Speeching: Mobile Crowdsourced Speech Assessment to Support Self-Monitoring and Management for People with Parkinson's

Róisín McNaney[1], Mohammad Othman[1], Dan Richardson[1], Paul Dunphy[1], Telmo Amaral[1], Nick Miller[2], Helen Stringer[3], Patrick Olivier[1] and John Vines[1]

[1]Open Lab, Newcastle University, Newcastle upon Tyne, UK
[2]Institute of Ageing, Newcastle University, Newcastle upon Tyne, UK
[3]School of Education, Communication and Language Sciences, Newcastle University, UK
{r.mcnaney, m.othman1, patrick.olivier, john.vines}@ncl.ac.uk

ABSTRACT

We present Speeching, a mobile application that uses crowdsourcing to support the self-monitoring and management of speech and voice issues for people with Parkinson's (PwP). The application allows participants to audio record short voice tasks, which are then rated and assessed by crowd workers. Speeching then feeds these results back to provide users with examples of how they were perceived by listeners unconnected to them (thus not used to their speech patterns). We conducted our study in two phases. First we assessed the feasibility of utilising the crowd to provide ratings of speech and voice that are comparable to those of experts. We then conducted a trial to evaluate how the provision of feedback, using Speeching, was valued by PwP. Our study highlights how applications like Speeching open up new opportunities for self-monitoring in digital health and wellbeing, and provide a means for those without regular access to clinical assessment services to practice—and get meaningful feedback on—their speech.

Author Keywords

Crowdsourcing; Healthcare; Self-monitoring and Management; Speech and Language Therapy; Parkinson's.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

Within the medical domain, crowdsourcing is emerging as a way to both collect [48] and analyze [13] large data sets. Although the benefits of crowdsourcing as a research tool are well-acknowledged, the role it might play to extend and enhance everyday healthcare remains underexplored. This is particularly so in the context of personal health, where the benefits of self-care, including self-management and monitoring are widely advocated [4,42].

In this paper, we examine the role crowdsourcing could play in personal health through its application to speech and language therapy (SLT) for people with Parkinson's (PwP). Within adult services, SLT involves the training and implementation of specific skills and strategies that have been lost or diminished due to degenerative (e.g. Parkinson's, motor neurone disease, dementia) or acquired (e.g. stroke, traumatic brain injury) health conditions. Personal health practices are of particular importance here due to a need for consistent repetition of exercises in order to extend and maintain gains made in clinical settings. However, there are motivational barriers to the self-directed practice of speech [41] and often treatment effect does not persist in the long-term following therapy [23,53]. This is a concern magnified by an acknowledgement that the number of Speech and Language Therapists (SLTs) required for adequate therapy provision are lower than demand requirements, both in developed [35] and developing [33] countries. As such, there is a need for new approaches that better scaffold self-directed therapeutic practices.

We developed Speeching, a crowdsourced analysis system to support the provision of feedback on speech directly to the individual, as a means to facilitate self-care practices. We took the case of Parkinson's as a focused example, as PwP commonly experience a range of speech difficulties as a result of their condition [24,34]. The system comprises a smartphone application that allows individuals to practice a series of speech tasks and to upload these to a remote server. The recordings are then rated by crowd workers for ease of listening, speaking rate, pitch variability and volume. These ratings are then fed back to participants in order to provide therapeutic targets to support home practice of SLT tasks. In this paper we first demonstrate the feasibility of using crowdsourced judgments on the properties of recorded speech, compared to expert judgments. Based on these

CHI'16, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-3362-7/16/05.
http://dx.doi.org/10.1145/2858036.2858321


results we developed and deployed Speeching in a real world pilot study with PwP to establish its acceptability.

We highlight the potential that crowdsourcing offers to support new forms of self-care practices. Our paper provides several contributions to HCI. First, we demonstrate the feasibility of crowdsourcing as a means of producing quantitative ratings of PwP impaired speech, comparable to expert judgments. Second, we provide an example of crowdsourcing being used in the wild to present data directly back to patients and the benefits and challenges that our methods posed. We next offer an enhanced understanding of the impact that having crowdsourced ratings had on a group of PwP, the value they placed in the system and how it prompted their practice of therapeutic tasks. Finally, we offer insights for future researchers wishing to further explore the space of crowdsourcing for personal health.

BACKGROUND

Crowdsourcing Health

Much crowdsourcing research in healthcare has focused on the collection of data. For example: to understand whether members of large online health communities can act as representative patients of wider populations [7]; to utilize the personal data already being collected by health communities about themselves, to gain new understandings into preventative medicine [48,49]; to look at how existing online communities (e.g. [44]) can provide new sources of patient data for research; or simply to understand how online communities function in a supportive role among specific patient groups [52]. Others have used crowdsourcing to facilitate the analysis of patient data. In some cases, this involves the outsourcing of data to a crowd of experts. For example Crowdmed [58] allows people to post information about their medical condition to be 'solved' by medical experts, while [55] explored the crowdsourced analysis of medical imaging data by General Practitioners. There has also been work focusing on the use of non-expert crowds in the analysis of large scale clinical data including: the use of online games to support the identification of malarial parasites within blood samples [13], the prediction of genomic protein structures [14], and the classification of colonic polyps within radiography scans [40].

Beyond the healthcare context, crowdsourcing has been used within several interactive, user-supporting systems (see [5] for a wider overview of multiple works focusing on human powered assistive technologies). For example, *VizWiz* [6,9] is a smartphone application that provides near real-time feedback on visual information to blind people, while the ASL-STEM Forum is an online portal for contributing sign language describing scientific terminology for deaf or hard of hearing people [12]. However, as of yet, there has been relatively little work that has examined how non-expert crowd workers might support health self-management in real-world settings. In this work, we explore this gap in the literature by leveraging the crowd to provide feedback to PwP, to support their self-monitoring and management practices in daily life.

Crowdsourcing for Speech Data

A number of researchers have explored how crowdsourcing can be applied to speech analysis problems. Crowdsourcing has been used for the collection [32] and transcription [3,31,43,54] of speech data, as well as enabling the refinement of speech recognition systems [22]. Others have examined the use of crowdsourcing techniques to measure the quality of speech samples. Parent and Eskenazi [43] highlight the value of using reductive measures of intelligibility in their study, where they invited Amazon Mechanical Turk (AMT) workers to transcribe and classify short utterances produced by users of a transport information system. Marge et al. looked at the reliability of using AMT for transcribing spontaneous speech samples [30]. They found accuracy to be approaching expert agreement, and that using small segments of speech might yield faster turnaround time and better transcription accuracy. In regards to the rating of perceptual aspects of speech, Evanini [19] studied the use of crowdsourcing for annotating prosodic stress and boundary tones on a corpus of spontaneous speech on non-native speakers and found high levels of agreement when compared to experts.

These studies provide a range of examples of crowdsourcing for speech data, and have highlighted a number of methodological considerations in this domain. However, there are specific complexities to take into account when rating impaired speech for clinical populations. Therefore, it is important to understand existing clinical literature within Speech and Language Therapy and the current methods and practices being employed within speech and voice measurement for PwP.

Parkinson's Speech

It is approximated that 90% of all PwP will experience speech and voice issues at some point [24]. Common changes include a reduction in volume and changes to prosody (stress and intonation patterns in speech), which is associated with a tendency to speak on one loudness level (monoloudness) with little variation in pitch (monopitch). In addition, perceptual vocal quality can become impaired, leading to a hoarse, rough, breathy or tremulous speaking voice [25,50]. These characteristics can cause loss of confidence and embarrassment, particularly when speaking with strangers and can lead to a tendency to avoid social situations altogether [34,36,37].

The process of measuring speech and voice difficulties in Parkinson's is generally conducted by a qualified SLT. In a typical assessment session, this involves the collection of a variety of speech samples, which are then subjected to formal and/or informal testing. One issue with this procedure is the fact that SLTs are highly specialized and experienced in listening to impaired speech performance. It

has been argued that this familiarity can lead to higher scoring during testing [39]. To mitigate this, best practice guidelines suggest the use of naive listeners to provide a representative rating. However, implementing this within everyday clinical practice is difficult due to time and resource constraints [56]. Furthermore, research notes that access to SLT is low within PwP [38], meaning that many would not reach these services in the first instance.

Measuring intelligibility

In response to these challenges, researchers have started to explore the potential for speech intelligibility testing to be conducted remotely via online digital platforms. Ziegler and Zierdt [57] proposed the *Munich Intelligibility Profile (MVP) online* system as a means to remotely provide SLTs with intelligibility judgments on dysarthric[1] speech. While MVP online proved successful (showing a decrease in individual deviation from the mean with increased numbers of listeners), the system still required a level of external control—speech samples that were submitted for analysis were collected in a clinical setting and reviewed by an SLT. In addition, moderators assigned speech samples to listeners and collated and reviewed listeners' responses.

Within the context of speech intelligibility testing, the availability of large, affordable and spontaneous workforces through crowdsourcing platforms allows for a large number and variety of non-expert listeners. While no previous work has yet examined the potential for crowd workers to provide speech analysis that can feed into a program of speech therapy, the use of pre-existing crowdsourcing platforms in providing diagnostic speech ratings is emerging. Byun et al. [10] asked untrained listeners, recruited on AMT, to classify speech samples from children with /r/ misarticulation as either correct or incorrect, and compared those judgments to those of experienced listeners. They found that the agreement between those two groups of listeners was extremely high (r=0.98) and highlighted the potential for crowdsourcing to play a greater role in SLT practice. However, while this binary approach holds promise, there is currently little understanding of how more intricate measures of intelligibility can be elicited through crowdsourcing.

In our development and evaluation of Speeching we address these gaps by: (1) exploring novel methods towards both eliciting and collecting real world speech samples; and (2) exploring the potential for crowdsourcing to provide feedback on PwP speech. Our study was conducted in two phases. The first aimed to demonstrate the feasibility of using an anonymous online crowd to rate impaired speech. The second involved a real world deployment of Speeching: the collection of samples from, and provision of feedback to, PwP, unsupervised, in their home environment.

PHASE 1: TESTING SPEECHING FEASIBILITY

Selecting the sample dataset

In this first phase, the main aim was to explore the development of crowdsourcing tasks which might elicit ratings of Parkinson's speech equivalent to expert ratings. In order to cover the main elements of impairment within Parkinson's speech, as identified in the literature [25], the issues of rate, pitch variability and volume were selected to investigate. A sample of 12 speakers were selected from a pre-existing dataset compiled of 125 PwP, collected in a lab setting [34]. In order to select this sample dataset we asked an SLT, experienced in Parkinson's speech, to navigate the 125 samples and select a representative sample. Speakers made up equal categories of mild, moderate and severe intelligibility problems, with 2 male and 2 female speakers in each category. Each speaker provided 10 single word reading samples (unconnected speech) and 9 sentences (connected speech) taken from a reading sample; the Grandfather Passage [17].

Designing the mini-tasks

In order to design the Speeching analysis tasks we worked alongside an expert in Parkinson's speech. The tasks were designed in a manner similar to standard SLT assessment in which a therapist will listen to a range of single words, sentences and longer samples of speech, which are produced by asking the individual to read a word or piece of text, describe a picture or engage in free flowing discussion about a topic. The SLT then makes a decision about the prominent difficulties being experienced—whether this be volume control, alterations in speech rate, vocal qualities like breathiness, or otherwise. The SLT will use a range of standardized assessment to objectively measure these problems, alongside unstandardized methods which rely on their expertise. Often recordings of the PwP speech will be made as a record of their pre-therapy voice.

We chose to study two categories of speech sample. The first was unconnected speech, or single words. These were chosen as they provide a measure of intelligibility in isolation, without any additional context that might add to a listener's ability to make sense of the message being expressed by the speaker. This type of task is widely used in SLT assessment and was thus included in our exploratory work as it has the extended potential to allow for a more fine grained analysis of the specific sound contrasts a speaker is having difficulty with, providing direction for therapeutic input. Although we did not explore this fully in our own work, we wanted to include this task for crowd analysis to scope wider, future potential for the system. Crowd members were asked to select the target word from a set of 10 similar words (e.g. coop, cup, cape, cope). There were a total of 10 single words that were subjected to this word recognition task within each assessment (this test was

[1] 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].

part of an assessment conducted by [34], designed to target specific sound contrasts).

The second set of speech samples were sentence level (connected speech) utterances. For these we used two types of rating measures applied to each sample. The first was an Ease of Listening (EOL) rating, to provide a subjective measure of how much effort it took to understand the speaker. This five-point scale has been used successfully in the past with novice listeners unfamiliar with dysarthric speech and was found to have a strong correlation to intelligibility scores [27,34]. The second set of ratings, addressing perceptual measures of rate, pitch variance and volume, involved more complex judgments. When rating speech quality, Likert scales lack sensitivity [15,39] but the use of continuous scaling systems can mitigate some of this difficulty [15]. Miller suggests the use of Direct Magnitude Estimation (DME) for perceptual intelligibility measures, whereby an anchor, or midrange exemplar, of impaired speech is played to the listener to allow for an estimation of the magnitude of difference [39,51]. As our crowd workers were not experienced in disordered speech, and thus were likely to exhibit variability in their judgments of volume, pitch variance and rate, we used a continuous scale of 0-100. This allowed for a larger range of variability amongst raters to be exhibited, without impacting the sensitivity of ratings that may have been observed in a discrete scale.

In order to select the mid-range exemplar sample, our experienced SLT selected one male and one female speaker representing a moderate impairment in each measure (pitch variance, rate and volume) from the larger dataset of 125 speakers. These mid-range exemplar samples were not from speakers who had been included in our subsample of 12 speakers for analysis. Mid-range exemplar samples were gender matched to the participant samples that we used in our final dataset. Crowd workers were asked to rate the speech, out of 100, for volume, rate and pitch variance using the midrange exemplar as a reference point for a score of 50.

Participants

We opportunistically recruited 33 crowd workers from AMT (who were from the UK) to complete the tasks and the obtained ratings from two highly experienced experts in Parkinson's speech to act as a gold standard. Our experts completed the entire dataset (282 speech samples). Crowd workers were automatically assigned a crowdsourcing task in a random order. We sourced a minimum of 3 ratings for each speech sample. Tasks were therefore assigned to allow for this, whilst ensuring that the same listener did not rate a sample twice. Because there was variation in the number of tasks that listeners completed, samples which had received 3 different ratings were indexed and the task assignment was re-randomized until the entire dataset was complete. Listeners were provided with a progress bar so that they could see how much of the full dataset they had rated, but were only required to complete 25% (70 tasks) in order to receive payment for their time. Crowd workers were paid at the UK minimum wage based on an estimate by the research team of the average time to complete the tasks.

Phase 1 Analysis

Spearman's Rho was conducted to calculate the correlation between the crowd and experts on single word recognition and EOL tasks, as well as the measures of pitch, rate and volume. We selected Spearman's Rho taking into account that our observations were based on independent samples. Each sample was rated by a different set of raters, albeit in a structured manner. We felt that Spearman's Rho would be the best measure to capture potential differences between groups, as we were interested in exploring the correlation (even if not linear) between the experts and the crowd. To prepare our dataset we first calculated the success rates for the word recognition task. Successful recognition of the target was given a binary correct/ incorrect score which was then aggregated into a total % of words correct score across each of the speakers. For the measures of pitch, rate and volume we took the median score (out of 100) from each group of 3 raters who had analyzed the speech sample and then compared these to the median scores of the experts (median was chosen over mean to due to the nature of the continuous rating scale, to account for possible outliers in the data).

Phase 1 Findings

See Table 1 for a summary of results. Besides the interquartile range (IQR) of observed scores, the table also shows the values of the lower (Q_1) and upper quartiles (Q_3). For the word recognition task a strong correlation [20] was found between the scores of the experts and the crowd indicating that crowd members selected similar options to the expert during the word recognition task. Strong correlations between the experts and crowd on the measures of pitch and rate were also observed, indicating that the crowd scored these perceptual measures within a similar trend to the experts. For EOL a substantial agreement was found [28]. Overall, these scores indicate that non-expert workers, anonymously recruited via an online crowdsourcing platform, can provide equivalent ratings to experts in the measurement of speech and voice changes in a subsample of PwP speech. They also provide evidence to support the feasibility of our crowdsourcing method.

However, one measure that did not correlate well was volume, which provided only a weak correlation with the expert. One reason for this could be that the quality of the recordings; although they were collected in a systematic way, were not consistent between speakers. Given the comparative nature of the DME-style task, it is easy for the quality of a recording to be rated over the actual speaking volume. This was a limitation of the study and would need further consideration if volume judgment were to be included in the phase 2 study, to ensure that it is the measure being tested causing the rating and not other external elements, such as recording environment,

	Measure: **Volume**			Measure: **Pitch**			Measure: **Rate**		
	Median (IQR; Q_1, Q_3)	Range of scores	Spearman's *r* *(p)*	Median (IQR; Q_1, Q_3)	Range of scores	Spearman's *r* *(p)*	Median (IQR; Q_1, Q_3)	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 1: Summary of results for phase 1 study on the measures of volume, pitch variability and rate.

equipment, external noise, other vocal elements. Extending this into the phase 2 study, a direct comparison of the users own voice, collected in the same way each time, could alleviate some of these issues. It is also worth noting that the range of scores provided by the expert (60-120) and crowd workers (50-123) were similar, and there was a smaller range of scores for volume than the other measures. This question asked raters to think about the differences between the midrange sample and the sample being scored purely in terms of volume. It 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 midrange), with the reverse being stated for *less than 100* (less severe). We were considering that a low volume indicated impairment, however this was not explicitly stated to raters. As such, it is possible that they were rating on the lower end of the scale to indicate any difference in volume, where the experts may have had an internalized perception of impaired volume and how this might affect the speaker. There were several instances where the crowd rated the samples in the 60-85 category (less severe problem), where experts were rating 100-120 (more severe problem). It is possible that crowd workers were rating lower for lower volumes, while the experts were rating severity. In light of this, it was decided that the questions would be revised for phase 2, to ensure full transparency of what was being asked.

PHASE 2: IMPLEMENTATION OF SPEECHING

The Speeching system is made up of several components. The individual, who is using the system to self-monitor their speech issues, accesses an application (app) on their mobile phone. The app is used to collect a variety of speech samples through an assessment task. This task is then uploaded to the Speeching service, which packages the separate recordings into a 'job' for the crowd and uploads it to Crowdflower (chosen over AMT due to imposed financial restrictions in the UK). Five crowd members are requested to complete the job. They are asked to listen to and analyze two types of speech samples. When analyzing single words (n=10) the crowd worker is asked to select the word they have heard from a choice of 10 similar sounding words. When analyzing sentences (n=3) the crowd worker is asked to provide an overall rating of understandability and provide ratings on the volume, rate and pitch variability of the sample. The individual analyses are sent back to the Speeching service, aggregated and then the median score of the ratings is sent back to the user, through the app, as feedback on their speech performance. The user can then use this feedback to inform the areas of their speech that require practice, and can then use the app to conduct targeted exercises on their speech with the aim of improving their intelligibility.

The Speeching App

The Speeching app has three functions: to collect speech data for analysis by the crowd in the *assessment* area; to enable users to receive feedback on their speech; and to allow for the self-directed practice of speech issues common to Parkinson's (in the *practice* area of the app).

Assessment area

The assessment tasks prompted the elicitation of several types of speech sample (see Figure 1a for the types of task presented to the user). The first of these was unconnected speech, or single words (derived from Miller et al. [34]), which asks participants to read a word as it is presented on the screen. Users are asked to read 10 single words as they appear on the screen, recording each one individually by pressing a start/stop button. The second type of sample is connected, or sentence level, speech. This task requires users to either read a sentence as it appears on the screen; or describe a picture, or answer an open question in order to elicit free speech. In order to provide structure to this task, on-screen prompts are presented as scenarios, such as ordering a pizza or taking a bus ride. Subsequently, there is a combination of reading and free speech collected as users make their way through the scenario. Each scenario asks for two reading samples and one free speech sample. Again, each separate sentence is recorded individually using the start/stop button. Users are prompted each time they make a recording to hold the phone "one hand's distance away" from their mouth before speaking, to ensure a consistency of recording quality. Following the completion of the assessment, the 13 separate samples are packaged together into a 'job' and sent for analysis by 5 crowd workers.

Practice Area

In a separate tab is the *practice area* where users can access a daily task. These tasks are added for practice only and samples captured cannot be uploaded to the crowd for analysis, although users have the option to listen back to their sessions. We focused on two types of practice tasks,

improving loudness and improving rate, which along with pitch variance issues are the most common issues in Parkinson's speech [11]. In addition, previous research has noted the benefits of improving loudness for other areas of speech and voice, such as intonation, which is associated with pitch variance [21]. For both practice exercises a video tutorial from an expert in SLT and Parkinson's was created explaining why the exercise was being carried out and how it should be completed. In order to improve volume, users are asked to set a target, i.e. by counting to 10 in their loudest speaking voice, and attempt to maintain their volume level to an equivalent or higher volume while reading a segment of text on-screen. A numeric visualization of their decibel level is provided on the screen and green/red system is used to indicate when the user is above or below their target level, respectively.

The second practice task focuses on slowing rate of speech. In the first stage of the task users are presented with an auditory metronome and prompted to speak a word per beat, to begin getting used to slowing their speech down (e.g. WHAT-TIME-WILL-THE-TRAIN-BE-COME-ING). The metronome can be made faster or slower depending on a user's personal preference and skill level. Once this skill has been mastered, this task progresses towards using the metronome in a more naturalistic way, using natural intonation and stress patterns that would be seen in everyday speech. In this case the important words are spoken on the beat to add a natural stress pattern (what TIME will the TRAIN be COMING).

Integration with Crowdsourcing Services

The Speeching app was linked with Crowdflower, an online crowdsourcing service. A Speeching API was created in Microsoft C# consisting of a web service (ASP.NET Web API) that links the app and crowdsourcing platform together. Once an assessment is uploaded by the user it is posted to the Crowdflower site with a unique identifier code. Each job is assigned to 5 crowd workers for analysis. Ratings from the crowd are aggregated and the median score (to account for outliers) is delivered back to the user.

Micro-task design

Tasks were carried over from phase 1 with minimal changes. Single word samples were subjected to a selection task, with crowd workers being asked to select the word that they though the person was saying from a set of 10 similar words (e.g. sheep, keep, heap). For the sentence level data, the ease of listening (EOL) rating from phase 1 was carried out again and the measures of pitch, rate and volume were adapted from phase 1 by providing a comparative element for the crowd workers to use in their ratings. Rather than using a random mid-range example, we used the user's own speech as a comparative sample. In this case, when users upload their first assessment for analysis, crowd members are asked to rate speech, out of 100, for volume, rate and pitch variance (e.g. for the volume rating participants were asked; "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'"). However, in subsequent ratings the crowd workers are provided with the user's previous speech sample to listen to, and the median rating that this sample was given for each measure by the last group of crowd workers who rated it. This allowed for quality control within our own analysis, since crowd workers were given an exemplar 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.

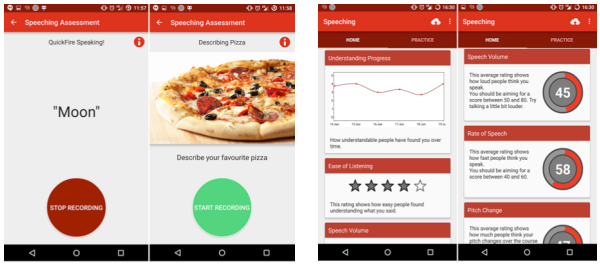

Figure 1: Screenshots for Speeching; a) Speeching assessment (left) and b) feedback screens (right)

Providing Feedback to Users

Within the Speeching app, users are provided with a graph of their EOL score over time, the EOL score of the sample they have just submitted, along with its volume, rate and pitch scores (Figure 1b). In order to allow users to make sense of the median data being presented back to them via the app, 'goals' are assigned to the measures to give them something to aim for and improve upon. Users are advised that both volume and pitch scores should fall within 50 and 90, while scores for rate should be between 40-60. These scores were chosen by the research team and expert to explore the impact that providing suggested goals might have on the participants. If participants scored below or above the threshold values, a prompt was added to their feedback to suggest how they should modify their speech.

REAL WORLD DEPLOYMENT WITH PWP

We trialed the Speeching app in a real world context with PwP. The purpose of this was to test our crowdsourcing approach on a group of PwP who could receive and react to feedback being generated by the crowd. Six PwP were recruited to take part, through local Parkinson's UK support groups following a presentation about the research aims. Participants of any age or stage of Parkinson's were considered for the study, so long as they reported issues with their speech. A profile of the individual participants and their main reported speech issues can be found in Table 2. Participants were visited by a member of the research team in their own home and given a smartphone with the Speeching app pre-installed on it. The researcher demonstrated how to use Speeching and participants were given an instruction manual bringing them through each step of the process for both the assessment and practice areas. They were asked to complete an assessment task

Name	Age	Years Since Diagnosis	Participant perception of speech severity	Main issues as reported by participant	No. Uploads	Mean range of pitch (SD)	Mean range of rate (SD)	Mean range of volume (SD)	Mean EOL with 1 being most severe (SD)
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)
Robert	61	11	Moderate	Volume	31	43.4 (18.8)	44.5 (19.0)	50.6 (21.6)	2.8 (1.3)
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)

Table 2: Speeching participant information and phase 2 quantitative results.

during the initial visit so that any issues with the app could be discussed with the researcher and a baseline measure of their speech could be collected. Participants were informed that they could not retry individual assessment items but that they had the choice whether or not to upload their session for crowd rating (this was to mirror traditional SLT assessment techniques, which often do not allow retries). Following this, they were instructed that they should receive feedback within 1 hour of completing a task. The researcher then helped them to navigate to the practice area and showed them the types of practice tasks that they could complete. Participants were asked to trial Speeching for one week, during which time they could use the app as little or often as they wished, though we requested that on at least one day they used the practice area and completed one other assessment before the end of the deployment. They were advised that they could upload their speech for analysis at any point during the deployment phase. Participants were additionally contacted via telephone at the midway point of the deployment to discuss any issues they might be having.

Following the deployment each participant took part in a semi-structured interview. Interviews lasted between 19 and 45 minutes (average 30 minutes) and included open questions on topics surrounding: their experiences of using the app over the week (frequency and ease of use, features they liked and disliked) and their opinions on the feedback from the crowd (if they found it useful, whether or not it motivated change, how they felt about being anonymously rated). Interviews were audio recorded and were transcribed verbatim for later analysis.

Quantitative data collected during the study included the number of tasks uploaded for analysis to the crowd each day of deployment, and the ratings that were provided by each crowd worker for each of the rated measures. There were 122 jobs in total uploaded to the crowd for analysis during the course of the study. A total of 6,306 ratings were completed by the crowd, comprising scores for volume, rate, pitch variance, EOL and single word recognition.

Overall, participants were varied in the amount that they used the app, with uploads ranging from 2-39 over the 7 days of deployment. A full breakdown of their individual engagement can be viewed in Table 1. This also details the different range of speech issues and severity across participants. As such, we looked at the data for each of the 6 participants individually.

Phase 2 Quantitative Analysis

Figure 3 shows a comparison of descriptive data for both Jerry (who rated himself as having severe speech difficulties and issues in multiple speech elements) and Jill, who had mild issues with her volume and voice clarity. For this stage of the analysis we were interested in extending our findings from phase 1 by exploring: a) how the word recognition task might be utilized in the future to inform therapy goals and b) the effect that receiving information about perceptual speech measures might have on facilitating home practice of speech. To explore the first question we constructed a confusion matrix for each participant, to visualize the error rate in the single word recognition (see Figure 3). This allowed us to look at the types of errors that were being made by speakers, as determined by the crowd's selections.

For the second question, we wanted to look at the speaker's scores over time and look at the extent to which the raters were providing similar scores for each measure. We took the mean range over speech samples (i.e. we computed the range covered by the 5 scores of each analyzed sample and then averaged all the ranges), as well as the mean standard deviation (SD) over samples (i.e. we computed the SD of the 5 scores of each sample and then averaged all the SDs). Table 1 shows the mean range and mean SD for each participant. This method was chosen due to the fact that each speech sample had the potential to be rated by 5 completely different raters at each point of submission. In addition, we had already established through our initial feasibility study that crowd workers could provide equivalent ratings, to experts in Parkinson's speech, in the measure of pitch, rate and EOL. However, we were required to test our theory that changing the recording procedure of the 'in the wild' data collection and the structure of the question for the raters would make a change to volume ratings. As such, we asked an expert to rate

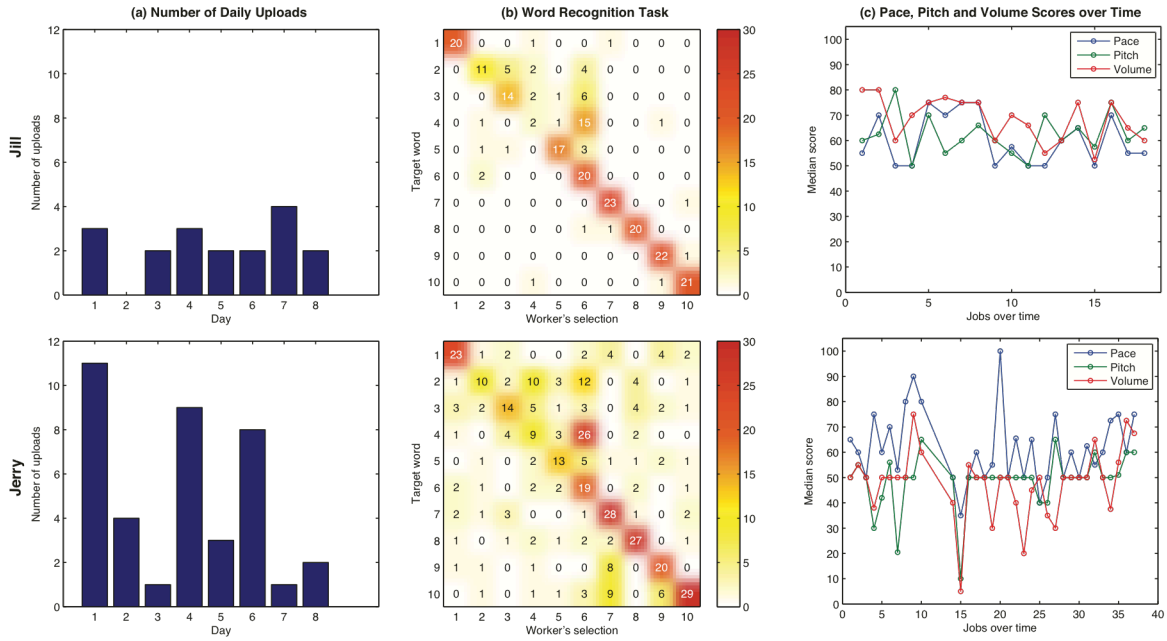

Figure 1: this figure provides comparative descriptive results for Jill (top) and Jerry (bottom); a) presents the number of daily uploads each participant provided over the course of the deployment; b) shows confusion matrices detailing the number of times that single words were recognized as either the correct target, or another word entirely (from 1-10 the words were, cape, carp, coop, cop, cub, cup, heap, keep sheep); c) shows the median scores for rate, pitch and volume presented to the participants following each upload in the deployment

volume on a small subsample of our entire data set, equaling 28 speech samples, and ran a correlation of this data and the median rating provided by the crowd. The 28 samples selected were randomly chosen to include 5 samples from each participant (2 of the selected samples did not contain audio were thus removed from the analysis).

Quantitative findings

Analysis of the word recognition task showed how more severe intelligibility issues (Damien and Jerry) were effectively identified by the crowd when compared to participants with milder speech impairment (Jill and Robert). This exercise was expressive enough to capture a variety of participant performance and provides useful direction for future work aimed at using tasks such as this to provide specified therapeutic direction for speech self-management. For example, Jerry's confusion matrix, shows many more instances of crowd workers identifying words that are different to the target. In the case of Jill, the bulk of errors stemmed around the misinterpretation of vowel contrasts (e.g. cup heard as cop 15 times) which could be an artifact of her accent, however Jerry, who had much more severe intelligibility issues, had a similar profile of errors, but was also experiencing word initial sound contrast difficulties (e.g. coop to hub 4 times, or sheep to heap 9 times). This indicates a more severe intelligibility difficulty which is indicative of Jerry's issues. He spoke at a very fast pace, and often ran out of breath, making it difficult to project his voice and position his articulators (e.g. tongue, palate, lips) into the correct position at times, which caused a slurred, imprecise quality to his speech. This suggests an opportunity for future functionality in the form of automatic provision of materials that target the repeated practice of these word initial sounds, with a view to improving his intelligibility (without the need for therapist input).

For the perceptual measure of pitch, volume and rate, participants displayed the highest range of scores within the measures that they perceived to be their biggest problem (see table 1). For example, Aaron had a higher range in his volume and rate scores, where Robert had the highest range in his volume scores.This is possibly due to the fact that the untrained listeners had more difficulty quantifying more severe problems with the speech. This is a problem which has been documented in previous studies such as Landa et al [27], who found that listeners can struggle to agree on speech ratings with increasing severity. Future research is required in order to scope this question further and draw out the best possible ways to train listeners to rate increasingly impaired speech. One possible solution worth exploring might be to draw much more attention to the measure being explored in isolation. For example, presenting the listener with a speech sample and asking them to focus only the volume in relation to a standardized tone (beep sounding at 60dB) which they increase or decrease to equal to volume in the speech sample equally; or asking a listener to draw a line to represent the speech sample, with increases and decreases in pitch being represented as peaks and troughs. Finally, following changes to research methods, a Pearson's Correlation Coefficient was used to explore the correlation between the experts and crowd on the measure of volume

and showed an *r= 0.57* indicating a moderate, almost high, positive correlation [20].

Engagement and Cost Analysis

A total of 86 crowd raters were included in the study, with an average of 8.9 jobs per rater. On average, it took 59 minutes to complete each job package for each set of 5 workers, from submission of the tasks by the participant to the provision of feedback. Crowd workers were paid an average of $0.42 per packaged job. With 5 workers per job this meant we paid a total of $2.10 per job, equaling a total of $256.20 spent on the 122 jobs submitted to the study. As a means of comparison, a specialist SLT in the UK is paid at approximately $109 per hour, which is the approximate amount of time taken to complete an assessment session with one client (not accounting for travel time).

Qualitative Findings

We conducted an inductive thematic analysis on the qualitative interview data using methods outlined by [8], by coding data at the sentence to paragraph level and drawing out themes across the data set. Three major themes were identified from the qualitative data; appreciation of the anonymous crowd; feedback and self-understanding; and problems with practicing and tasks. These are discussed in detail below.

Appreciation of the anonymous crowd

Participants responded well to considering crowdsourcing as a method of obtaining feedback about their speech. There was discussion around how people within their social networks are often not good markers of their ability. Damien compared the crowdsourced feedback to that he would normally receive from friends and neighbors: *"It was interesting to see how people rate you, because people don't usually tell you what they think"*. The app was valued in its capability to provide a sense of how speech was being perceived by others, without necessarily having to ask the question to friends and family. Robert echoed this sentiment: *"sometimes I just talk to people and they just look at me"*. He discussed the fact that gaining feedback about his speech from others can cause embarrassment and drew comfort from 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's 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."* Robert found the ratings from the crowd a motivator to 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

Most of the participants found the feedback features helpful as a means of understanding their speech and targeting improvements. Robert used the feedback from the crowd as a way to challenge himself to improve: *"I kept wanting to get to 5* [in EOL]. *And then speech volume, I wanted to increase that one, as well."* He also enjoyed the speed that he received his feedback *"getting it within, say, half an hour, an hour, is good…being so instantaneous"* Damien echoed the positive view that he saw the feedback as a *"challenge"*. His wife described the process Damien went through to improve his scores if the crowd rated him lower that his previous attempt: *"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"*. Due to having only limited Internet connectivity during the trial, Aaron only used the app minimally during his deployment. However, despite only using the app for at a couple of different time points, he did find that the feedback gave him insight on his speech rate *"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 that overall the app provided him with a way to monitor improvement *"this tells me that I can improve if I'm willing to change…Being reflective is enough for me"*.

While the feedback from the crowd was, for the most part, found helpful, three participants (Jill, Neil and Jerry) frequently used the listen back function within the practice area as a way to self-monitor their speech. For Jill, who was the most avid user of the function, she found it most useful for practicing and making changes to 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?"* Jill practiced particular elements of her speech which she felt were unclear, helping her to focus specifically on words or phrases that were affecting her clarity. For Neil, the listen back function gave him a tool for realizing and accepting 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"* Impaired volume perception is a common issue in Parkinson's speech [47], so supporting an increased understanding of how the voice actually sounds is particularly positive.

Problems with practicing and tasks

The two practice tasks, metronomic pacing and volume monitoring, were discussed at length by participants. Several issues were identified with these, particularly with the pacing exercise: *"He was going faster…He's way ahead of what the beeps were."* (Damien's wife). Robert and Neil similarly had difficulties: *"I didn't like the pacing... I understand it theoretically, but I can't do it practically"* (Robert). Robert also discussed the fact that he struggled to monitor his volume during the task due to the placement of the db level monitor at the bottom of the screen *"The text is here, and the green light's there. So you've got to try and concentrate."* There also discussion around how modifying the materials to be used within the practice exercises could increase motivation and improve engagement with the app. Aaron wished to use his own material to read, while Damien noted the scenarios were not relevant to him: *"I wouldn't get on the bus*". For Jerry, the scenarios were just too simple: *"it asks you stupid questions*". Robert and Jill however liked the scenarios due to their everyday nature *"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), although both reported that more variability in their content would be appreciated.

DISCUSSION

Crowdsourcing the Analysis of Impaired Speech

In our phase 1 feasibility study we demonstrated that anonymous crowd workers, recruited opportunistically via an online crowdsourcing platform could provide equivalent ratings on impaired speech to that of an expert. We additionally resolved the issues around volume by providing a consistent way of collecting speech data in the wild. Our findings also indicated that our Speeching system could prove useful within the area of speech diagnostics in the future. Future work of this kind might serve to leverage this diagnostic potential of the crowd through the careful restructuring of crowd tasks with SLTs, providing a cheap and abundant task force to aid in the diagnosis of speech and voice issues. In addition, further training of the crowd, and the implementation of binary selection tasks such as that used by Byun [10] could quickly and easily highlight areas of issue from voice collected in the wild. Although unrelated to crowdsourcing, relevant work conducted by Arora et al [2] has additionally studied the diagnostic potential of using automatic voice analysis on speech collected, over the phone, in the identification of undiagnosed Parkinson's. Supporting automatic diagnostic tools with therapeutic input provided by the crowd could greatly enhance the access to SLT level input, without the need for SLT resources. Considering that SLT uptake for PwP is though the be less than 40%, despite 90% of all individuals experiencing problems [38], digital technology could serve to fill a much needed therapeutic gap.

Trust and appreciation of the crowd

There was much appreciation for the fact that our crowdsourcing method employed real people to conduct the ratings. Participants used the crowdsourced ratings to gain insight into the ways that they were being understood by others and to achieve a baseline for themselves upon which to improve their speech upon. This, in itself, is a benefit for the Speeching system. Through the process of self-monitoring their speech, participants were able to engage more holistically in self-management practices. Without conducting a larger scale trial, it is unclear whether this method would be a motivator for a second group of participants, and indeed, what their reactions would be if their results increasingly worsened. This is a direction for future work, given that degeneration in ability is an almost inevitable concern for PwP. However, for those participants who are motivated in their rehabilitation efforts, using a method such as this could prove beneficial.

There was also a level of appreciation surrounding the anonymity of the crowd, which could absolve feelings of embarrassment surrounding speech and what others might think of it. Similarly, participants expressed how they felt this anonymity led to a more truthful measure of their speech, which could not be obtained from friends and family (who remain polite) or professionals (who are trained in listening to Parkinson's speech). This last point was a reason why a non-expert crowd was chosen in the first place, as the 'familiarity effect' has been widely researched in the past [27,39,56]. Aaron however had another option on this matter, feeling that his friends and family would be better objective raters of his speech as they would be "hard" on him. Although his ideas contrast with the other participant's views his opinion sits within a line of thinking around leveraging a person's social capital to help support sustainable systems within healthcare[46].

Despite participants' optimism with the Speeching system, there are several privacy and security concerns surrounding the crowd and their access to personal data uploaded by PwP. It's important to note this as a limitation of the current system, especially as Lasecki at al. [29] describe the vulnerabilities of crowdsourcing systems to unwanted information extraction and malicious manipulation. They suggest the design of workflows that leverage key reliable workers, to screen and alert the researchers to data which might be open to malicious attack. This is certainly an area which deserves further attention in the future. One possible solution could be to shift from making use of the anonymous crowd to one that is formed by connected individuals, within national charities and support groups, leveraging individual and collective capital. This would also have an added benefit around the resource implications of paying crowd workers (and indeed ongoing ethical questions over the economics and labour of crowd work [18]).

CONCLUSIONS

The work we 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 we have highlighted the validity of using a crowd as lay listeners and raters of Parksonian 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

ACKNOWLEDGEMENTS

This research was funded through the EPSRC Digital Economy Research Centre (EP/M023001/1). Data supporting this publication is not openly available due to ethical considerations. Access may be possible under appropriate agreement. Additional metadata record at 10.17634/141304 -1. Please contact Newcastle Research Data Service at rdm@ncl.ac.uk for further information or access requests.

REFERENCES

1. Melissa M Ahern and Michael S Hendryx. 2003. Social capital and trust in providers. *Social Science & Medicine* 57, 7: 1195–1203. http://doi.org/10.1016/S0277-9536(02)00494-X
2. S Arora, V Venkataraman, A Zhan, et al. 2015. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism & Related Disorders* 21, 6: 650–653. http://doi.org/10.1016/j.parkreldis.2015.02.026
3. Kartik Audhkhasi, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2011. Reliability-weighted acoustic model adaptation using crowd sourced transcriptions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3045–3048.
4. Julie Barlow, Chris Wright, Janice Sheasby, Andy Turner, and Jenny Hainsworth. 2002. Self-management approaches for people with chronic conditions: A review. *Patient Education and Counseling 48*, 177–187. http://doi.org/10.1016/S0738-3991(02)00032-0
5. Jeffrey P. Bigham, Richard E. Ladner, and Yevgen Borodin. 2011. The design of human-powered access technology. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '11*, ACM Press, 3. http://doi.org/10.1145/2049536.2049540
6. Jeffrey P. Bigham, Samual White, Tom Yeh, et al. 2010. VizWiz. *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*, ACM Press, 333. http://doi.org/10.1145/1866029.1866080
7. Riley Bove, Elizabeth Secor, Brian C Healy, et al. 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: e59707. http://doi.org/10.1371/journal.pone.0059707
8. V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3: 77–101. http://doi.org/10.1191/1478088706qp063oa
9. Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz. *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '12*, ACM Press, 135. http://doi.org/10.1145/2384916.2384941
10. Tara McAllister Byun, Peter F Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders* 53, 0: 70–83. http://doi.org/http://dx.doi.org/10.1016/j.jcomdis.2014.11.003
11. Gerald J. Canter. 1963. Speech Characteristics of Patients with Parkinson's Disease: I. Intensity, Pitch, and Duration. *Journal of Speech and Hearing Disorders* 28, 3: 221. http://doi.org/10.1044/jshd.2803.221
12. Anna C. Cavender, Daniel S. Otero, Jeffrey P. Bigham, and Richard E. Ladner. 2010. Asl-stem forum. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, ACM Press, 2075. http://doi.org/10.1145/1753326.1753642
13. Rumi Chunara, Vina Chhaya, Sunetra Bane, et al. 2012. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. *Malaria Journal* 11, 1: 43. http://doi.org/10.1186/1475-2875-11-43
14. Seth Cooper, Firas Khatib, Adrien Treuille, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–60. http://doi.org/10.1038/nature09304
15. Nicolas Côté. 2011. *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer Science & Business Media. Retrieved September 12, 2015 from https://books.google.com/books?id=-utLeUB2H34C&pgis=1
16. F Darley, A Aronson, and J Brown. 1969. Differential Diagnostic Patterns of Dysarthria. *Journal of Speech Language and Hearing Research* 12, 2: 246. http://doi.org/10.1044/jshr.1202.246
17. F Darley, A Aronson, and J Brown. 1975. *Motor speech disorders*. W.B. Saunders Company., Philadelphia, PA.
18. Julie McDonough Dolmaya. 2011. The ethics of crowdsourcing. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*. Retrieved September 25, 2015 from https://lans-tts.ua.ac.be/index.php/LANS-TTS/article/view/279
19. Keelan Evanini and Klaus Zechner. 2011. Using crowdsourcing to provide prosodic annotations for non-native speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3069–3072.
20. James Evans. 1996. *Straightforward statistics for the behavioral sciences*. Brooks/Cole Pub. Co., Pacific Grove.
21. C Fox, C Morrison, L Ramig, and S Shapir. 2002. Current Perspectives on the Lee Silverman Voice Treatment (LSVT) for Individuals With Idiopathic Parkinson Disease. *American Journal of Speech-Language Pathology* 11: 111–123.
22. Masataka Goto and Jun Ogata. 2011. PodCastle: Recent advances of a spoken document retrieval service

improved by anonymous user contributions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3073–3076.

23. John Green, Anne Forster, Sue Bogle, and John Young. 2002. Physiotherapy for patients with mobility problems more than 1 year after stroke: A randomised controlled trial. *Lancet* 359, 9302: 199–203. http://doi.org/10.1016/S0140-6736(02)07443-3
24. Aileen K. Ho, Robert Iansek, Caterina Marigliani, John L. Bradshaw, and Sandra Gates. 1998. Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology* 11: 131–137. http://doi.org/10.1155/1999/327643
25. R Holmes, J Oates, D Phyland, and A Hughes. 2000. Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders* 35, 3: 407–418. http://doi.org/10.1080/136828200410654
26. I Kawachi, B P Kennedy, and R Glass. 1999. Social capital and self-rated health: a contextual analysis. *American Journal of Public Health* 89, 8: 1187–1193. http://doi.org/10.2105/AJPH.89.8.1187
27. Sophie Landa, Lindsay Pennington, Nick Miller, Sheila Robson, Vicki Thompson, and Nick Steen. 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: 408–16. http://doi.org/10.3109/17549507.2014.927922
28. J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1: 159–174. http://doi.org/10.2307/2529310
29. Walter S. Lasecki, Jaime Teevan, and Ece Kamar. 2014. Information extraction and manipulation threats in crowd-powered systems. *17th ACM conference on Computer supported cooperative work & social computing*, 248–256. http://doi.org/10.1145/2531602.2531733
30. Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. 2010. Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, 99–107.
31. Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 5270–5273. http://doi.org/10.1109/ICASSP.2010.5494979
32. Ian McGraw, Alexander Gruenstein, and Andrew Sutherland. 2009. A self-labeling speech corpus: Collecting spoken words with an online educational game. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3031–3034.
33. J A McKenzie. 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* 39: 50–4. Retrieved September 21, 2015 from http://europepmc.org/abstract/med/1345506
34. Nick Miller, Liesl Allcock, Diana Jones, Emma Noble, Anthony J Hildreth, and David J Burn. 2007. Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *Journal of neurology, neurosurgery, and psychiatry* 78, 11: 1188–1190. http://doi.org/10.1136/jnnp.2006.110171
35. Nick Miller, Katherine H O Deane, Diana Jones, Emma Noble, and Catherine Gibb. 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* 46, 2: 189–201. http://doi.org/10.3109/13682822.2010.484849
36. Nick Miller, Emma Noble, Diana Jones, Liesl Allcock, and David J Burn. 2008. How do I sound to me? Perceived changes in communication in Parkinson's disease. *Clinical rehabilitation* 22, 1: 14–22. http://doi.org/10.1177/0269215507079096
37. Nick Miller, Emma Noble, Diana Jones, and David Burn. 2006. Life with communication changes in Parkinson's disease. *Age and Ageing* 35, 3: 235–239. http://doi.org/10.1093/ageing/afj053
38. Nick Miller, Emma Noble, Diana Jones, Katherine H O Deane, and Catherine Gibb. 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 / Royal College of Speech & Language Therapists* 46, 2: 179–188. http://doi.org/10.3109/13682822.2010.484850
39. Nick Miller. 2013. Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders 48*, 601–612. http://doi.org/10.1111/1460-6984.12061
40. Tan B Nguyen, Shijun Wang, Vishal Anugu, et al. 2012. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT

colonography. *Radiology* 262, 3: 824–33. http://doi.org/10.1148/radiol.11110938

41. M J Nijkrake, S H J Keus, J G Kalf, et al. 2007. Allied health care interventions and complementary therapies in Parkinson's disease. *Parkinsonism & related disorders* 13 Suppl 3: S488–S494. http://doi.org/10.1016/S1353-8020(08)70054-3
42. Francisco Nunes and Geraldine Fitzpatrick. 2015. Self-care technologies and collaboration. *International Journal of Human-Computer Interaction*: 150730080814008. http://doi.org/10.1080/10447318.2015.1067498
43. Gabriel Parent and Maxine Eskenazi. 2011. Speaking to the Crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3037–3040.
44. Patientslikeme. 2015. Live Better, Together! Retrieved from https://www.patientslikeme.com/
45. Megan Perry, Robert L Williams, Nina Wallerstein, and Howard Waitzkin. 2008. Social capital and health care experiences among low-income individuals. *American journal of public health* 98, 2: 330–6. http://doi.org/10.2105/AJPH.2006.086306
46. R Putnam. 2001. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster. Retrieved September 25, 2015 from http://bowlingalone.com/
47. L O Ramig, S Sapir, S Countryman, et al. 2001. *Intensive voice treatment (LSVT) for patients with Parkinson's disease: a 2 year follow up.* http://doi.org/10.1136/jnnp.71.4.493
48. M Swan, K Hathaway, C Hogg, R McCauley, and A Vollrath. 2010. Citizen science genomics as a model for crowdsourced preventive medicine research. *J Participat Med* 2: e20. Retrieved from http://www.jopm.org/evidence/research/2010/12/23/citizen-science-genomics-as-a-model-for-crowdsourced-preventive-medicine-research
49. M Swan. 2012. Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen. *Journal of personalized medicine* 2, 3: 93–118. http://doi.org/10.3390/jpm2030093
50. Kris Tjaden. 2008. Speech and Swallowing in Parkinson's Disease. *Topics in geriatric rehabilitation* 24, 2: 115–126. http://doi.org/10.1097/01.TGR.0000318899.87690.44
51. Gary Weismer and Jacqueline S Laures. 2002. Direct magnitude estimates of speech intelligibility in dysarthria: effects of a chosen standard. *Journal of speech, language, and hearing research : JSLHR* 45, 3: 421–433. http://doi.org/10.1044/1092-4388(2002/033)
52. Paul Wicks, Dorothy L Keininger, Michael P Massagli, et al. 2012. Perceived benefits of sharing health data between people with epilepsy on an online platform. *Epilepsy & behavior : E&B* 23, 1: 16–23. http://doi.org/10.1016/j.yebeh.2011.09.026
53. Sheila Wight and Nick Miller. 2015. Lee Silverman Voice Treatment for people with Parkinson's: audit of outcomes in a routine clinic. *International journal of language & communication disorders / Royal College of Speech & Language Therapists* 50, 2: 215–25. http://doi.org/10.1111/1460-6984.12132
54. Maria K. Wolters, Karl B. Isaac, and Steve Renals. 2011. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. Retrieved August 27, 2015 from https://www.era.lib.ed.ac.uk/handle/1842/4660
55. Xian-Hong Xiang, Xiao-Yu Huang, Xiao-Ling Zhang, Chun-Fang Cai, Jian-Yong Yang, and Lei Li. 2014. Many Can Work Better than the Best: Diagnosing with Medical Images via Crowdsourcing. *Entropy* 16, 7: 3866–3877. http://doi.org/10.3390/e16073866
56. Wolfram Ziegler and Andreas Zierdt. 2008. Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders*. http://doi.org/10.1016/j.jcomdis.2008.05.001
57. Wolfram Ziegler and Andreas Zierdt. 2008. Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders* 41, 6: 553–577. http://doi.org/10.1016/j.jcomdis.2008.05.001
58. Crowdmed. Retrieved from https://www.crowdmed.com/