye tracking processes like real time pupil detection demand a lot of computation, all features other than recordings were turned off throughout the eye tracking recording sessions, allowing the utilisation of affordable pocket PCs.

4.5.3 Crowd tasks

Data volume

The proposed crowdsourcing pipeline, if run by brute-force localising pupil and annotating fixations will incur a high cost. This is because there are (ex. 30Hz camera) 30 frames captured per second, there are roughly 108,000 frames to process by crowd workers. Both world and eye scenes must be recorded during eye tracking sessions, so the number of captured frames is doubled. Nevertheless, at such high frame rates, recordings contain redundant frames where the target (e.g., pupil or fixation) has not moved significantly. Consequently, to reduce the overall running costs to a minimum, redundant frames must be identified and alienated before performing the crowdsourcing tasks.

Presentation

CrowdEyes proposes two tasks to complement eye tracking detection algorithms accurately and robustly. The first task is to localise the target centre (pupil or calibration target), while the second is to validate and refine rejected crowd inputs. However, it is necessary to carefully design the localisation tasks to avoid increasing cognitive workload, which might influence completion time and decisions. As such, localisation tasks must i) keep to the minimum the time and effort required to visually locate the target's centre across presented images, and ii) minimise page scrolling as well as mouse movements from one image to another. Moreover, the validation and refinement task (for rejected and poor crowd inputs) should be designed to enable a quick overview of all workers' annotated images and provide easy access to those that require refinement.

Quality vs. costs

While accurate target localisation is key for robust eye tracking, and since CrowdEyes proposes leveraging commercial crowdsourcing platforms, it is crucial to keep data processing costs low. Existing crowdsourcing platforms (e.g., CrowdFlower, Amazon Mechanical Turk) provide basic built-in quality control measures, with most common methods being test question injection, and aggregation of multiple worker judgments. However, test questions may become easily detectable, giving worker's sheer amount of likewise completed tasks, while on the other hand multiple judgments aggregation method raises the data processing costs. Accordingly, CrowdEyes will have to introduce additional quality measures to guarantee not only high quality but also low cost. Moreover, crowdsourcing platforms empower requesters to either accept workers' responses or reject them if they do not achieve the required minimum quality threshold. At the same time, it is unfair to instantly reject workers who spent time and effort completing the tasks but did not achieve high accuracy, especially when centre localisation tasks require highly accurate, yet challenging, responses. Besides, applying stricter quality

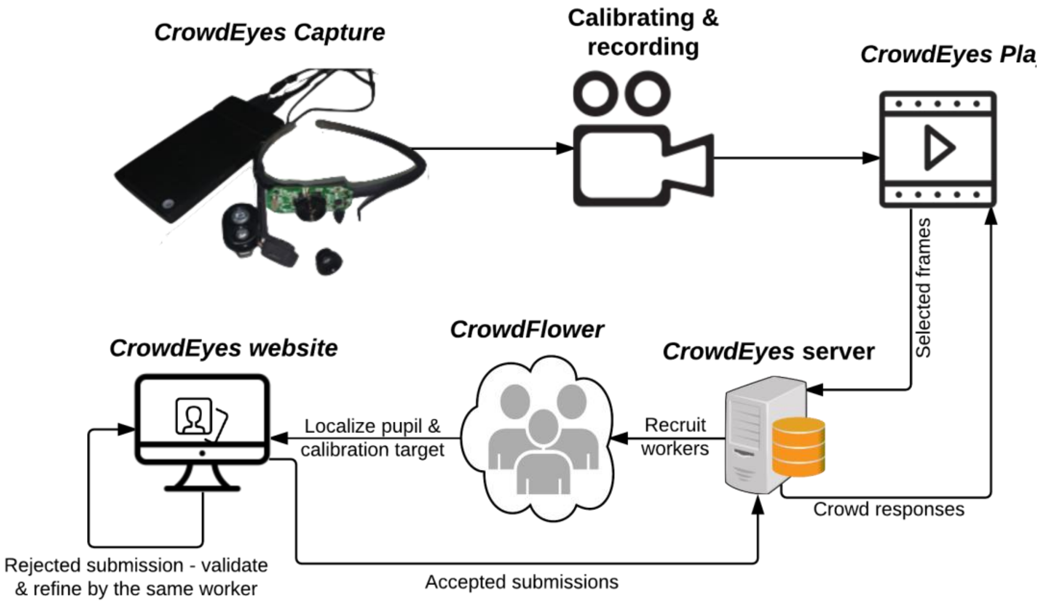


Figure 4-2 CrowdEyes system architecture

measures and rejecting workers failed to meet the required accuracy from the first try will also result in additional costs. And since crowdsourcing commercial platforms give limited control over the process pipeline and quality measures, *CrowdEyes* instead direct recruited workers from the commercial crowdsourcing platform, CrowdFlower, to complete tasks away on CrowdEyes tasks website. Unlike in crowdsourcing platforms, CrowdEyes tasks websites does not reject workers who fail to achieve high quality from the first try. Instead, it gives such workers further opportunities to validate and refine their responses before receiving their reward.

4.6 The CrowdEyes System

The *CrowdEyes* system is composed of: (i) the *Pupil* open source eye tracking software, (ii) a 3D printed head-mounted eye tracker frame fitted with two low cost off-the-shelf web cameras (30Hz) to capture the eye and world scene (Figure 4-2); (iii) portable system processing hardware in the form of a portable pocket PC (Figure 4-2) running Ubuntu 16.04, and a Bluetooth remote button to control the start and end of calibration and recording sessions; (iv) software plugins that mediate between the eye tracking software and a crowdsourcing server; and (vi) the crowdsourcing server where recruited workers will be redirected to complete CrowdEyes tasks. The total cost of construction of the eye tracker is approximately US $270 (not including crowdsourcing costs).

The *CrowdEyes* system has extended the open-source *Pupil* platform and integrated two new components, *Capture* and *Player* both written in Python. *Capture* is a lightweight plugin developed to capture eye and world scenes. It considers the current performance limitations of pocket PCs, and disables *Pupil's* functionalities (i.e., runtime detection processes) other than video capture. *Capture* also saves information about the start and end time of each calibration procedure. On the other hand, *Player* is the plugin that processes *CrowdEyes* captured data offline by harnessing the crowd. When recording is complete, *Player* communicates with the crowdsourcing server to perform two crowdsourcing tasks. Whereas the first is mandatory and is to localise the pupil and calibration target; the second task is optional and is to label the detected fixations. The utilisation of *Pupil's* open-source software is crucial to the *CrowdEyes* system. While the *CrowdEyes Player* carries out the localisation process, *CrowdEyes* also leverages *Pupil* software to instantly access standard eye tracking functionalities, like gaze detection, saccades, as well as fixation positions and durations.

The crowdsourcing server manages the crowdsourcing process pipeline and quality measures of pupil and calibration target localisation in addition to fixations labelling. It consists of three components: i) an online web service that facilitates the communications between *Player* and CrowdFlower, and manages worker recruitment to the platform; ii) a web application where the recruited workers carry out the required tasks; and iii) a database server storing the responses gathered from the crowd.

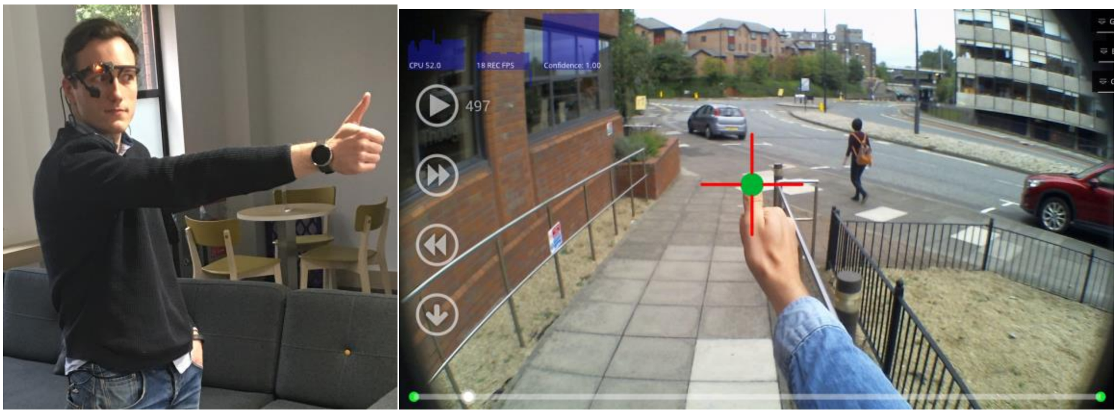


Figure 4-3 Calibrating while looking at the thumbnail (left), and a player screenshot with overlaid crosshair gaze position (right).

4.7 Data capture

Just like conventional eye trackers, recording with CrowdEyes begins with a user-controlled calibration process, which in CrowdEyes begins automatically when powering up the processing unit. However, it has become evident from the initial trials that, in real-world scenarios, computer vision algorithms not only fail to detect difficult targets like the centre of eye pupil, but also machine-known calibration targets. Giving the required high precision outcome of calibration process, and since computer vision algorithms are not reliable in real-world scenarios, the nature of CrowdEyes means that any identifiable object can be used as the calibration target for workers to localise its centre. Consequently, the eye tracker user (wearer) may carry out the calibration process based on features found in the surrounding environment, like a wall corner, or (conveniently) their own thumbnail (Figure 4-3). For instance, a wearer can complete calibration by focusing their gaze at the nail of their thumb while moving their head (thumb-static), or alternatively, keeping their head static while visually following their moving thumb (head-static), such that the target occupies different positions in their visual field. But just like in the typical 9-point calibration method, the wearer is required to make a short pause between each calibration movement. This allows for the collection of a sufficient number of calibration samples, and ensures target is fixated on accurately. The calibration process takes on average one minute depending on the wearer and how many pauses (points) they cover. When completed the wearer clicks a Bluetooth connected remote button to mark the end of the calibration process and the beginning of the eye tracking recording session. While recording, CrowdEyes imposes no further constraints. *CrowdEyes* enables the wearer to wear their spectacles, contact lenses, and eye make-up, and to record under any illumination level and under other uncontrolled real-world conditions. To stop recording the wearer clicks the remote button once more, which marks the end of recordings and powers off the processing unit.

4.8 Pupil and Calibration Target localisation

The localisation of the centre of pupil and the centre of the calibration target is performed post-hoc after the recording session is complete. It is performed by *CrowdEyes Player* and carried out in three stages: i) frame selection; ii) per frame pupil and calibration target localisation; and iii) gaze mapping.

4.8.1 Step1: Frame selection

The *CrowdEyes Player* plugin decomposes the recorded videos into single frames, after identifying calibration and post-calibration recording session (based on the remote button markers). Since flat decomposition of videos yields a large number of single frames, *CrowdEyes* classifies similar frames (e.g., wherein the pupil has not moved significantly) and chooses the mid-point frame for analysis by the crowd workers. Unlike (Laput et al., 2015), applications are limited to almost-static environments, and a frames similarity check is performed periodically (every n-minutes) on a cropped part of the frame, *CrowdEyes* captures and looks for a rapidly moving eye pupil in changing environment (e.g., lighting reflections). Consequently, *CrowdEyes* constantly checks for similarities within all sequential frames. Whereas (Sandhu & Anupam Agarwal, 2015) searches for substantial differences amongst video frames to summarise it, *CrowdEyes* focuses on eye pupil position and searches for minor changes between sequential frames. And in contrast to both (Laput et al., 2015; Sandhu & Anupam Agarwal, 2015), *CrowdEyes* utilises the *multi-scale structural similarity index* (MS-SSIM) method (Wang et al., 2003) for sequential frames, giving more weight to changes in pupil position than lighting reflections and other irrelevant noise factors.

MS-SSIM compares changes in luminance, contrast and correlation between two images, and repeats it over multiple scales of the original images (i.e., 0.75x the original size, 0.50x, etc). It has a maximum value of 1, which indicates two images are identical, while a value of 0 indicates no similarity. Since MS-SSIM is a multi-scale processing method, it requires substantial processing power and time to compare thousands of eye tracking pupil-frames with each other. Since *CrowdEyes* is only interested in changes to pupil positions, *Player* simplify

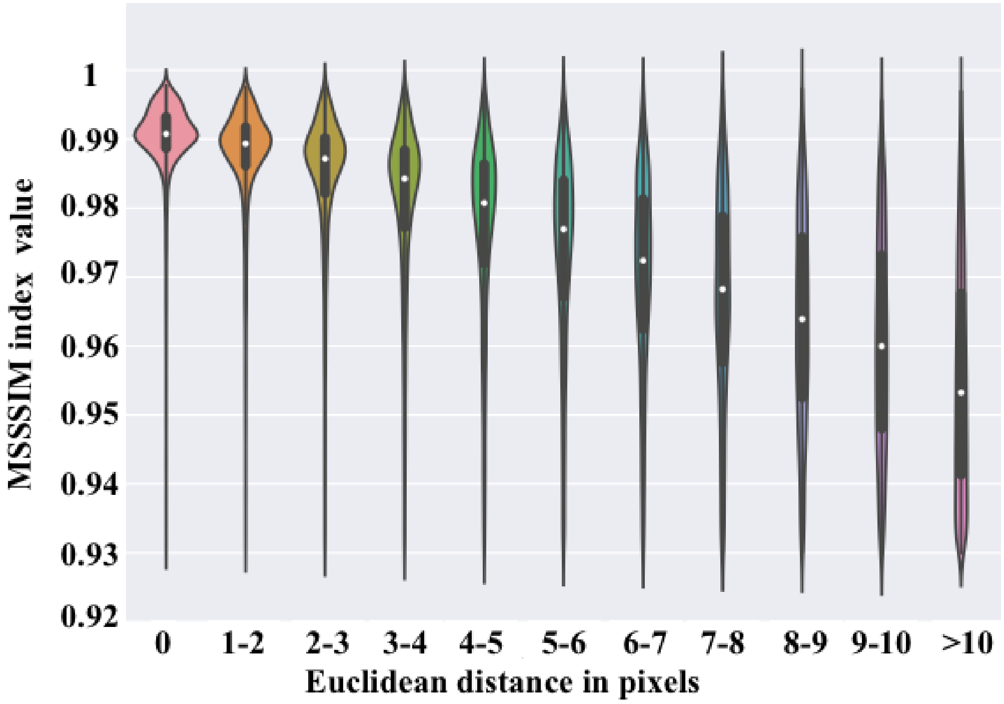
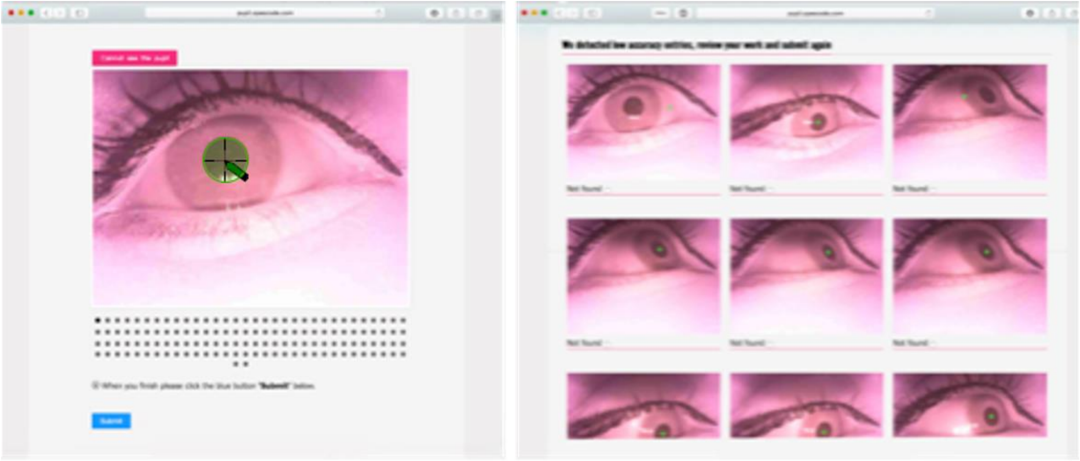


Figure 4-4 MS-SSIM distribution

all frames, blurring and converting them to grey scale and resizing them down to 160×120px before conducting the similarity test. Sequential frames with MS-SSIM values >= 0.98 are clustered and the middle frame (by MS-SSIM value) is added to the crowd job list. The violin plots in Figure 4-4 illustrate the distribution and probability density of MS-SSIM at different values for changes in pupil position (in pixels) in sequential frames using the LPW dataset (Tonsen et al., 2016). The plots indicate the uppermost probability density amongst sequential frames is when the MS-SSIM is > 0.985 with no distance differences. Moreover, the plots also suggest the distance difference between pupil positions in most of the consecutive frames is less than 5px with MS-SSIM value greater than or equal to 0.98. Once the similarity test is complete and frames are clustered, *Player* prepares the localisation crowd job—a set of selected frames and the related crowd task description and configuration (i.e., pupil or calibration target localisation, payment in cents, number of judgments)—and then submits it to the crowdsourcing server for processing by the crowd.



☑ When you finish please click the blue button "Submit" below.

We detected low accuracy entries, review your work and submit again

Figure 4-5 CrowdEyes web pages to localise pupil (left) as well as validate and

4.8.2 Step 2: Pupil and calibration target localisation

Once the server receives the job from *Player*, the server recruits workers from CrowdFlower to complete the tasks simultaneously on the *CrowdEyes website*. Workers are asked to localise the centre of the pupil or the calibration target for 130 consecutive images (640×480px) per task (including 30 gold standard test images), clicking on the corresponding point in the image (Figure 4-5 left). To simplify the localisation process, I have replaced the default mouse cursor with a crosshair pointer surrounded by a green circle (Figure 4-5 left), making it visually easier for workers to identify the centre. And to address the challenge of annotating many images as quick as possible (keeping annotation cost minimum) *CrowdEyes website* presents all task's images in a slider, one image at a time. Once a worker clicks on the anticipated target centre, the next image is randomly selected—impose random mouse movement—to locate the next target centre and so on.

Quality throughput

While reliable eye tracking depends on high quality pupil and calibration target localisations, crowd workers are usually after maximising their financial reward. As such, workers are more inclined to favour speed over quality to increase their daily monetary compensation, which usually result in various mistakes. On the other hand, strict quality measures increase the chance to block honest workers who make unintentional mistakes while completing a task, resulting in unfairness towards workers and incurring further expenses for requesters. Since high quality localisation is essential, in such context, and low cost as well as fair payment are

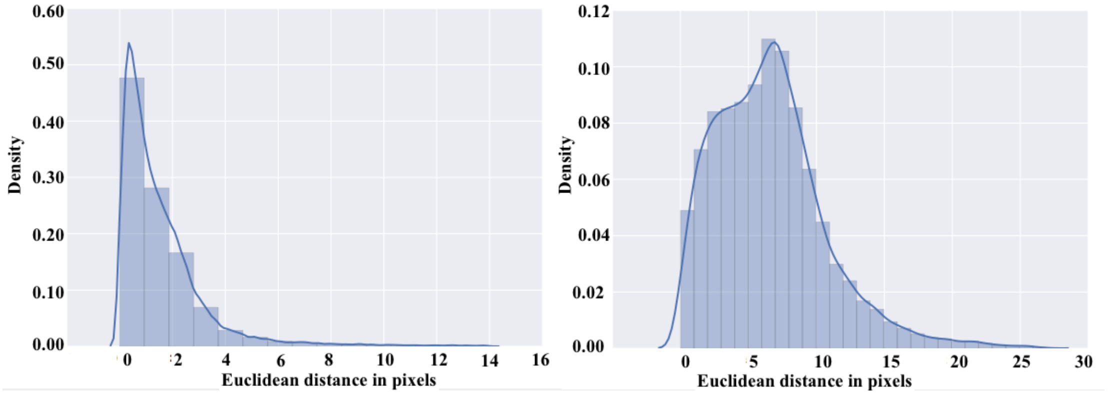


Figure 4-6 The density of Euclidean distance between consecutive frames (left), and sequential MSSSIM-selected frames (right).

necessary, *CrowdEyes* follows three quality control methods:

Injecting gold standard reference images: In crowdsourcing it is a common quality check practice to inject questions with known answers in order to test and track workers performance. *CrowdEyes*, as such, insert images with known target centres (eye or calibration frame for relevant localisation task) for which a worker must achieve an accuracy (Euclidean distance from the predetermined target centre) of no more than 10 pixels. Besides, to keep crowdsourcing costs minimal, *CrowdEyes* builds on a single judgement, instead of the traditional multi judgements per task, but increases the gold standard data percentage. Each task includes 30% gold standard sequential images selected randomly from our manually pre-labelled images pool. The percentage of ground truth data is purposefully high, so each test image is judged once (no repetition), making it impossible to identify them among the others, and to compensate the single judgment per task.

Euclidean distance between two sequential clicks: As the target moves swiftly, its centre shifts gradually in consecutive frames. Accordingly, computing the farthest Euclidean distance between two consecutive frames can be used as a quality measure. It will detect random as well as robot responses, and detect unintentional false responses too. When assessing the LPW dataset (Tonsen et al., 2016) I found the longest Euclidean distance between two consecutive frames to be shorter than 15px (Figure 4-6 left), while less than 30px between two sequential MS-SSIM-selected frames (Figure 4-6 right). Consequently, a worker fails to meet this quality measure when any Euclidean distance between annotated sequential frames of their work was found longer than 15px (all frames task) or 30px (MS-SSIM-selected frames task). For instance, in an MS-SSIM-selected frame job, an annotation is rejected if a worker localises pixel position (130, 110) and for the following sequential frame localises pixel position (95, 110)—a Euclidean distance of over 30px. This same method has been applied to localising calibration target tasks too.

Time spent: Considering that the nature of task assignments on crowdsourcing platforms is first-come-first-served, and the small tasks size as well as low payment, workers often tend to multitask (sign-up for, and undertake, multiple crowd tasks at the same time). Therefore, workers are given 10 minutes each to complete the localisation task (3 × average completion time), or their job is reassigned to the next available worker. This is so to avoid malicious workers or attacks, which aim to prevent tasks from being completed, or to prevent multi-tasking or lazy workers from locking the task indefinitely. Late or inactive workers whose session expired and their job has been reassigned to others lose their session and receive no reward.

Entry validation and refinement: Contrary to the conventional crowdsourcing guidelines where workers with low quality responses are either blocked (Silberman et al., 2010) without compensation, or allowed and their work is accepted regardless of quality, *CrowdEyes* empowers workers to validate and refine their quality rejected entries. Where workers fail to pass quality measures (other than time spent) they are given the option to review and refine their entries and submit again. Unless the worker gives up, the refinement task may be completed several times until entries meet the imposed quality measures. If a worker gives up, they will receive no compensation and all their responses will be dismissed and removed. Thus, workers are given 5 additional minutes every time they are instructed to refine their entries before the job is reassigned to the next available worker. As such, every time a worker fails to satisfy the quality measure, they are redirected to the refinement page where their annotations are overlaid on the task images. Images are presented in a grid (Figure 4-5 right) and the failing worker is instructed to review and improve their annotations so they are as close to target centre as possible. Unlike the initial annotation task that presents images in a slider for faster annotation with minimal cursor moves, the refinement task presents all images in a grid. The grid enables workers to quickly browse through all images including accepted ones to identify and correct or improve less accurate annotations.

4.8.2.1 Recruitment and Payment

For every eye-tracking job, *CrowdEyes* automatically advertises the job on CrowdFlower. The recruitment advert page includes the job description with a hyperlink to the *CrowdEyes* tasks website; a text input field to enter the payment redeem code (rewarded to successful workers); and a hidden client-side script that validates entered redeem codes with the *CrowdEyes* server. On every successful job completion on *CrowdEyes* tasks website, *CrowdEyes* issues a payment redeem code assigned to that particular worker. Workers are then instructed to enter the redeem code into the dedicated text input field on CrowdFlower job page to receive the promised monetary compensation. The payment code is unique per worker per task and cannot be redeemed twice, preventing workers from over redeeming or sharing the code. Once a worker enters the code on the CrowdFlower job page, our client-side script confirms with the *CrowdEyes* server the code’s authenticity and validity.

4.8.3 Step 3: Gaze mapping

While recruited workers are completing the *CrowdEyes* target localisation tasks, *CrowdEyes Player* plugin provides a live job completion status via its graphical user interface. *Player* also has a feature to retrieve the crowd responses, and to identify outliers in the annotated data for which it recompenses by the computed mean of the localised centre in preceding and following frames. Following that, *Player* uses the standard *Pupil* methods (Kassner et al., 2014) to compute the common eye tracking analysis metrics including gaze positions, saccades as well fixations. At this stage, the user can watch their recordings on which gaze, saccades and fixations are overlaid. But detected fixations are not labelled yet, users will need to use the optional labelling feature provided by the *CrowdEyes Player* plugin to label everything that they fixated upon.

4.8.4 Step 4: Labelling fixations (Optional)

Eye trackers capture users’ points of interest in which the person’s eyes are relatively stable (fixations) for a minimum duration of 100-200ms (Jacob, R.J.K., and Karn, 2003). The technology yields a large number of fixations, and each fixation corresponds to number of related world scene frames (e.g., a 200ms long fixation captured by 30Hz camera is composed of 6 frames). To label detected fixations, first *Player* picks the middle frame out of each corresponding world scene frame set, eliminating repetitive frames (Munn et al., 2008) and overlaying a crosshair over the detected fixation points. *Player* then bundles selected frames and pushes them to *CrowdEyes* server to organise the crowd fixations labelling job. Each crowd job comprises a maximum of 10 tasks, and each task consists of a fixation image and a set of questions. All questions are related to the marked point of interest (e.g., object being looked at) and surrounding area for workers to answer (e.g., describe what is being looked at). Upon the completion of the crowd labelling job, *CrowdEyes* server aggregates crowd responses and pushes them back to *Player*, which overlays them on the relevant eye tracking recording to appear near the fixations' crosshair. And for any other further independent work, *Player* extracts the aggregated results and their corresponding timestamps into a spreadsheet document.

4.9 Stage 1: Evaluation of pupil localisation

I evaluated *CrowdEyes* in two stages. In the first stage, I evaluated the accuracy and costs to crowdsource the pupil localisation task. As such, I leveraged the large-scale (manually labelled) open dataset LPW (Tonsen et al., 2016) to ensure crowd workers could sufficiently localise the centre of the eye pupil in comparison to those of existing measures (Tonsen et al., 2016). This stage helped us in tuning our crowd localisation tasks to increase accuracy and reduce costs, and it enabled us to conduct the second stage. In the second stage, I have applied the tuned *CrowdEyes* solution to a real-world scenario, evaluating its entire pipeline from calibration, over data captures, to analysis.

CrowdEyes proposes crowdsourcing methods in order to overcome the shortcomings in existing pupil localisations algorithms and improve the efficiency of mobile eye tracking technology. The accuracy and robustness of the proposed *CrowdEyes* methods to localise target centre (i.e., eye pupil) are evaluated against the overall costs of crowsdourcing. The assessment was carried out over 130,856 pupil frames captured in unrestricted environments (22 minutes of footage captured at 95fps) of 66 diverse recordings of 22 participants (5 different nationalities to count for race/phenotypes and eye shape)—a dataset manually labelled by a researcher and used to assess five state-of-the-art pupil localisation algorithms (Tonsen et al., 2016). The dataset was chosen for its challenging and distinct conditions, such as, users wearing spectacles, and eye make-up; for recordings captured outdoor and indoor with mixed source of light. The crowdsourcing of 66 recordings were evaluated in two ways: all frames were crowdsourced i) without MSSSIM frame selection in the initial run (R1), and ii) with MSSSIM frame selection in the second run (R2), and compared to (Tonsen et al., 2016)'s reported measures of the selected state-of-the-art algorithms.

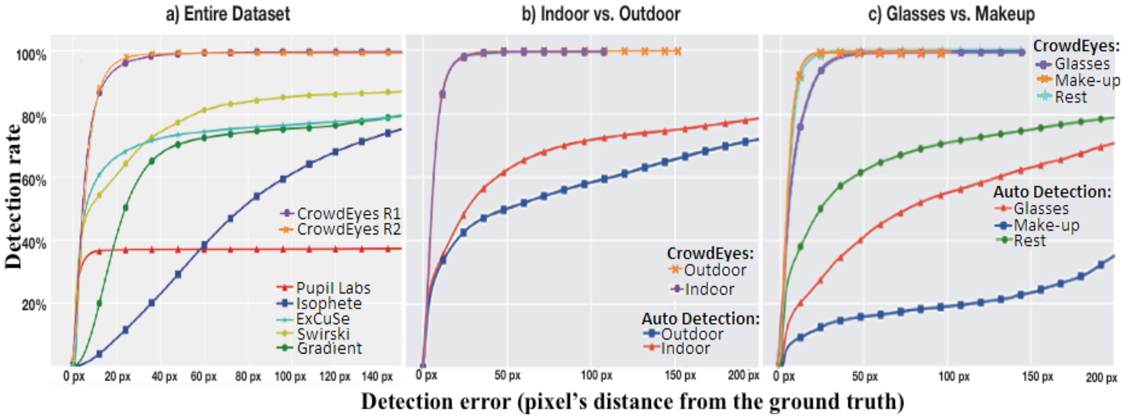


Figure 4-7 Cumulative distribution of the mean error: a) comparison of *CrowdEyes* method (run 1 (R1) and run 2 (R2) without and with MSSSIM frame selection respectively) and 5 common algorithms (adapted from (Tonsen et al., 2016)); b) comparison of *CrowdEyes* method and automatic detection using frames collected indoors and outdoors; c) comparison of *CrowdEyes* method and automatic detection using frames representing glasses and eye make up

4.9.1 Accuracy and robustness

The outcome of both initial run (R1) and the second run (R2) on the LPW dataset clearly indicates that *CrowdEyes* outperforms all five algorithms for cumulative distribution (CD) of the mean error in pixels. Unlike the best (Tonsen et al., 2016) two evaluated algorithms *Swirski* and *ExCuSe*, *CrowdEyes* accomplished 100% pupil localisation rate on all frames in both runs under all conditions, with accuracy (pixel's distance from the ground truth) under 10px for 80% and under 20px for 97% of all frames, see Figure 4-7a. Whereas, *Swirski* and *ExCuSe*, fall behind with under 80% pupil detection rate on all frames (in all conditions), and with low accuracy over 20px distance error for more than 35% and over 100px for more than 20% of all frames. In addition, both of the CrowdEyes runs demonstrated the capability of the crowd to incomparably localise pupil centre under challenging conditions. Notwithstanding the indoor and outdoor mixed lighting (Figure 4-7b) and eye make-up (Figure 4-7c), CrowdEyes methods resulted in CD mean pupil localisation error under 25px for 99% of the data. However, Figure 4-7c also suggests crowd workers yield less accurate pupil centre localisation with data of participants wearing spectacles, which resulted in CD mean localisation error under 25px for 90% of the data. It is almost certain that this lower accuracy is a result of the spectacle's frame occluding (partially) the pupil captured in the eye camera field of view. In contrast, all five

	R1	R2
Frames	130,856	27,230
Micro-tasks	1309	273
Workers (no tasks completed)	1375 (39)	305 (17)
Refined (success rate)	120 (77.5%)	59 (74.6%)
Cost	$523	$109
Time Mean (STD)	175s (56s)	179s (63)

Table 4-1 R1: Crowdsourcing all video frames; R2 with frame selection using MSSSIM, for 66 recordings with 95Hz cameras (about 23 minutes)

algorithms yield low accuracy when detecting pupil under the same challenging conditions. They produced a CD mean detection error over 50px for 40% of indoors and 50% of outdoors data; and over 100px of more than 80% of the recordings for participants wearing eye make-up—let alone over 60% of eye make-up frames remained undetected. On the other hand, our results also suggest alienating redundant frames using MSSSIM index method in CrowdEyes second run reduces the overall costs with no notable compromise on accuracy.

4.9.2 Results and analysis: Time and costs

Throughout the first run R1 of the *CrowdEyes* experiment, 130,856 frames were labelled in 1375 assigned micro-tasks. Among the micro-tasks assignments 39 workers failed to complete any task and a further 27 micro-tasks workers either gave up on refining their responses or were timed out. Amongst the successfully completed 1309 micro-tasks assignments, 93 workers efficaciously refined their responses for micro-tasks after one or more refinement trials. On the other hand, the application in R2 of MSSSIM on the captured eye tracking frames resulted in approximately 80% less frames to crowdsource, hence an 80% reduction in the costs of crowdwork. Throughout R2 27,230 frames were labelled in 305 assigned micro-tasks, of which 17 assignments workers failed to complete any tasks, and 15 others either gave up on refining their entries or were timed out. Whereas 44 crowd-task assignments workers efficaciously refined their responses for micro-tasks after one or more refinement trials. Multiple micro-tasks were running in parallel and took in total 57 minutes to complete R1 compared to 27 minutes for R2, with the mean time taken to complete one task in both runs being just under 3 minutes.

Considering fair crowdsourcing pay, workers who completed their tasks received a pay rate equivalent to the UK minimum wage (at the time of the experiment). Each successful worker received US $0.4 per task (130 frames including gold standards). Whereas as the total cost came to US $523 to crowdsource all R1 frames (US $22.7 per 95Hz eye tracking minute), it only costed US $109 to crowdsource all MSSSIM selected R2 frames (US $4.7 in average per 95Hz eye tracking minute), see Table 4-1. Above and beyond, the overall costs may be reduced using lower fps eye trackers. Crowdsource recordings of a 30Hz eye tracker results in costs of approximately US $7.2 per eye tracking minute (all frames), or US $1.4 per eye tracking minute (MSSSIM selected frames). As such, *CrowdEyes* doesn't only ensure accurate and robust data capture, it also costs as little as US $87 per hour of data using 30Hz sampling rate cameras.

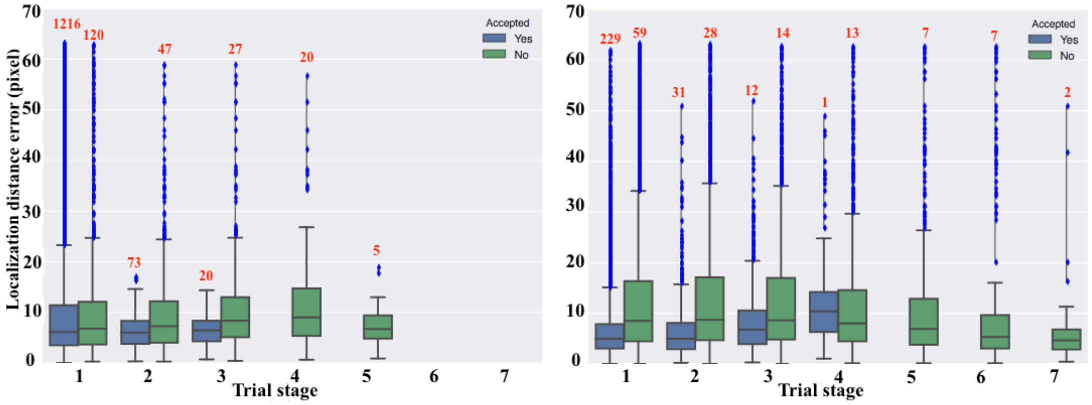


Figure 4-8 Refinement trials vs. workers' localisations distance error in pixels for accepted (blue box) and rejected (green box) submissions (total in red) from first run R1 (left) and second run R2 (right)

4.9.3 Results and analysis: Refinement

Figure 4-8 illustrates the distance error distribution for the pupil centre localisation tasks. Blue boxes represent accepted submissions, whereas the green boxes represent rejected submissions during the initial pupil localisation trial (trial 1) as well as the refining trials that follows (trail 2 and above)—the total submission number per trial is highlighted in red. Responses that don't meet the minimum localisation quality standards are rejected and responsible workers are requested to refine their work—unlike the traditional crowdsourcing where workers are straight removed and unpaid if their responses were rejected. During the trial 1 of *CrowdEyes* R1 (Figure 4-8 left trial 1) the distribution error for accepted workers' responses was found to be equivalent to the rejected ones, suggesting workers may fail one or more quality standards notwithstanding their generally good responses. On the other hand, 73 workers in trial 2 and another 20 workers in trial 3 successfully refined their responses, confirming their genuineness and desire not to waste their effort made in previous trials. Those workers achieved an even better distance error distribution than that accepted in the initial trial, with outliers almost eliminated in the accepted refinement trials. But Figure 4-8 left also indicates workers either timed out or gave up on refining, or continued to fail one or more quality standards after trial 3. Similar result is also notable in *CrowdEyes* R2, see Figure 4-8 right. This suggests that trusting workers to validate and refine their responses not only increases the quality of their work; it also offers them fair compensation for their effort and time, and keeps crowdsourcing costs to a minimum. Subsequently, 137 (out of 179) workers (across R1 and R2) successfully completed their refinement tasks and guaranteed their compensation. Finally, it appears that the refinement quality method could also serve as a motivation factor. It encourages workers to visit our job again. Approximately 33% and 51% of workers who were accepted after the refinement trials in R1 and R2 respectively returned to complete more tasks.

4.10 Stage 2: Applying *CrowdEyes* within a real-life scenario

Following on the successful and reassuring results from stage 1 which proved the eye tracking accuracy and robustness of CrowdEyes, here I demonstrate its utility and extensibility to accommodate other types of crowd tasks, such as fixation labelling. Fixation labelling is a common interest in the analysis of eye tracking, which provides eye tracking (*CrowdEyes*) users with summaries of where they gazed.

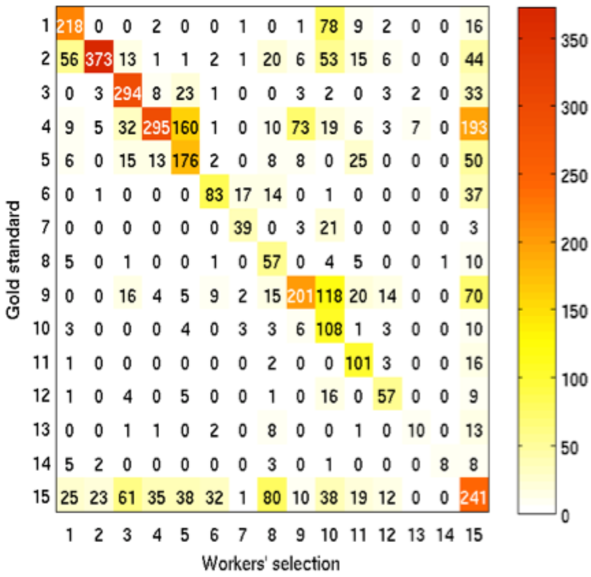
4.10.1 Methods & procedure

In this case study, I captured where and what participants paid their attention to with their gaze while purchasing lunch. I recruited 8 university employees and students (6 male and 2 females; 4 with spectacles, and 1 with eye make-up) to wear *CrowdEyes* when purchasing food in their university's cafeteria. I trialled the initial calibration procedure with each participant and asked them to use their thumbnail instead of a machine-known calibration marker (e.g., marked paper). Since this solution relies on crowd workers to localise the calibration target, this method eliminates the need for using special calibration markers and avoids the inaccurate auto-detection methods in light-filled and object crowded environment. All 8 participants were instructed to use their thumbnail to simulate the traditional 9-point calibration method—where a user visually tracks a machine-known marker moving towards predefined points relative to the world camera's field of view. Accordingly, participants were given two calibration choices, either to calibrate with head stationary while moving hand (thumbs up) or vice versa (see Figure 4-3). After the trials were completed, all participants opted to record with stationary hand (thumbs up) while moving their head (fixating their gaze on their stationary thumbnail, while moving their head to cover upper, middle, and lower rows of their field of view and pausing three times on each row). For the actual recording session, three participants were instructed to calibrate their eye tracker outdoors (in sunlight-filled environments) while the remaining participants calibrated indoor in mixed light-filled and objects crowded environment. Henceforth, participants were instructed to first power up the *CrowdEyes* device and carry out the calibration process before entering the cafeteria, choose and pay for their lunch, and turn off the recording on completion of their purchase. In total, participants have recorded 28:56 minutes, averaging 03:36 minutes (the shortest recording was 01:50 and the longest was 05:37 minutes). The overall calibration time was 08:25 minutes, averaging 01:03 minutes (the shortest calibration took 42 seconds, while the longest took 01:23 minutes).

4.10.2 Results and analysis: Time and costs

The overall number of captured eye frames was 52,093, of which 10,104 frames were selected by applying the MSSSIM index method (a reduction by roughly 81%) to crowdsource pupil localisation. As a result, 102 workers successfully completed 102 pupil localisation tasks, of whom 9 workers had to refine their responses before being accepted. Furthermore, I also crowdsourced the calibration world frames to localise the calibration target (finger thumb as mentioned earlier). However, the application of MSSSIM to eliminate redundant calibration world frames was less effective than applying it to eye frames, and resulted in a reduction by 42% (8,723 frames crowdsourced in 88 crowd tasks). A likely explanation is that these lower reduction rate in world frames compared to eye frames is a result of higher noise factors present in the world frames' field of view (e.g., mixed light and object crowded environment). Nevertheless, the total cost for crowdsourcing the selected eye frames (pupil localisation) was US $41 (averaging US $1.4 per minute), plus US $35 for crowdsourcing the calibration world frames (thumbnail localisation), averaging US $4.4 per calibration session.

All recordings yield 1406 fixations, excluding those captured during calibration. As a result, 141 micro-tasks were created for labelling where participants fixated their gaze on. Each micro-task was judged by 3 workers and comprised of a maximum 10 world scene frames plus two pre-labelled gold injected frames. *CrowdEyes* server recruited in total 456 workers from CrowdFlower to carry out the fixations labelling micro-tasks on *CrowdEyes* website. Among the workers 12 timed out or gave up on refining their answers, 21 quit their tasks too early, and 49 others had to successfully refine their answers to meet with the quality standard. For each frame, workers were given a predefined list of categories and instructed to choose the category that best match the object being fixated upon (identified by a crosshair). The categories were given after going through every item and object in the cafeteria that participants may fixate their gaze upon. They were: 'Man', 'Woman', 'Group of people', 'Drink', 'Sandwich', 'Chocolate bar, crisps, chips, biscuits', 'Cash register', 'Display Screen', 'Table or chair', 'Sign, post or advertisement', 'Wall', 'Floor', 'Gate or Door', 'Fruit', 'Other'. The mean time



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	218	0	0	2	0	0	1	0	1	78	9	2	0	0	16
2	56	373	13	1	1	2	1	20	6	53	15	6	0	0	44
3	0	3	294	8	23	1	0	0	3	2	0	3	2	0	33
4	9	5	32	295	160	1	0	10	73	19	6	3	7	0	193
5	6	0	15	13	176	2	0	8	8	0	25	0	0	0	50
6	0	1	0	0	0	83	17	14	0	1	0	0	0	0	37
7	0	0	0	0	0	0	39	0	3	21	0	0	0	0	3
8	5	0	1	0	0	1	0	57	0	4	5	0	0	1	10
9	0	0	16	4	5	9	2	15	201	118	20	14	0	0	70
10	3	0	0	0	4	0	3	3	6	108	1	3	0	0	10
11	1	0	0	0	0	0	0	2	0	0	101	3	0	0	16
12	1	0	4	0	5	0	0	1	0	16	0	57	0	0	9
13	0	0	1	1	0	2	0	8	0	0	1	0	10	0	13
14	5	2	0	0	0	0	0	3	0	1	0	0	0	8	8
15	25	23	61	35	38	32	1	80	10	38	19	12	0	0	241

Figure 4-9 Confusion matrix illustrating the agreement between the categories selected by the crowd workers and the correct (gold standard) categories.

taken of all micro-tasks was ~01:54 minutes (STD=01:01 minutes), and successful workers were paid US $0.3 in return per micro-task (US $127 per full job).

4.10.3 Results and analysis: Accuracy and robustness

Previously in this chapter I demonstrated the robustness and accuracy of centre pupil localisation by crowd workers, next I evaluate their responses in relation to fixation labelling tasks. I used Fleiss' kappa to measure the inter-rater reliability (IRR) between crowd workers, since I used nominal categories for labelling fixations to be judged by at least three workers. Although the micro-tasks were designed to employ conventional quality control only (injecting gold standard images), the measured IRR scored a Fleiss' kappa of 0.6671, suggested substantial levels of agreement between crowd workers. I have manually labelled detected fixations in the captured world frames before crowdsourcing, and selected 10% of them to inject (as the gold standard) into the fixations labelling micro-tasks for quality measures. Consequently, I compared the crowd responses with our gold standard labels, and illustrated the agreement between the categories workers selected and the gold standard (see the confusion matrix in Figure 4-9). The confusion matrix highlights in the diagonal the cases where fixations were correctly labelled by the crowd. I also computed the unweighted Cohen's kappa coefficient from this confusion matrix and found it to be moderate (Cohen's Kappa 0.49). This moderate level of agreement suggests that some crowd workers failed to distinguish between

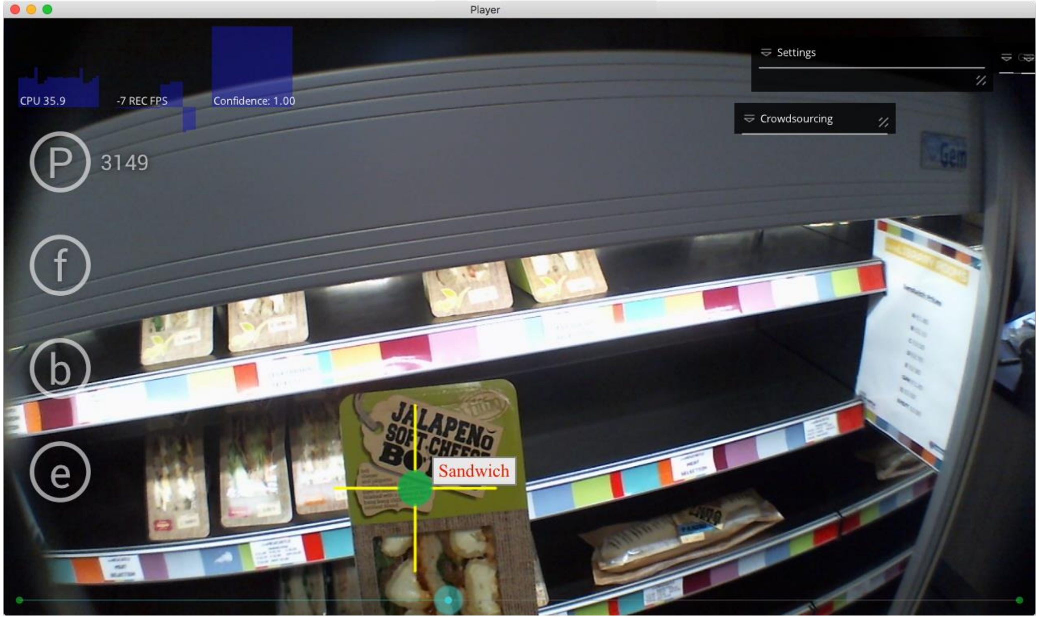


Figure 4-10 CrowdEyes Player plugin integrated into the open-source Pupil Player software to show the labelled fixations

some of the categories, which results in some high values outside of the diagonal in Figure 4-9. A likely explanation is that such difficulties in distinguishing between categories are the result of giving workers limited training; the quality of captured images; the distance of the target object from the camera; or the object being unknown to workers. For instance, workers seem to fall for category 15 ("Other") whenever they fail to recognise the object being fixated on. Taking this category out would result in a substantial level of agreement (Cohen's Kappa of 0.61).

Finally, the processed data is presented using *Pupil Player* software and *CrowdEyes Player* plugin. Figure 4-10 presents a selected frame from the lunch purchase process with the crowd-labelled fixation (Sandwich).

4.10.4 Discussion

CrowdEyes demonstrates that crowdsourcing (human-computation) can be employed to improve data processing and analysis for wearable mobile eye trackers. Our studies deliver robust comparative findings, showing high pupil tracking accuracy and suggest that fixation labelling can also be automated to deliver reliable and telling outcomes. While employing workers for these tasks does come at a cost, projections including broader worker audiences and a tolerable reduction in key frames that are sent out for manual detection suggest that eye tracking data analysis with *CrowdEyes* can be efficient and scale to a low per minute cost, while delivering a level of quality that is unparalleled by purely computational approaches. As evidenced by our findings, giving workers further opportunity to validate and refine their entries yielded better levels of performance, higher rates of task completion, more compensation awarded to workers and, importantly, more workers revisiting the job.

Whereas the self-reporting gaze recall methods (Rudoy et al., 2012; S. Cheng et al., 2015) require no other special hardware than a display screen, *CrowdEyes* requires a head-mounted video-based eye tracker. In turn, *CrowdEyes* expands the *Pupil* platform, adding a human-computation plugin, and using a pocket PC, two off-the-shelf webcams and a 3D printed head-mounted frame—low-cost and hackable. However, unlike (Rudoy et al., 2012) and (S. Cheng et al., 2015), which must be performed on screen while workers complete number of memory-dependent tasks to recall gaze positions, *CrowdEyes* enables robust as well as mobile eye tracking under real-world conditions, with few constraints regarding locations, lighting conditions, or eyewear. This means that eye tracking can be used, for instance, to efficiently evaluate outdoor activities (e.g., visual attention for cyclists when cycling on or off road) and technologies (e.g., the impact of using mobile phones on situational awareness during a walk). Moreover, it can also be used as a lifelogging tool that video captures and labels the surrounding area as well as the wearer gaze and fixations, adding more depth to lifelogging captured data. As a result, *CrowdEyes* could eventually be used to drive recommender systems based on what a wearer looked at.

4.10.5 Limitations and future work

While the evaluations presented here were designed to include realistic use cases, the approach does require ecological validation, which is especially relevant to gauging the value of future applications of the fixation labelling process. The durations of the eye tracking recordings employed in these studies were substantial, but the question of how easily the approach scales to longer duration recordings does require further evaluation, as do considerations related to potential near real-time analysis through further parallelization. Furthermore, the process for pupil localisations partially relies on gold-standard data. It can be argued that it will likely not be necessary to employ novel gold data samples for the analysis of future recordings, since existing gold data frames could simply be reused. The gold standard data itself, however, also poses a limitation on the study. Given that some of Tonsen’s dataset was human annotated, there was possibly a bias towards human annotation methods. In addition, images within Tonsen’s dataset were captured with a 95Hz camera, whereas *CrowdEyes* only employed a

30Hz camera. Since the workers' localisation accuracy is independent of camera frame rate, unlike the costs, I evaluated the localisation accuracy and costs of our approach with Tonsen's dataset (95Hz) in Stage 1 compared to costs only in Stage 2 (30Hz). Consequently, I reported the costs difference in running *CrowdEyes* with 30Hz cameras (~US $85 for localisations) compared to 95Hz cameras (~US $280). However, to reduce the costs, speed up the process and ensure higher labelling agreement, in our future work I will look at training crowd workers and create a pool of trained workers available on demand. Lastly, the promising outlook of improving automated methods through crowdsourced high-quality results, e.g., by training modern deep learning networks, certainly warrants further study.

4.11 Conclusions

In this chapter, I have presented the motivation, design, and evaluation of *CrowdEyes*, a hybrid eye tracking system that employs crowdsourcing for pupil and calibration target localisations, combined with automatic data processing (e.g., gaze mapping) provided by standard functionalities of the *Pupil* framework. *CrowdEyes* leverages the crowd to provide a robust and reliable mobile eye-tracker that functions under real-world conditions, a feat that has so far remained elusive. The high accuracy of *CrowdEyes* in localising pupil center highlights the potential it holds for enabling a broad variety of applications beyond those that are available when using regular contemporary eye tracking only. Moreover, in this chapter I have presented a novel crowd quality measure, which relies on workers to self-validate and refine their entries. This method yields more accurate entries, encourages workers to perform better, and prevents honest workers from being rejected or unpaid. The results of this work suggest our approach is robust, accurate, and cost effective.

5 Discussion and Conclusion

This thesis addressed challenges of obtaining high quality contributions from crowdsourcing approaches. To this extent, I presented various methods for improving the quality of crowdsourcing the analysis and translation of speech and eye tracking data collected in the wild. This research aimed to address the design of crowdsourcing speech and visual real-world mobile systems that maximise the quality of crowd responses, while minimising cost and crowd effort. To achieve this, I have designed and developed a set of crowdsourcing tasks for two novel mobile solutions to be evaluated in two case studies. The first case study aimed to crowdsource the collection and analysis of disordered speech of people with Parkinson's to support the self-management and monitor of their speech condition. The second case study focused on eye tracking and the technical challenges it faces, which drastically limits its applications to controlled scenarios (e.g., in lab, users cannot wear eye makeup). Both case studies explored various design aspects and factors to achieve quality in real-world mobile crowdsourcing systems and answer the three research questions.

RQ1: What is the implication of self-verification as a quality control method on improving accuracy with no additional costs?

Imposed standard quality measures often result in expelling, not only unsatisfactory, but sometimes quality honest workers from the job and consequent tasks (e.g., due to poor tasks design). The outcome of being rejected does not end by expelling a worker from performing a task, a worker's reputation on crowdsourcing platforms is dependent on their rejection ratio. More rejections lead to lower reputation and thus limited access to the tasks pool. For requesters, the impact includes additional costs, inaccurate responses, and extra post-processing effort. Thus, I proposed a self-verification method that offers workers multiple chances to complete tasks they have started and receive their rewords without compromising on quality or costs.

I evaluated the self-verification quality control method in a crowdsourcing solution that I designed and developed to power a DIY mobile video eye-tracking system. Video eye-tracking systems require high accurate localisation of the centre of eye pupil and other targets (i.e., calibration markers), a task that is very challenging for current state-of-the-art algorithms, especially when recording in real-world scenarios. As such, I introduced highly strict crowdsourcing quality measures to detect any worker's responses (e.g., pupil centre localisation) that are not highly accurate. Until the desired accuracy is achieved, or the worker gives up, failing workers receive unlimited chances to verify and improve their low-accuracy responses. This way, workers protect their reputation in the crowdsourcing platform and contribute to improving their own responses. The images in the self-verification task are presented in a grid view layout to enable rapid visual search for responses (provided in previous trials) that appear less accurate. When any less accurate response is recognised the worker uses the mouse to correct and override their previous response. When a worker hover over an image the mouse pointer changes to a crosshair, so it may aid workers localise the centre faster and more accurately. Chapter 4 reported how the self-verification task not only improves the accuracy of crowd responses, but also retains workers and encourages re-preparticipation in further tasks. Together with other quality measures, the self-verification method enabled the crowdsourcing eye-tracking solution to obtain quality responses from one worker per task (unlike the traditional 3 or more responses per task) and kept costs to minimum.

RQ2: How to design crowdsourcing tasks to achieve expert-comparable input when working with speech and visual data?

While crowdsourcing may seem like a low-cost approach, maintaining good quality crowd responses is very challenging, let alone expert-comparable quality responses. The answer to this research question was explored in chapter 3 and 4.

Chapter 3 explored designing crowdsourcing speech assessment tasks to support personal care around speech. Not only high-quality crowd responses are necessary here, but the quality has to match that of an expert to be meaningful to those being assessed. In two iterations, this chapter presented the feasibility of the crowd to measure perceptual speech including pitch, rate, and volume, and judge speech intelligibility. It explored various crowd tasks designs that closely simulate clinical speech assessments and strived to obtain quality judgements comparable to that of an expert. Just like in a clinical set up, the crowd task focused on connected speech and single words tasks to assess a recorded impaired speech. Only workers based in the UK were permitted to perform the tasks to overcome language and accents barriers.

For connected speech assessment, a worker has to listen to a short recording of impaired speech then transcribe it and judge its ease of listening. The transcription task aimed to highlight misheard or unintelligible words. Whereas the ratings offer a measurable method to inform speakers about their progress. Using five-point Likert scale each worker rates ease of listening of the impaired speech and the influence of accent on their judgement. The five-point Likert scale was found to have a strong correlation to intelligibility scores, especially when listeners (the crowd workers in this study) are novice and unfamiliar with speech disorders (Landa et al., 2014; Miller et al., 2007). The worker has to also listen to two other audio recordings of the same speaker and rate their perception of volume, pitch, and pace on a continuous scale of 0–100. The continuous scale allows for large range of variability between listeners without impacting the sensitivity of ratings that may have been observed in a discrete scale. To aid workers giving quality ratings, one of the two audio recordings is a baseline (previously rated) to allow for an easy estimation of the magnitude of variance in the connected speech tasks (Gary & S., 2002; Miller, 2013).

On the other hand, the single word task consists of an audio recording of a person saying one word and the worker has to select the word he/she actually heard from 10 options (e.g. cup, cop, coop), without guessing. And using five-point Likert scale, the worker rates the intelligibility of what they heard and the influence of the speaker’s accent on their judgement. Single word tasks provide a measurement of intelligibility in isolation by removing context and flow that may add to the worker’s ability to hear and relate the words together in an intelligible message. The quality measures for this study were kept intentionally simple, five crowd judgements per task and gold standard questions. This enabled me to study the impact of designing crowd tasks that simulate clinical practices on achieving quality judgements comparable to that of an expert.

As such chapter 3 reported on the level of agreement among the crowd and between the crowd’s aggregated-judgements and the experts’. In the first iteration, despite the weak correlation was observed in measuring speech volume, high correlation was observed between the crowd and experts in measuring speech pitch and rate, and in recognising single words. Substantial agreement in the ease of listening tasks was also observed in the first iteration. These findings suggest that anonymously recruited crowd workers are capable to offer experts' like ratings when tasks are considerately designed. However, it became essential for the second iteration to address the observed lower correlation in ease of listening and volume rating tasks. And unlike the first iteration in which data was collected by experts in controlled environment, the second iteration focused on data collected by smart phones in the wild (using Speeching app). As such, the second iteration explored further the capability of the crowd to produce experts' comparable judgements and provide meaningful feedback to the Speeching app users. This iteration, improved on the tasks' instructions to measure the perception of speech volume, which led to moderate, almost high positive correlation between the crowd workers' ratings and the experts'. Furthermore, a comparative element (made of the user's own speech and previously judged by the crowd) was added to the tasks to help the crowd provide a comparable measure of pitch, rate, volume, and ease of listening. The implemented tasks modifications, in return, led to further correlation improvement between crowd's judgements and the experts'.

Whereas Chapter 4 explored designing crowdsourcing tasks to overcome eye tracking's quality and robustness issues that limit its applications in real-world scenarios. The Chapter identified number of challenges, such as mixed lightings and user's eye wear (e.g., spectacles or makeup) and proposed a crowdsourcing approach as an alternative to existing automatic target detection methods. Thus, I examined the influence of task's layout on crowd efficiency and effectiveness to achieve expert's comparable accuracy and higher performance than the current state-of-the-art algorithms. Consequently, two tasks were designed, and each worker was given 130 images (frames collected by the eye tracker camera) to accurately localise and verify the centre of a target displayed on each image.

Workers are first informed about how many images to be completed in a task, the average time it takes to complete all of them (3 minutes) and the reward. This enabled workers to decide whether the reward is worth the effort. Workers were instructed to submit their responses within a pre-determined duration (10 minutes) or their task is reassigned to another worker. This is to prevent any task from being indefinitely locked down to any worker or automated malicious tools.

The first task presents all images in a slider that flips on worker's mouse click. This layout was selected to reduce page scrolling and mouse movement across the page, reducing time and effort it takes a worker to move from one image to another. Besides, all crowdsourced images per task were sequential (based on timestamp), which lessen mouse movement even further (considering 30fps and the short eye movement within one or two seconds). To aid workers effectively localise the centre of an eye pupil, the mouse cursor was replaced with a crosshair.

The quality measures implemented in this study were a mix of traditional and advanced. I used 30% gold standard injected data (images with known target centre) to make it harder for workers to identify whether an image is a gold standard. I also designed the task to crowdsource sequentially captured images where the target makes a very small shift in its location (few pixels) between these images. Consequently, a wide Euclidean distance between two sequential responses is identified as low a quality response. So, if a submission does not meet the imposed quality standards, the worker is transferred to the second task to self-verify their responses.

The self-verification task layout was designed differently, but used the same quality measures. Instead of a slider where a worker has to click on each image to see the next one, this task presents all images in a grid view with the worker's responses already laid out. This layout enables workers to quickly browse all images and look for those that appear less accurate. Like in the first task, the mouse cursor turns onto crosshair when moving over the images to aid workers verify their response and provide a more accurate localisation of the centre of the target. As a result of the carefully designed task layout and the chosen strict quality measures, each task required the responses of a single worker (unlike the traditional 3 or more workers), so no post-crowdsourcing data aggregation and no additional costs.

The chapter established that crowd workers are capable in producing high quality annotations equal to that of an expert annotator. Besides, the overall outcome shows that the proposed crowdsourcing solution outperformed all state-of-the-art algorithms under controlled as well as real-world conditions.

RQ3: How to develop low-cost crowd-powered solutions that directly benefit end-users?

In chapter 3, Speeching app was deployed in the wild to enable participants with Parkinson's to self-manage and- monitor their speech. The app offered participants the possibility to practice their speech from anywhere anytime, and offered them the choice to complete an assessment on their smart phones. Assessments were submitted to the crowd for ratings and their responses were then aggregated and presented in a statistical form on how their speech was perceived by naïve anonymous raters unrelated to them (the crowd workers). Each assessment took less than an hour to complete and costed nearly $2.10 USD. In a post-deployment interview with the participants, they emphasized how the received feedback helped them better understand their speech condition and their progress. Some participants also suggested the solution motivated them to practice more and improve their speech.

Similarly, in chapter 4, the crowd-powered mobile eye-tracking system was deployed in the wild to establish its usability and benefit to end users. I recruited 8 participants from Newcastle University to use the eye tracker during lunch purchase activity. After crowdsourcing the localisation of eye pupil and calibration target, for robust eye tracking results (e.g. gaze, fixations, saccades), the system identified fixations (what a participant looked at and thought about) and crowdsourced the labelling of these fixations. The crowd identified labels were then laid out on the eye tracking recordings for participants to watch anytime and increase their knowledge about their lives and the decisions they make throughout the day. The results show that the approach is functional and can be used in many domains, particularly in research domains (e.g., marketing or product evaluation studies) given its robustness and low cost.

5.1 Research limitations and future directions

Both case studies presented in chapter 3 and 4 have potential limitations. The Speeching case study was deployed for a week with six participants with Parkinson’s whose speech disorder conditions vary from low to sever. A larger and longer scale study is needed to thoroughly evaluate the proposed Speeching solution, exploring its sustainability, and investigating long term impact of the provision of feedback on its users. One may investigate alternative motives than monetary and the potential to utilise people’s social capital to rate the speech assessment and support their friends or relatives with speech disorder. On the other hand, one of the participants failed to use the mobile solution due to broadband connectivity issues. Broadband connectivity in mobile services present a challenge that may be addressed by offering offline (no connectivity) services that sync with the Speeching server once back online, for example. Although crowdsourced speech assessments were, in average rated in less than an hour for a nearly $2.10 USD, crowd tasks may be refactored to lessen completion time and lower the cost further. Only UK based crowd workers could rate the uploaded assessments, which restricted the access to the wider crowdsourcing market and contributed to the longer average completion time. Since you don’t need to understand English to judge perceptual speech (pitch and volume at least) another design iteration is needed to explore the potential for utilising all available crowd workers (not only UK based). Further design iterations research is beneficial to investigate, for example the golden ratio between number of raters and the length of speech recordings per assessment. Finally, practices and assessments were hard coded in the Speeching app, so it may contain irrelevant content to some users. As such, the app should be designed in a way that enables users or therapists to set up more relevant content to their conditions and goals they wish to achieve.

Although CrowdEyes solution (from the second case study) presents to its users descriptive labels of everything they gazed on while wearing the eye tracker, yet it offers no further information or interactions. This limitation is intentional as to keep the scale of this study under control and more focused on the quality and accuracy of the crowd. As such, further studies needed to explore the design of crowdsourcing tasks for supporting the provision of feedback based on what we gaze on and the impact on its users. Also, the current capability of mobile computing (e.g., mobile devices, pocket PC) is still inadequate to support long constant eye tracking sessions. And when that is possible, the amount of video frames from constantly recording eye tracking data present another challenge for crowdsourcing. As such, another study is needed to investigate the role of crowdsourcing approaches and task design to support automated deep learning eye tracking algorithms that may offer highly accurate and cheap as well as interactive eye trackers.

Finally, as this research did not investigate the privacy and security issues related to crowdsourcing personal data, since that is out of the scope of this study, further study is required. The study may explore the impact of making speech data incomprehensible before obtaining rating of the crowd for perceptual speech elements (pitch, volume, and rate).

6 REFERENCES

Abeliuk, A. & Masuda, N. (2014) Iterated crowdsourcing dilemma game. *Scientific reports*. 4 (February), 8–15.

von Ahn, L. (2013) Duolingo: Learn a Language for Free while Helping to Translate the Web. *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*. 1.

Von Ahn, L. (2005) Human computation. Synthesis Lectures on Artificial Intelligence and Machine Learning. 131–120.

Alallah, F., Neshati, A., Sheibani, N., Sakamoto, Y., Bunt, A., Irani, P. & Hasan, K. (2018) 'Crowdsourcing vs Laboratory-Style Social Acceptability Studies?: Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. [Online]. 2018 New York, NY, USA: ACM. pp. 310:1--310:7.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E. & Dustdar, S. (2013) Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*. 17 (2), 76–81.

Allen, Z. (2015) GAZE : Using Mobile Devices to Promote Discovery and Data Collection. CHI EA '15 Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. 97–102.

Alonso, O. (2013) Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*. 16 (2), 101–120.

Alonso, O. & Lease, M. (2011) Crowdsourcing 101 : Putting the WSDM of Crowds to Work for You. *Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11*.

Alonso, O. & Mizzaro, S. (2012) Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*. 48 (6), 1053–1066.

André, P., Kraut, R.E. & Kittur, A. (2014) Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. *Conference on Human Factors in Computing Systems - Proceedings*. 139–148.

Antin, J. & Shaw, A. (2012) 'Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. [Online]. 2012 New York, NY, USA: Association for Computing Machinery. pp. 2925–2934.

Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K.M., Dorsey, E.R. & Little, M.A. (2015) Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism & Related Disorders*. 21 (6), 650–653.

Audhkhasi, K., Georgiou, P.G. & Narayanan, S.S. (2011) Reliability-weighted acoustic model adaptation using crowd sourced transcriptions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. (August), 3045–3048.

Aydin, B.I., Yilmaz, Y.S., Li, Y., Li, Q., Gao, J. & Demirbas, M. (2014) 'Crowdsourcing for multiple-choice question answering', in *Proceedings of the National Conference on Artificial Intelligence*. [Online]. 2014 pp. 2946–2953.

Balicki, J., Brudło, P. & Szpryngier, P. (2014) Crowdsourcing and Volunteer Computing as Distributed Approach for Problem Solving. *Applications of Information Systems in Engineering and Bioscience*. 115–121.

Barbier, G., Zafarani, R., Gao, H., Fung, G. & Liu, H. (2012) Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*. 18 (3), 257–279.

Barlow, J., Wright, C., Sheasby, J., Turner, A. & Hainsworth, J. (2002) Self-management approaches for people with chronic conditions: A review. *Patient Education and Counseling*. 48 (2), 177–187.

Bederson, B.B. & Quinn, A.J. (2011) Web workers unite! Addressing challenges of online laborers. *Conference on Human Factors in Computing Systems - Proceedings*. 97–105.

Bengoechea, J.J., Villanueva, A. & Cabeza, R. (2012) Hybrid eye detection algorithm for outdoor environments. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. 685.

Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S. & Yeh, T. (2010) 'VizWiz: Nearly Real-time Answers to Visual Questions', in *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*. UIST ’10. [Online]. 2010 New York, NY, USA: ACM. pp. 333–342.

Bigham, J.P., Ladner, R.E. & Borodin, Y. (2011) The design of human-powered access technology. The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS ’11. 3.

Bigham, J.P., White, S.S., Yeh, T., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C.R., Tatarowicz, A. & White, B. (2010) *VizWiz: Nearly real-time answers to visual questions.* [Online] [online]. Available from: http://portal.acm.org/citation.cfm?doid=1866029.1866080%5Cnhttp://dl.acm.org/citation.cfm?id=1866029.1866080 (Accessed 6 January 2015).

Bockes, F., Edel, L., Ferstl, M. & Schmid, A. (2015) Collaborative landmark mining with a gamification approach. *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia - MUM ’15*. (Mum), 364–367.

Bove, R., Secor, E., Healy, B.C., Musallam, A., Vaughan, T., Glanz, B.I., Greeke, E., Weiner, H.L., Chitnis, T., Wicks, P. & De Jager, P.L. (2013) Evaluation of an Online Platform for Multiple Sclerosis Research: Patient Description, Validation of Severity Scale, and Exploration of BMI Effects on Disease Course. *PLoS ONE*. 8 (3), .

Brabham, D.C. (2013) *Crowdsourcing*. The MIT Press.

Brabham, D.C. (2008) Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*. 14 (1), 75–90.

Brabham, D.C. (2010) Moving the Crowd At Threadless. *Information, Communication & Society*. 13 (8), 1122–1145.

Bragg, J., Mausam & Weld, D.S. (2018) Sprout: Crowd-powered task design for crowdsourcing. *UIST 2018 - Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 165–176.

Brambilla, M., Ceri, S., Mauri, A. & Volonterio, R. (2015) An Explorative Approach for Crowdsourcing Tasks Design. *Www 2015*. 1125–1130.

Braun, V. & Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in*

Psychology. 3 (2), 77–101.

Bulling, A. & Gellersen, H. (2010) Toward mobile eye-based human-computer interaction. *Pervasive Computing, IEEE*. 8–12.

Burton, M.A., Brady, E., Brewer, R., Neylan, C., Bigham, J.P. & Hurst, A. (2012) 'Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities', in *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '12. [Online]. 2012 New York, NY, USA: ACM. pp. 135–142.

Cai, C.J., Iqbal, S.T. & Teevan, J. (2016) Chain reactions: The impact of order on microtask chains. *Conference on Human Factors in Computing Systems - Proceedings*. 3143–3154.

Callison-Burch, C. & Dredze, M. (2010) 'Creating Speech and Language Data With Amazon's Mechanical Turk', in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. [Online]. June 2010 Los Angeles: Association for Computational Linguistics. pp. 1–12.

Candido dos Reis, F.J., Lynn, S., Ali, H.R., Eccles, D., Hanby, A., Provenzano, E., Caldas, C., Howat, W.J., McDuffus, L.A., Liu, B., Daley, F., Coulson, P., Vyas, R.J., Harris, L.M., Owens, J.M., Carton, A.F.M., McQuillan, J.P., Paterson, A.M., Hirji, Z., et al. (2015) Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer. *EBioMedicine*. 2 (7), 681–689.

Canter, G.J. (1963) Speech Characteristics of Patients with Parkinson&#x2019;s Disease: I. Intensity, Pitch, and Duration. *Journal of Speech and Hearing Disorders*. 28 (3), 221–229.

Catallo, I. & Martinenghi, D. (2017) 'The Dimensions of Crowdsourcing Task Design', in Jordi Cabot, Roberto De Virgilio, & Riccardo Torlone (eds.) *Web Engineering*. [Online]. 2017 Cham: Springer International Publishing. pp. 394–402.

Cavallo, R. & Jain, S. (2012) 'Efficient crowdsourcing contests', in *11th International Conference on Autonomous Agents and Multiagent Systems*. [Online]. 2012 pp. 677–686.

Cavender, A.C., Otero, D.S., Bigham, J.P. & Ladner, R.E. (2010) 'Asl-stem Forum: Enabling Sign Language to Grow Through Online Collaboration', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. [Online]. 2010 New York, NY, USA: ACM. pp. 2075–2078.

Chan, K.T., King, I. & Yuen, M. (2009) 'Mathematical Modeling of Social Games', in *2009 International Conference on Computational Science and Engineering*. [Online]. August 2009 pp. 1205–1210.

Chanal, V. (2008) How to invent a new business model based on crowdsourcing : the Crowdspirit ® case.

Chandler, J., Mueller, P. & Paolacci, G. (2014) Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*. 46 (1), 112–130.

Cheng, J., Teevan, J., Iqbal, S.T. & Bernstein, M.S. (2015) Break It Down: A Comparison of Macro- and Microtasks. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. 4061–4064.

Cheng, S., Sun, Z., Ma, X. & Forlizzi, J. (2015) Social Eye Tracking: Gaze Recall with Online Crowds. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 454–463.

Ching, A., Zegras, P.C. & Kennedy, S. (2012) A User-Flocksourced Bus Experiment in Dhaka : New Data Collection Technique with Smartphones Muntasir Imran Mamun. *Journal of the Transportation Research Board*. (November 2012), 1–41.

Chittilappilly, A.I., Chen, L. & Amer-Yahia, S. (2016) A Survey of General-Purpose Crowdsourcing Techniques. *IEEE Transactions on Knowledge and Data Engineering*. 28 (9), 2246–2266.

Chunara, R., Chhaya, V., Bane, S., Mekaru, S.R., Chan, E.H., Freifeld, C.C. & Brownstein, J.S. (2012) Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. *Malaria Journal*. 11 (1), 43.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. & Players, F. (2010) Predicting protein structures with a multiplayer online game. *Nature*. 466 (7307), 756–760.

Côté, N. (2011) Integral and Diagnostic Intrusive Prediction of Speech Quality.

Dai, P., Rzeszotarski, J.M., Paritosh, P. & Chi, E.H. (2015) And now for something completely different: Improving crowdsourcing workflows with micro-diversions. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. 628–638.

Dalton, N.S., Collins, E. & Marshall, P. (2015) Display Blindness? Looking Again at the

Visibility of Situated Displays using Eye Tracking. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. 3889–3898.

Damessie, T.T., Järvelin, K., Scholer, F. & Culpepper, J.S. (2016) The effect of document order and topic difficulty on assessor agreement. *ICTIR 2016 - Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. 73–76.

Damessie, T.T., Kim, J., Shane Culpepper, J. & Scholer, F. (2018) Presentation ordering eects on assessor agreement. *International Conference on Information and Knowledge Management, Proceedings*. 723–732.

Darley, F.L., Aronson Arnold E. (Arnold Elvin), 1928- (joint author.) & Brown Joe Robert, 1911- (joint author.) (1975) References interspersed. *Motor speech disorders*. Philadelphia : Saunders.

Deng, X.N., Joshi, K.D. & Galliers, R.D. (2016) The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful Through Value Sensitive Design. *MIS Q.* 40 (2), 279–302.

Dergousoff, K. & Mandryk, R.L. (2015) Mobile Gamification for Crowdsourcing Data Collection : Leveraging the Freemium Model. *CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1065–1074.

Deterding, S., Dixon, D., Khaled, R. & Nacke, L. (2011) 'From Game Design Elements to Gamefulness: Defining 'Gamification'', in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. MindTrek '11. [Online]. 2011 New York, NY, USA: ACM. pp. 9–15.

Difallah, D., Demartini, G. & Cudre-Mauroux, P. (2012) Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. *CrowdSearch*. 84226–30.

Difallah, D., Filatova, E. & Ipeirotis, P. (2018) Demographics and dynamics of Mechanical Turk workers. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 2018-Febua (August 2017), 135–143.

Doan, A., Ramakrishnan, R. & Halevy, A.Y. (2011) Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*. 54 (4), 86.

Dolmaya, J.M. (2011) The Ethics of Crowdsourcing. *Linguistica Antverpiensia*. (10), 97–110.

Doroudi, S., Kamar, E., Brunskill, E. & Horvitz, E. (2016) Toward a learning science for complex crowdsourcing tasks. *Conference on Human Factors in Computing Systems - Proceedings*. 2623–2634.

Dow, S., Kulkarni, A., Klemmer, S.R. & Hartmann, B. (2012) Shepherding the crowd yields better work. *CSCW*. 1013–1022.

Eickhoff, C. & de Vries, A.P. (2013) Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*. 16 (2), 121–137.

Estellés-Arolas, E. & González-Ladrón-de-Guevara, F. (2012) Towards an integrated crowdsourcing definition. *Journal of Information Science*.

Evanini, K. & Zechner, K. (2011) 'Using Crowdsourcing to Provide Prosodic Annotations for Non-Native Speech.', in [Online]. 2011 pp. 3069–3072.

Evans, J.D. (1996) *Straightforward statistics for the behavioral sciences*. Belmont, CA, US: Thomson Brooks/Cole Publishing Co.

Evans, K.M., Jacobs, R.A., Tarduno, J.A. & Pelz, J.B. (2012) *Collecting and Analyzing Eye-tracking Data in Outdoor Environments*. [Online] [online]. Available from: http://www.bcs.rochester.edu/people/robbie/Evans-etal_JEMR2012.pdf (Accessed 14 January 2015).

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P. & Cox, A.L. (2014) 'Designing for dabblers and deterring drop-outs in citizen science', in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. [Online]. 2014 pp. 2985–2994.

Felstinerf, A. (2011) Working the Crowd : Employment and Labor Law in the Crowdsourcing Industry. *Berkeley Journal of Employment & Labor Law*. 32 (1), 143–204.

Finnerty, A. & Kucherbaev, P. (2013) Keep it simple: Reward and task design in crowdsourcing. *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*. 2–5.

Fox, C. (2002) Current Perspectives on the Lee Silverman Voice Treatment (LSVT) for Individuals With Idiopathic Parkinson Disease. *American Journal of Speech-language Pathology - AM J SPEECH-LANG PATHOL*. 11111–123.

Fuhl, W., Kübler, T., Sippel, K., Rosenstiel, W. & Kasneci, E. (2015) Excuse: Robust pupil detection in real-world scenarios. *Proc. CAIP*. 9256.

Fuhl, W., Santini, T.C., Kübler, T. & Kasneci, E. (2016) ElSe : Ellipse Selection for Robust

Pupil Detection in Real-World Environments. *Eye Tracking Research & Applications*. 123–130.

Fuhl, W., Tonsen, M., Bulling, A. & Kasneci, E. (2016) Pupil detection in the wild : An evaluation of the state of the art in mobile head-mounted eye tracking. *Machine Vision and Applications*.

Fung, G. (2011) Active Learning from Crowds. *Icml 2011*. 1161–1168.

Gadiraju, U., Demartini, G., Kawase, R. & Dietze, S. (2015) Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems*. 30 (4), 81–85.

Gadiraju, U., Fetahu, B., Kawase, R., Siehndel, P. & Dietze, S. (2017) Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction*. 24 (4), .

Gadiraju, U., Kawase, R. & Dietze, S. (2014) 'A Taxonomy of Microtasks on the Web', in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT '14. [Online]. 2014 New York, NY, USA: ACM. pp. 218–223.

Gadiraju, U., Kawase, R. & Dietze, S. (2015) 'Understanding Malicious Behavior in Crowdsourcing Platforms : The Case of Online Surveys', in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. [Online]. 2015 pp. 1631–1640.

Gadiraju, U., Yang, J. & Bozzon, A. (2017) Clarity is a worthwhile quality - On the role of task clarity in microtask crowdsourcing. *HT 2017 - Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 5–14.

GALTON, F. (1907) Vox Populi. *Nature*. 450–451.

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A. & Verroios, V. (2016) Challenges in Data Crowdsourcing. *IEEE Trans. on Knowl. and Data Eng.* 28 (4), 901–911.

Gary, W. & S., L.J. (2002) Direct Magnitude Estimates of Speech Intelligibility in Dysarthria. *Journal of Speech, Language, and Hearing Research*. 45 (3), 421–433.

Geiger, D. & Schader, M. (2014) Personalized task recommendation in crowdsourcing information systems — Current state of the art. *Decision Support Systems*. 653–16.

Gerber, E. & Hui, J. (2013) Crowdfunding: Motivations and deterrents for participation. *ACM Transactions on Computer-Human Interaction (…*. 20 (6), .

Goldberg, J.H. & Wichansky, A.M. (2003) 'Eye Tracking in Usability Evaluation: A Practitioner's Guide', in J Hyönä, R Radach, & H Deubel (eds.) *The Mind's Eye*. [Online]. Amsterdam: North-Holland. pp. 493–516.

Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E. & Kostakos, V. (2015) Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks*. 9034–48.

Goto, M. & Ogata, J. (2011) 'PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions.', in [Online]. 2011 pp. 3073–3076.

Green, J., Forster, A., Bogle, S. & Young, J. (2002) Physiotherapy for patients with mobility problems more than 1 year after stroke: a randomised controlled trial. *The Lancet*. 359 (9302), 199–203.

Gurari, D. & Grauman, K. (2016) Visual Question: Predicting If a Crowd Will Agree on the Answer. *Chi17*. 3511–3522.

Halder, B. (2014) Crowdsourcing Collection of Data for Crisis Governance in the Post-2015 World : Potential Offers and Crucial Challenges. *ICEGOV '14 Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*. 1–10.

Hansen, D.L., Schone, P., Corey, D., Reid, M. & Gehring, J. (2013) Quality control mechanisms for crowdsourcing: Peer review, arbitration, & expertiseat familysearch indexing. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 649–660.

Hansen, D.W. & Pece, A.E.C. (2005) Eye tracking in the wild. *Computer Vision and Image Understanding*. 98 (1), 155–181.

Hara, K., Milland, K., Hanrahan, B. V., Callison-Burch, C., Adams, A., Savage, S. & Bigham, J.P. (2019) Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. *Conference on Human Factors in Computing Systems - Proceedings*. 1–6.

Harris, C.G. (2011) You're Hired ! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. *Proceedings of the Workshop on Crowdsourcing for …*. (Csdm), 15–18.

Harrison, L., Skau, D., Franconeri, S., Lu, A. & Chang, R. (2013) Influencing visual judgment through affective priming. *Conference on Human Factors in Computing Systems - Proceedings*. 2949–2958.

Hipp, J.A., Manteiga, A., Burgess, A., Stylianou, A. & Pless, R. (2015) Cameras and Crowds in Transportation Tracking. *Proceedings of the conference on Wireless Health - WH '15*. 1–8.

Ho, A.K., Iansek, R., Marigliani, C., Bradshaw, J.L. & Gates, S. (1999) Speech Impairment in a Large Sample of Patients with Parkinson's Disease. *Behavioural Neurology*. 11 (3), 131–137.

Holmqvist, K., Nyström, M. & Mulvey, F. (2012) Eye tracker data quality: what it is and how to measure it. *Eye Tracking Research & Applications*. 1 (212), 45–52.

Hosseini, M., Phalp, K., Taylor, J. & Ali, R. (2014) 'The four pillars of crowdsourcing: A reference model', in *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. [Online]. 2014 pp. 1–12.

Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S. & Keimel, C. (2014) Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force '"Crowdsourcing"'. *COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services*.

Howe, J. (2008) C ROWDSOURCING Why the Power of the Crowd is Driving the Future of Business. *Achievement THE International INSTITUTE*. unedited e9.

Howe, J. (2006) The Rise of Crowdsourcing. *Wired*. 14.

Hu, Q., Wang, S., Author, C., Ma, L., Bie, R. & Cheng, X. (2017) Anti-Malicious Crowdsourcing Using the Zero-Determinant Strategy. *IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 1137–1146.

Huang, S.W. & Fu, W.T. (2013) Enhancing reliability using peer consistency evaluation in human computation. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 639–647.

Ipeirotis, P. (2010) Analyzing the Amazon Mechanical Turk Marketplace. *ACM Crossroads*. 17 (December 2010), 16–21.

Irani, L.C. & Silberman, M.S. (2013) 'Turkopticon : Interrupting Worker Invisibility in Amazon Mechanical Turk', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [Online]. 2013 New York, NY, USA: Association for Computing Machinery. pp. 611–620.

J. Holmes, R., M. Oates, J., J. Phyland, D. & J. Hughes, A. (2000) Voice characteristics in the progression of Parkinson’s disease. *International Journal of Language & Communication Disorders*. 35 (3), 407–418.

Jacob, R.J.K., and Karn, K.S. (2003) “Eye tracking in human computer interaction and usability research: Ready to deliver the promises”, In The Mind’s Eye: Cognitive and Applied Aspects of Eye Movement Research. *The Mind’s Eye: Cognitive and Applied Aspects of Eye Movement Research.* 573–605.

Jain, A., Das Sarma, A., Parameswaran, A. & Widom, J. (2017) Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment*. 10 (7), 829–840.

Javadi, A.-H., Hakimi, Z., Barati, M., Walsh, V. & Tcheang, L. (2015) SET: a pupil detection method using sinusoidal approximation. *Frontiers in neuroengineering*. 8 (April), 4.

Johnstone, D., Tate, M. & Fielt, E. (2018) Taking rejection personally: An ethical analysis of work rejection on Amazon Mechanical Turk. *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018*. 1–12.

Jovian, L.T. (2011) OCR Correction via Human Computational Game National University of Singapore Olivier Amprimo. *Sciences-New York*. 1–10.

Jung, H.J., Park, Y. & Lease, M. (2014) 'Predicting Next Label Quality : A Time-Series Model of Crowdwork Method : Latent Autoregressive Model', in *HComp’14 Proceedings of the AAAI Conference on Human Computation*. [Online]. 2014 Association for the Advancement of Artificial Intelligence. p.

Kaspar, A., Patterson, G., Kim, C., Aksoy, Y., Matusik, W. & Elgharib, M. (2018) 'Crowd-Guided Ensembles: How Can We Choreograph Crowd Workers for Video Segmentation?', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. [Online]. 2018 New York, NY, USA: ACM. pp. 111:1--111:12.

Kassner, M., Patera, W. & Bulling, A. (2014) Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. *arXiv preprint*. 10.

Kaufmann, N. & Veit, D. (2011) More than fun and money . Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. (2009), 1–11.

Kazai, G., Kamps, J. & Milic-Frayling, N. (2011) Worker types and personality traits in crowdsourcing relevance labels. *International Conference on Information and Knowledge Management, Proceedings*. (December 2013), 1941–1944.

Ke, W. & Zhang, P. (2008) Motivations in OSS Communities : The Mediating Role of Effort Intensity and Goal Commitment Weiling Ke School of Business Clarkson University School of Information Studies Syracuse University Motivations in OSS Communities : The Mediating Role of Effort I. 1–40.

Khazankin, R., Schall, D. & Dustdar, S. (2012) 'Predicting QoS in Scheduled Crowdsourcing', in *Proceedings of the 24th international conference on Advanced Information Systems Engineering*. [Online]. 2012 Springer Berlin Heidelberg. pp. 460–472.

Kim, J., Cheng, J. & Bernstein, M.S. (2014) Ensemble: Exploring Complementary Strengths of Leaders and Crowds in Creative Collaboration. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. 745–755.

Kim, J., Leksikov, S., Thamjamrassri, P., Lee, U. & Suk, H.J. (2015) Crowd color: Crowdsourcing color perceptions using mobile devices. *MobileHCI 2015 - Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 478–483.

Kittur, A. (2010) Crowdsourcing, collaboration and creativity. *XRDS: Crossroads, The ACM Magazine for Students*. 17 (2), 22.

Kittur, A., Chi, E.H. & Suh, B. (2008) 'Crowdsourcing user studies with Mechanical Turk', in *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems*. [Online]. 2008 pp. 453–456.

Kittur, A., Nickerson, J. V., Bernstein, M.S., Gerber, E.M., Shaw, A., Zimmerman, J., Lease, M. & Horton, J.J. (2013) *The Future of Crowd Work*. [Online] [online]. Available from: https://www.lri.fr/~mbl/ENS/CSCW/2012/papers/Kittur-CSCW13.pdf (Accessed 8 January 2015).

Kittur, A., Smus, B. & Kraut, R. (2011) CrowdForge Crowdsourcing Complex Work. Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11. 1801.

Klaib, A.F., Alsrehin, N.O., Melhem, W.Y. & Bashtawi, H.O. (2019) IoT smart home using eye tracking and voice interfaces for elderly and special needs people. *Journal of Communications*. 14 (7), 614–621.

Klin, A., Jones, W., Schultz, R., Volkmar, F. & Cohen, D. (2002) Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*. 59 (9), 809–816.

Komarov, S., Reinecke, K. & Gajos, K.Z. (2013) 'Crowdsourcing Performance Evaluations of User Interfaces', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’13. [Online]. 2013 New York, NY, USA: ACM. pp. 207–216.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W. & Torralba, A. (2016) Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2176–2184.

Kritikos, K., Pernici, B., Plebani, P., Cappiello, C., Comuzzi, M., Benrernou, S., Brandic, I., Kertész, A., Parkin, M. & Carro, M. (2013) A Survey on Service Quality Description. *ACM Comput. Surv*. 46 (1), .

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A. & Duin, R.P.W. (2003) Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*. 6 (1), 22–31.

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V. & Steen, N. (2014) Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology*. 16 (4), 408–416.

Landis, J.R. & Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 33 (1), 159–174.

Laput, G., Lasecki, W.S., Wiese, J., Xiao, R., Bigham, J.P., Harrison, C. & Boulevard, J.C.W. (2015) Zensors : Adaptive , Rapidly Deployable , Human - Intelligent Sens or Feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1935–1944.

Lasecki, Walter S, Gordon, M., Dow, S.P., Bigham, J.P., Koutra, D., Jung, M.F., Dow, S.P. & Bigham, J.P. (2014) Glance: Rapidly Coding Behavioral Video with the Crowd. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. 551–562.

Lasecki, Walter S., Homan, C. & Bigham, J.P. (2014) Architecting Real-Time Crowd-Powered Systems. *Human Computation*. 1 (1), 67–93.

Lasecki, W.S., Rzeszotarski, J.M., Marcus, A. & Bigham, J.P. (2015) The Effects of Sequence and Delay on Crowd Work. *Proceedings of the ACM CHI’15 Conference on Human Factors in Computing Systems*. 11375–1378.

Lasecki, Walter S., Teevan, J. & Kamar, E. (2014) Information extraction and manipulation threats in crowd-powered systems. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW ’14*. 248–256.

Lasecki, Walter S, Teevan, J. & Kamar, E. (2014) 'Information Extraction and Manipulation Threats in Crowd-powered Systems', in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*. CSCW ’14. [Online]. 2014 New York, NY, USA: ACM. pp. 248–256.

Lee, J.W., Heo, H. & Park, K.R. (2013) A novel gaze tracking method based on the generation of virtual calibration points. *Sensors (Basel, Switzerland)*. 13 (8), 10802–22.

Leimeister, J.M., Huber, M., Bretschneider, U. & Krcmar, H. (2009) Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition. *Journal of Management Information Systems*. 26 (1), 197–224.

Li, D., Babcock, J. & Parkhurst, D.J. (2006) openEyes: a low-cost head-mounted eye-tracking solution. *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications*. 95–100.

Liang, H., Wang, M.-M., Wang, J.-J. & Xue, Y. (2018) How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: A mediated moderation model. *Computers in Human Behavior*. 81168–176.

Licklider, J.C.R. (1960) Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*. HFE-1 (1), 4–11.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P. & Vandenberg, J. (2008) Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*. 389 (3), 1179–1189.

Luz, N., Silva, N. & Novais, P. (2015) A survey of task-oriented crowdsourcing. *Artificial Intelligence Review*. 44 (2), 187–213.

Mackenzie, A.K. & Harris, J.M. (2014) Characterizing Visual Attention During Driving and Non-driving Hazard Perception Tasks in a Simulated Environment. *Proceedings of the Symposium on Eye Tracking Research and Applications*. 2011127–130.

Majaranta, P. & Bulling, A. (2014a) Advances in Physiological Computing Stephen H. Fairclough & Kiel Gilleade (eds.). *Advances in Physiological Computing*. 39–65.

Majaranta, P. & Bulling, A. (2014b) 'Eye Tracking and Eye-Based Human-Computer Interaction', in Stephen H Fairclough & Kiel Gilleade (eds.) *Advances in Physiological Computing*. [Online]. London: Springer London. pp. 39–65.

Malone, T.W., Laubacher, R. & Dellarocas, C. (2010) The collective intelligence genome. *IEEE Engineering Management Review*. 38 (3), 38–52.

Malone, T.W., Laubacher, R. & Dellarocas, C.N. (2011) Harnessing Crowds: Mapping the Genome of Collective Intelligence. *SSRN Electronic Journal*.

Mantuano, A., Bernardi, S. & Rupi, F. (2016) Cyclist gaze behavior in urban space: An eye-tracking experiment on the bicycle network of Bologna. *Case Studies on Transport Policy*.

Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M.E., Lintott, C.J. & Smith, A.M. (2013) Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing. *First AAAI Conference on Human Computation and Crowdsourcing*. 94–102.

Mao, K., Harman, M. & Jia, Y. (2017) Crowd intelligence enhances automated mobile testing. ASE 2017 - Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering. 16–26.

Marcus, A., Karger, D., Madden, S., Miller, R. & Oh, S. (2012) Counting with the crowd. *Proceedings of the VLDB Endowment ,*. 6 (2), 109–120.

Marge, M., Banerjee, S. & Rudnicky, A.I. (2010) Using the Amazon Mechanical Turk for transcription of spoken language. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 5270–5273.

Marlow, J. & Dabbish, L. (2014) Who’s the boss? Requester transparency and motivation in a microtask marketplace. *Conference on Human Factors in Computing Systems - Proceedings*.

Mason, W. & Watts, D.J. (2009) Financial incentives and the performance of crowds. *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 11 (2), 77–85.

Massung, E. & Preist, C. (2013) *Normification: using crowdsourced technology to affect third-party change*. [Online] [online]. Available from: http://dl.acm.org/citation.cfm?id=2468356.2468615&coll=DL&dl=ACM&CFID=506957690&CFTOKEN=83409190 (Accessed 9 January 2015).

McAllister Byun, T., Halpin, P.F. & Szeredi, D. (2015) Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*. 5370–83.

McGraw, I., Gruenstein, A. & Sutherland, A. (2009) 'A self-labeling speech corpus: Collecting spoken words with an online educational game', in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. [Online]. 1 January 2009 pp. 3031–3034.

Mcinnis, B., Cosley, D., Nam, C. & Leshed, G. (2016) Taking a HIT : Designing around Rejection , Mistrust , Risk , and Workers ' Experiences in Amazon Mechanical Turk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2271–2282.

McKenzie, J.A. (1992) The provision of speech, language and hearing services in a rural district of South Africa. *The South African journal of communication disorders = Die Suid-Afrikaanse tydskrif vir Kommunikasieafwykings*. 3950—54.

Mcnaney, R., Othman, M., Richardson, D., Dunphy, P., Amaral, T., Miller, N., Stringer, H., Olivier, P. & Vines, J. (2016) Speeching : Mobile Crowdsourced Speech Assessment to S upport Self - Monitoring and M anagement for People with Parkinson ' s. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 4464–4476.

Meade, A.W. & Craig, S.B. (2012) Identifying careless responses in survey data. *Psychological Methods*. 17 (3), 437–455.

Miller, N. (2013) Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders*. 48 (6), 601–612.

Miller, N., Allcock, L., Jones, D., Noble, E., Hildreth, A.J. & Burn, D.J. (2007) Prevalence and pattern of perceived intelligibility changes in Parkinson{\textquoteright}s disease. *Journal of Neurology, Neurosurgery & Psychiatry*. 78 (11), 1188–1190.

Miller, N., Deane, K.H.O., Jones, D., Noble, E. & Gibb, C. (2011) National survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: therapists' practices. *International journal of language & communication disorders / Royal College of Speech & Language Therapists*. 46 (2), 189–201.

Miller, N., Noble, E., Jones, D., Allcock, L. & Burn, D.J. (2008) How do I sound to me? Perceived changes in communication in Parkinson's disease. *Clinical Rehabilitation*. 22 (1), 14–22.

Miller, N., Noble, E., Jones, D. & Burn, D. (2006) Life with communication changes in Parkinson's disease. *Age and Ageing*. 35 (3), 235–239.

Miller, N., Noble, E., Jones, D., Deane, K.H.O. & Gibb, C. (2011) Survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: patients' and carers' perspectives. *International Journal of Language & Communication Disorders*. 46 (2), 179–188.

Minder, P. & Bernstein, A. (2012) 'CrowdLang: A Programming Language for the Systematic Exploration of Human Computation Systems', in Karl Aberer, Andreas Flache, Wander Jager, Ling Liu, Jie Tang, & Christophe Guéret (eds.) *Social Informatics*. [Online]. 2012 Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 124–137.

Mitra, T., Hutto, C.J. & Gilbert, E. (2015) Comparing person- And process-centric strategies for obtaining quality data on amazon mechanical turk. *Conference on Human Factors in Computing Systems - Proceedings*. 2015-April1345–1354.

Morris, R.R., Dontcheva, M. & Gerber, E.M. (2012) Priming for Better Performance in Microtask Crowdsourcing Environments. *IEEE Internet Computing*. 16 (5), 13–19.

Morrow, E. & Scorgie-Porter, L. (2017) Bowling Alone: The Collapse and Revival of American Community. *Bowling Alone: The Collapse and Revival of American Community*. 6 (2), 1–84.

Moussawi, S. & Koufaris, M. (2013) The crowd on the assembly line: Designing tasks for a better crowdsourcing experience. *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design*. 43745–3761.

Munn, S.M., Stefano, L. & Pelz, J.B. (2008) Fixation-identification in dynamic scenes. *Proceedings of the 5th symposium on Applied perception in graphics and visualization - APGV '08*. 1 (212), 9.

Naderi, B., Wechsung, I., Polzehl, T., Möller, S., Wechsung, I., Polzehl, T. & De, S.M. (2014) Development and Validation of Extrinsic Motivation Scale for Crowdsourcing Micro-task Platforms. *CrowdMM '14 Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*. 31–35.

Nakatsu, R.T., Grossman, E.B. & Iacovou, C.L. (2014) A Taxonomy of Crowdsourcing Based on Task Complexity. *J. Inf. Sci.* 40 (6), 823–834.

Naroditskiy, V., Jennings, N.R., Hentenryck, P. Van & Cebrian, M. (2014) Crowdsourcing contest dilemma. *Journal ofthe Royal Society*. 111–8.

Nebeling, M., To, A., Guo, A., de Freitas, A.A., Teevan, J., Dow, S.P. & Bigham, J.P. (2016)

WearWrite: Crowd-Assisted Writing from Smartwatches. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3834–3846.

Nguyen, T.B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., Burns, J.E. & Summers, R.M. (2012) Distributed Human Intelligence for Colonic Polyp Classification in Computer-aided Detection for CT Colonography. *Radiology*. 262 (3), 824–833.

Nijkrake, M.J., Keus, S.H.J., Kalf, J.G., Sturkenboom, I.H.W.M., Munneke, M., Kappelle, A.C. & Bloem, B.R. (2007) Allied health care interventions and complementary therapies in Parkinson’s disease. *Parkinsonism & Related Disorders*. 13S488–S494.

Niu, X.-J., Qin, S.-F., Vines, J., Wong, R. & Lu, H. (2019) Key Crowdsourcing Technologies for Product Design and Development. *International Journal of Automation and Computing*. 16 (February), 1–15.

Norcie, G. (2011) Ethical and Practical Considerations For Compensation of Crowdsourced Research Participants. CHI Extended Abstracts, Workshop on Ethics Logs and VideoTape: Ethics in Large Scale Trials & User Generated Content.

Nordin, A., Ahmad Zaidi, N.H. & Mazlan, N.A. (2017) Measuring software requirements specification quality. *Journal of Telecommunication, Electronic and Computer Engineering*. 9 (3-5 Special Issue), 123–128.

Nunes, F. & Fitzpatrick, G. (2015) Self-Care Technologies and Collaboration. *International Journal of Human–Computer Interaction*. 31 (12), 869–881.

Odobasic, D., Medak, D. & Miler, M. (2013) Gamification of Geographic Data Collection. *Gi_Forum 2013: Creating the Gisociety*. (July 2013), 328–337.

Ohm, C., Mueller, M., Ludwig, B. & Bienk, S. (2014) "'Where is the Landmark? Eye Tracking Studies in Large-Scale Indoor Environments', in *Proceedings of the 2nd International Workshop on Eye Tracking for Spatial Research*. [Online]. 2014 pp. 47–51.

Parent, G. & Eskenazi, M. (2011) Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges.

Patientslikeme (2015) *Live Better, Together!* [Online] [online]. Available from: https://www.patientslikeme.com/.

Pavisic, I., Primativo, S., Yong, K., Russell, L., J.D. Hardy, C., Bond, R., Marshall, C., Brotherhood, E., Warren, J., D. Rohrer, J. & Crutch, S. (2018) 'CAN EYETRACKING

METRICS PROVIDE INSIGHT INTO THE DIAGNOSIS OF DIFFERENT DEMENTIA TYPES? A SPATIAL ANTICIPATION TASK', in *Alzheimer's & Dementia*. [Online]. 2018 pp. P930–P931.

Peißl, S., Wickens, C.D. & Baruah, R. (2018) Eye-Tracking Measures in Aviation: A Selective Literature Review. *The International Journal of Aerospace Psychology*. 00 (00), 1–15.

Posch, L., Bleier, A., Lechner, C., Danner, D., Flöck, F. & Strohmaier, M. (2017) Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. *arXiv*. (September), .

Prestopnik, N. & Crowston, K. (2012) Citizen science system assemblages: understanding the technologies that support crowdsourced science. *Proceedings of the 2012 iConference*. 168–176.

Purucker, C., Naujoks, F., Prill, A. & Neukum, A. (2017) Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling. *Applied Ergonomics*. 58543–554.

Quinn, A.J. & Bederson, B.B. (2011) 'Human computation: A survey and taxonomy of a growing field', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [Online]. 2011 Association for Computing Machinery. pp. 1403–1412.

Quinn, E.J. & Bederson, B.B. (n.d.) B.B.: A taxonomy of distributed human computation.

Ramig, L.O., Sapir, S., Countryman, S., Pawlas, A.A., O\textquoterightBrien, C., Hoehn, M. & Thompson, L.L. (2001) Intensive voice treatment (LSVT®) for patients with Parkinson{\textquoteright}s disease: a 2 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry*. 71 (4), 493–498.

Riegler, A. & Holzmann, C. (2018) Measuring Visual User Interface Complexity of Mobile Applications With Metrics. *Interacting with Computers*. 30 (3), 207–223.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. & Vukovic, M. (2011) An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Fifth International AAAI Conference on Weblogs and Social Media*. (Gibbons 1997), 321–328.

Rudoy, D., Goldman, D.B., Shechtman, E. & Zelnik-Manor, L. (2012) Crowdsourcing gaze data collection. *Proceedings Collective Intelligence*.

Russell, B.C., Torralba, A., Murphy, K.P. & Freeman, W.T. (2005) LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*. 77 (1–3), 157–173.

Ryan, R. & Deci, E. (2000) Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary educational psychology*. 25 (1), 54–67.

Sampath, H.A., Rajeshuni, R. & Indurkhya, B. (2014) Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. *Proceedings of the 32th International Conference on Human Factors in Computing Systems, CHI 2014*. 3665–3674.

Sánchez-Ferrer, M., Grima Murcia, M., Sánchez Ferrer, F., Isabel Hernández-Peñalver, A., Fernandez, E. & del Campo, F. (2017) Use of Eye Tracking as an Innovative Instructional Method in Surgical Human Anatomy. *Journal of Surgical Education*. 74.

Sandhu, S.K. & Anupam Agarwal (2015) Summarizing Videos by Key frame extraction using SSIM and other Visual Features. *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015*. 209–213.

Sari, P.K., Alamsyah, A. & Wibowo, S. (2018) Measuring e-Commerce service quality from online customer review using sentiment analysis. *Journal of Physics: Conference Series*. 97112053.

Satzger, B., Psaier, H., Schall, D. & Dustdar, S. (2013) Auction-based crowdsourcing supporting skill management. *Information Systems*. 38 (4), 547–560.

Saxton, G.D., Oh, O. & Kishore, R. (2013) Rules of Crowdsourcing: Models, Issues, and Systems of Control. *Information Systems Management*. 30 (1), 2–20.

Schall, D., Truong, H. & Dustdar, S. (2008) 'The Human-Provided Services Framework', in 2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services. [Online]. July 2008 pp. 149–156.

Schenk, E. & Guittard, C. (2009) Crowdsourcing : What can be Outsourced to the Crowd , and Why? *Innovation*. 1–29.

Schenk, E. & Guittard, C. (2011) Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics*. 7 (1), 93.

Schnitzer, S., Rensing, C. & Schmidt, S. (2015) Demands on task recommendation in crowdsourcing platforms - the worker’s perspective. *In workshop of Crowdsourcing and* *human computation for recommender systems(CrowdRec2015), ACM RecSys*. 1–7.

Schrom-Feiertag, H., Schinko, C., Settgast, V. & Seer, S. (2014) Evaluation of guidance systems in public infrastructures using eye tracking in an immersive virtual environment. *Proceedings of the 2nd International Workshop on Eye Tracking for Spatial Research*. 124162–66.

Schulze, T., Nordheimer, D. & Schader, M. (2013) Worker Perception of Quality Assurance Mechanisms in Crowdsourcing and Human Computation Markets. *19th Americas Conference on Information Systems, AMCIS 2013*. 5 (August 2013), 4046–4056.

Shao, Y., Liu, Y., Zhang, F., Zhang, M. & Ma, S. (2019) On Annotation Methodologies for Image Search Evaluation. *ACM Trans. Inf. Syst.* 37 (3), 29:1--29:32.

Shaw, A.D., Horton, J.J. & Chen, D.L. (2011) Designing incentives for inexpert human raters. *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*. 45275.

Sheng, V.S., Provost, F. & Ipeirotis, P.G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 614–622.

Silberman, M.S., Irani, L. & Ross, J. (2010) Ethics and Tactics of Professional Crowdwork. *Xrds*. 17 (2), 39–43.

Singh, H., Bhatia, J.S. & Kaur, J. (2011) Eye tracking based driver fatigue monitoring and warning system. *India International Conference on Power Electronics, IICPE 2010*.

Singh, H. & Singh, J. (2012) Human Eye Tracking and Related Issues: A Review. *International Journal of Scientific and Research …*. 2 (9), 1–9.

Sinha, P., Balas, B., Ostrovsky, Y. & Russell, R. (2006) Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*. 94 (11), 1948–1961.

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A.Y. (2008) 'Cheap and fast - But is it good? Evaluating non-expert annotations for natural language tasks', in *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL*. [Online]. 2008 Honolulu: Association for Computational Linguistics. pp. 254–263.

Stol, K. & Fitzgerald, B. (2014) Two’s Company, Three’s a Crowd: A Case Study of Crowdsourcing Software Development. *Association for Computing Machinery*. (February), 187–198.

Su, H., Deng, J. & Fei-fei, L. (2012) Crowdsourcing Annotations for Visual Object Detection. *Proc. AAAI Human Computation’12*. 40–46.

Sugano, Y. & Bulling, A. (2015) Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency. *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 363–372.

Surowiecki, J. (2005) *The Wisdom of Crowds*. New York: Anchor Books.

Swan, M. (2012) Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen. Journal of Personalized Medicine 2 (3).

Swan, M., Hathaway, K., Hogg, C., McCauley, R. & Vollrath, A. (2010) Citizen Science Genomics as a Model for Crowdsourced Preventive Medicine Research. *Participatory Medicine*.

Swirski, L., Bulling, A. & Dodgson, N. (2012) Robust real-time pupil tracking in highly off-axis images. *Proceedings of the Symposium on Eye Tracking Research and Applications*. 1–4.

Terragni, V., Salza, P. & Pezzè, M. (2020) 'Measuring Software Testability Modulo Test Quality', in *Proceedings of the 28th International Conference on Program Comprehension*. ICPC ’20. [Online]. 2020 New York, NY, USA: Association for Computing Machinery. pp. 241–251.

Timm, F. & Barth, E. (2011) Accurate eye centre localisation by means of gradients. *Proc. Computer Vision Theory and Applications*. 125–130.

Tjaden, K. (2008) Speech and Swallowing in Parkinson’s Disease. *Topics in Geriatric Rehabilitation*. 24 (2), 115–126.

Tonsen, M., Zhang, X., Sugano, Y. & Bulling, A. (2016) Labeled pupils in the wild: A dataset for studying pupil detection in unconstrained environments. *Proc. Eye Tracking Research and Applications*. 139–142.

Vaish, R., Goel, S., Davis, J., Bernstein, M.S., Gaikwad, S. (Neil) S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I., Simoiu, C., Wilber, M. & Belongie, S. (2017) Crowd Research:

Open and Scalable University Laboratories. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17*. 829–843.

Valenti, R. & Gevers, T. (2012) Accurate eye center location through invariant isocentric patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 34 (9), 1785–1798.

Villanueva, A. & Cabeza, R. (2008) Gaze Estimation With One Calibration Point. *Proc. IEEE Transactions on Systems, Man, and Cybernetic*. 38 (4), 1123–1138.

Wang, Z., Simoncelli, E.P. & Bovik, A.C. (2003) Multi-scale structural similarity for image quality assessment. *IEEE Asilomar Conference on Signals, Systems and Computers*. 29–13.

Wicks, P., Keininger, D.L., Massagli, M.P., la Loge, C. de, Brownstein, C., Isojärvi, J. & Heywood, J. (2012) Perceived benefits of sharing health data between people with epilepsy on an online platform. *Epilepsy and Behavior*. 23 (1), 16–23.

Wight, S. & Miller, N. (2015) Lee Silverman Voice Treatment for people with Parkinson's: audit of outcomes in a routine clinic. *International Journal of Language & Communication Disorders*. 50 (2), 215–225.

Winfield, D. & Parkhurst, D.J. (2005) Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*. 379–79.

Wolters, M.K., Isaac, K.B. & Renals, S. (2010) Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk. *Proceedings of 7th Speech Synthesis Workshop (SSW7)*. 136–141.

Wu, M. & Quinn, A.J. (2017) Confusing the Crowd : Task Instruction Quality on Amazon Mechanical Turk. *The Fifth AAAI Conference on Human Computation and Crowdsourcing*. (Hcomp), 206–215.

Xiang, X.-H., Huang, X.-Y., Zhang, X.-L., Cai, C.-F., Yang, J.-Y. & Li, L. (2014) Many Can Work Better than the Best: Diagnosing with Medical Images via Crowdsourcing. *Entropy*. 16 (7), 3866–3877.

Xu, P., Ehinger, K. & Zhang, Y. (2015) TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *arXiv preprint arXiv: 1504.06755*.

Yang, Y., Zhu, B.B., Guo, R., Yang, L., Li, S. & Yu, N. (2008) 'A Comprehensive Human Computation Framework: With Application to Image Labeling', in *Proceedings of the 16th*

ACM International Conference on Multimedia. MM ’08. [Online]. 2008 New York, NY, USA: ACM. pp. 479–488.

You, C., Lane, N.D., Chen, F., Wang, R., Chen, Z., Bao, T.J., Montes-de-oca, M., Cheng, Y., Lin, M., Torresani, L. & Campbell, A.T. (2012) CarSafe App: Alerting Drowsy and Distracted Drivers using Dual Cameras on Smartphones Categories and Subject Descriptors. *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 1–14.

Yuen, M.-C., King, I. & Leung, K.-S. (2012) Task recommendation in crowdsourcing systems. *Proceedings of the First International Workshop on Crowdsourcing and Data Mining - CrowdKDD ’12*. 22–26.

Yuen, M.-C., King, I. & Leung, K. (2011) 'A Survey of Crowdsourcing Systems', in IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing. [Online]. 2011 pp. 766–773.

Zhang, Y.N. (2017) Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation. *Parkinson’s Disease*. 2017 (1), .

Zheng, H., Li, D. & Hou, W. (2011) Task Design, Motivation, and Participation in Crowdsourcing Contests. *International Journal of Electronic Commerce*. 15 (4), 57–88.

Zhu, H., Dow, S.P., Kraut, R.E. & Kittur, A. (2014) Reviewing versus doing: Learning and Performance in Crowd Assessment. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW ’14*. 1445–1455.

Ziegler, W. & Zierdt, A. (2008) Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders*. 41 (6), 553–577.