Assessing privacy-friendly local open-source voice annotation for participants with Parkinson's disease

Emma L. Tonkin[1][0000-0001-7405-4982] and Gregory J. L. Tourte[2][0000-0002-2819-392X]

[1] Digital Health, University of Bristol, UK
e.l.tonkin@bristol.ac.uk
[2] Advanced Research Computing, University of Oxford, UK
gregory.tourte@it.ox.ac.uk

Abstract There is significant potential clinical benefit to be gained in capturing symptom data from individuals with Parkinson's Disease (PD). For this purpose, sensor data is often collected. However, labels (ground truth) data is also beneficial, both to train (supervised learning) and to validate outcomes from automated monitoring systems. With the increasing use of voice assistants, this modality has been proposed for labelling. In this study, we examine some design patterns for voice-agent-supported labelling, identify failure modes, and make use of the MDVR-KCL dataset to benchmark a widely used key component, a speech-to-text pipeline. We identify that this component shows rapid increase in several error metrics (WER, CER, WIL) when employed on data from mildly symptomatic participants. We identify some potential mitigating steps and discuss potential future work.

Keywords: Parkinson's Disease · Privacy-first engineering · Voice assistants.

1 Introduction

The increasing use of in-home technology for assistance, support and monitoring, such as for the aged and for those living with conditions such as Parkinson's or Alzheimer's, has led to an ongoing need for training and validation data to ensure that these systems are functioning as planned. For this purpose, a broad swathe of technologies may be used, ranging from paper diaries, tablet, phone or wearable based questionnaires or voice assistants, through to post-hoc annotation of identifiable data such as video or audio. It has also led to ongoing concern about the potential costs and risks surrounding the externalisation of this data and its processing from the home, since a great deal of the data collected is likely to be identifying or contain re-linkable features. In recent years, the problem of privacy-preserving and decentralised approaches to data processing with machine learning (ML) has received a great deal of attention. For example, federated learning is often proposed, in which models are trained in-home and in principle

may be tuned by comparison to other models by centralising operations on model weights, rather than raw data and labels. Since the raw data and features do not leave the home, the residual risk of this data sharing is solely that attached to the model weights, and is consequentially lower.

In this paper, we examine a subset of the design challenges of building an affordable, auditable voice assistant suitable for supporting clinical trials within the home, with a particular focus on participants with Parkinson's Disease (PD). Our aim is to assess the current technology readiness of a key component widely used for open source voice assistants. Ultimately, we restrict our focus to privacy-friendly, decentralised approach: for the present study, our interest is in voice assistant implementations in which data does not leave the home (that is, in which audio is processed via on-device processing entirely within the home).

The paper is structured as follows: we begin by characterising PD, discuss a number of potential uses of data annotation/labelling, and review potential effects of the symptoms of PD on the individual's voice and hence on the technologies. We then discuss scenarios of use for a voice assistant suitable for data annotation purposes, drawn from individual and collaborative ideation and refined using a design-fiction approach to develop scenarios. Our methodology is to make use of an existing dataset created by King's College London (KCL) to characterise the performance of standard components of a voice assistant speech-to-text pipeline of changes in speech or voice due to PD. This enables us to establish the effects of performance limitations on the system requirements. In our discussion, we discuss the methodological limitations of this study and explore further work, discuss key findings, and briefly review potential mitigations of these issues.

2 Characterising PD

PD is a neurodegenerative disorder. Motor function is impaired, resulting in symptoms such as slowness of motion or decrease in amplitude of motion (bradykinesia), tremor, and disturbance to gait and balance [40]. The patient is likely to experience nonmotor symptoms such as pain [38], and individual patients report finding a range of symptoms particularly troublesome. Nonmotor symptoms such as abnormalities of sensation, behavioural changes and sleep disturbance are also important components of PD [37].

People with Parkinson disease are very likely to have identified, or had others notice, changes in their speech or voice [16]. In particular, Holmes *et al.* [22] identified that people with PD are likely to speak with lower intensity (quieter), have a breathier voice, and display reduced variation in pitch and loudness compared to a control group. People with longer disease duration are more likely to experience higher magnitude of frequency tremor [20]. However, it is important to be aware that the experiences of people with PD vary widely [12]. There is also evidence that voice self-assessment is complex for people with PD compared to individuals with general voice disorders [8]. Hence, in this study, we have chosen to work with a dataset that is labelled via expert assessment.

PD symptoms are likely to fluctuate [39]. This occurs in the short term, as in the case of motor fluctuations such as freezing – a sudden and temporary inability to move – and paradoxic kinesis [15], in which fluid motion is briefly exhibited, potentially as a consequence of external cueing [28]. Within the day, there is fluctuation – for example, in gait – potentially due to factors such as depletion of medication [35]. Additionally, the Hawthorne effect, change in performance due to the awareness that an activity is observed, is likely to be a factor in short-term observation of people with PD [33].

The variance described in the previous paragraph adds complexity to 'snapshot-based' evaluation of PD, meaning that ratings such as the Unified Parkinson's Disease Rating Scale (UPDRS) have difficulty capturing the full picture [21]. As a consequence of this complex picture, there is considerable interest in making use of mobile, pervasive and wearable sensors to characterise the ongoing activities of people with PD, and in particular, of the symptoms that they experience. This is perceived as likely to enable a fuller picture of the ebb and flow of symptoms throughout the day and over time, facilitating evaluation of interventions such as medication or other therapies.

2.1 Data labelling in PD

In this section, we briefly discuss the challenge of data labelling in PD.

Existing literature shows a wide range of potential approaches to data annotation in PD. One popular approach is app-based scripted data collection. For example, the mPower dataset [5] made use of a smartphone-based app strategy to implement and record a series of scripted activities, including participants with and without PD, with the intention of quantifying the fluctuation in PD symptoms. The activities included were well-chosen for an app-based delivery mechanism and for the available sensors, and included a memory task, a 'tapping' task to assess dexterity, a voice-based task to assess sustained phonation, and a walking activity involving walking 20 steps in a straight line, turning, standing still and walking back. However, the IMU data did not receive ground truth labelling within this task, possibly because there was no convenient mechanism to achieve this without disrupting the task itself. Similarly, Borzì *et al.* [4] describe the use of mobile phone sensors to record a scripted activity, with data recording started and stopped after each case. As Little [26] summarises the problem: labelling often requires expert decisions, and hence requires care, potentially training for the assessor, assessment of inter-rater variability, and a high degree of ongoing adherence, which may become a significant burden for the labeller. In-situ labelling also has the potential to disrupt activities, adding a further burden to the participant: on a very basic level, the need to carry a device with one and interact with it periodically is itself a burden.

By contrast, Morgan *et al.* [32] present a taxonomy consisting of a series of features in PD that may be of interest for annotation, including: activity level and intensity (walking, sitting, lying down), activities of daily living (e.g. watching television, food prep, cleaning, chatting,…), global spontaneity of movement (slowness), gait (from unproblematic independent walking to requiring assistance

or unable to walk), and level of impairment in the activity of sit-to-stand, going from a seated position to standing up. These are then labelled using RGB video data taken from within the home, by a medically trained specialist with support from other annotators. Such an approach has challenges, notably those related to the limited field of view of a camera: the body will tend to occlude some features, especially if the individual is not well placed in front of the camera, and especially when the activity takes place in a small space.

2.2 The challenge of voice annotation in PD

Partially in response to the challenges of other methodologies, it has been proposed that voice agents provide a potential solution for data labelling in the wild in the general case, whether via automated assessment of audio cues [18], via semi-automated means through interaction with a voice agent [10] or via straightforward voice-based logging [42]. Audio recording, however, is a privacy-intrusive approach, and hence there is the potential that at least some demographics of participant will react negatively to this as a consequence of privacy concerns, requiring appropriate mitigations to be put in place to safeguard privacy [17].

From a privacy perspective, voice annotation poses challenges. As Germanos *et al.* [19] indicate, typical voice assistants coexist with a wider ecosystem, in which a large quantity of data, some of which is classifiable as personal data, is either temporarily or permanently stored outside the user's immediate physical context. In particular, commercial voice assistants are increasingly moving toward LLM-based approaches and cloud architectures rather than on-device processing, a trend encapsulated by the recent decision by Amazon to discontinue the 'Do not send voice recording' option, meaning that recordings that may previously have been retained locally are now sent to remote processing venues. Where medical information is annotated and logged, this may be a greater concern than for other applications of this technology. Hence, in this paper, we focus on the challenge of voice annotation using technology situated within the home.

The use of voice assistants for data labelling in contexts such as human activity recognition is not widely explored, compared to the use of data labelling on voice commands (in order to improve performance), which is common place, or, in the case of PD, the analysis of voice recordings to directly assess the progression of PD symptoms. However, the designer is led to consider the possibility by observing, firstly, that voice assistants or agents are currently widespread and widely used, including amongst older adults [2], for purposes such as setting up reminders, weather information and search; secondly, that the technology is reasonably mature; thirdly, that voice agents are well-placed for use where the hands and eyes are busy [23]; and finally, that there is also some anecdotal evidence that the practice of making use of a voice agent may lead to improvements in the speech of people with PD [16]. A key question for our purposes is therefore whether the technology as implemented in the home is adequately mature to support this use case, and to what extent.

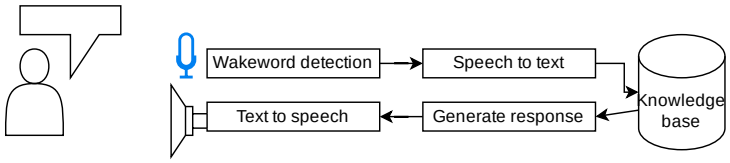


Figure 1. Simplified architecture for a voice agent

In the following section, we discuss generic interaction scenarios that are envisaged to be relevant to supporting voice annotation. These are included as a lens through which to consider the required quality of voice transcription (speech-to-text). In this paper, we primarily consider participant-initiated information logging. We acknowledge that more complex patterns may mitigate many quality concerns. Additionally, patterns that primarily involve information presented by the voice assistant are likely to be useful in many annotation scenarios, such as prompting the participant to engage in scripted activities.

3 Samples of design patterns for voice annotation

The methodology used for collecting scenarios was as follows: scenarios of use for voice annotation in PD are elicited from several sources, including: literature review; individual and group ideation; symptom-led discussion; clinical requirements for study data; and a design-fiction approach, using design stories to elicit and explore potential uses and accompanying concerns and risks. This last approach may be used informally in technical teams to communicate functionality and illustrate perceptions and expectations regarding factors such as user acceptance, functional constraints and practical risks. While these scenarios will not be reported in full in this workshop paper, as they are beyond the scope of this work, we provide a summary of the underlying design patterns here alongside an illustrative example for each case. By design patterns, we refer to generic, reusable patterns of dyadic [27] interaction [14] that can take place between the participant and the voice agent. These generic capabilities define the base functional requirements for a voice assistant framework in this context.

3.1 Pattern 1: Participant-initiated information logging

Example: Recording events and activities without system prompting.

> *Human speaker:* Assistant name/Wakeword, I'm taking my medication.
> *Voice assistant (VA):* Noted, you took your medication at nine fifty three AM. I'll add an entry to the log.

This type of speech act is described by Mahmood *et al.* [27] as a statement. The VA's response indicates that the relevant action is taken and that the interaction is at an end.

3.2 Pattern 2: Device-initiated query-response

Example: Activity recognition validation.

> *Voice assistant [triggered by ML device prediction]:* Hey, [participant name], are you cooking something?
> *Human speaker:* Yes, I'm making lunch. It's sardines on toast.

This is also potentially useful for many other types of validation task, such as re-identification, quality of motion and so forth. An **informational device-initiated announcement** is also possible that does not expect any kind of response: for example, a voice assistant announcing that there is a person at the door, that post has been received or that a device is low on battery and needs recharging. However, in the context of data annotation it is likely that dyadic interactions with confirmation will be preferred where possible, principally because the purpose is usually to elicit information from the participant, and secondly because this is likely to lead to more robust communication, in that it is more likely for communication failures to be identified and potentially repaired.

3.3 Pattern 3: Participant-led longer unscripted interaction or continued conversations

Example: participant chats to device

3.4 Pattern 4: Lengthier scripted interactions, such as completion of standard movement scripted activities

Example: Device leads participant through several tasks that provide adequate proxy data for estimating standardised results such as UPDRS score, or device talks participant through standardised instruments such as the Pittsburgh Sleep Quality Index (PSQI) test.

3.5 Failure pattern in participant-initiated interaction: system does not recognise that interaction is occurring

In some cases the system wake word is not recognised, resulting in the voice assistant failing to process the following statement at all.

It is also possible in device-initiated interactions that the participant will not recognise that the sound they are hearing is the voice assistant starting an interaction. Mitigating patterns that are often suggested for this are visual and audible indicators preceding the utterance to clearly link the utterance with the source.

3.6 Failure patterns: system misunderstands the participant, or vice versa

This is a key problem in voice assistant design. Appropriate recovery strategies are an issue.

3.7 Key concerns in voice assistant design

While the present study essentially examines feasibility and practicality by evaluating the performance of a key component, there are further major areas of concern for us as implementers. The first is usability and user acceptance. For example, according to Mahmood *et al.* [27], lag time in response is a significant concern. Voice assistants frequently implement a 'wait pattern' in the event that queries will take more than a couple of seconds. Similarly, a voice assistant with a high error rate is likely to frustrate the user, either because it does not detect that an error has occurred or, to a lesser extent, if it overcorrects [11]. Implementing repair in a domestic voice assistant is beyond the scope of this study, which aims only to explore the likelihood that such errors occur: however, we will touch on this topic again in the discussion of this paper. In general, voice assistants may be viewed as intrusive [36] in privacy terms, and this may affect user acceptance. Some of the design patterns mentioned above are potentially disruptive in nature, which, Jamshed, Nurain and Brewer [25] observes, may be beneficial depending on the precise purpose of use. For example, as Zargham *et al.* [43] find, proactivity in voice assistants is preferable during opportune moments, which implies a certain level of context-awareness in system design. Perceived appropriateness is of importance: however, in that study, participant opinions of proactive interventions vary between participants, and there is a negative correlation between perceptions of usefulness and invasiveness. A further system feature not examined within this study is identification of the speaker: speaker identification is a requested feature for home-deployable systems. Again, a detailed discussion of these concerns is beyond the scope of this study: however, it is useful to recognise that reduced-quality voice detection, keyword detection and transcription may have a significant effect on the ability of a system to achieve the desired design and user acceptance.

The second key subject area for implementers is security and privacy. Although the technical and organisational concerns of voice assistants are within the scope of our broader study, detailed discussion is beyond the scope of this paper. Perhaps the most commonly discussed issue with voice assistant implementations is the potential to retain data for system improvement purposes [7], for example, finetuning [30]: there is genuine potential benefit for system users in doing so, yet the retention of the data, even if appropriate technical and organisational measures are taken to protect it, raises significant risks for participant privacy. Cheng and Roedig [6] highlights many concerns with potential uses of participant data and the need to ensure transparency, such as participant awareness of recording, the destinations of recordings, and the potential for one's voice (and data) being picked up by others' devices, as well as the possibility for detection of activities (e.g. laughter, crying, or eating, or indeed medical state, such as a cold [1]), room characteristics, or even of active audio sensing of room features, occupancy and so on. Technologies such as speaker recognition offer the potential of access control on data input and system control, yet these are subject to attacks such as speech synthesis based on existing samples of data, known as replay attacks[13]. Furthermore, some uses of voice-based annotation for digital health purposes,

for participants in multi-occupancy homes, for example living with other family members or friends, have the effect of exposing others in the home to healthcare data: this may not be desired by participants, for example because there may be tensions around these topics, or participants may, as Binda *et al.* [3] describes, be concerned about causing other family members to worry. Voice assistant use in these contexts may limit participant control over disclosure, and hence it may be beneficial to provide parallel solutions, such as an app, according to participant preference. Crotty *et al.* [9] find that information sharing practices for older people are often fluid, designed to maximise autonomy, and that preferences vary significantly. To summarise, data from voice recordings has the potential for considerable abuse and information leakage, and the appropriateness of spoken interaction varies significantly between participants, so practical implementations of this kind require effective technical and organisational measures to be taken throughout the design, implementation and deployment processes, as well as involvement of stakeholder groups (for example, through co-design) to assess and address concerns.

4 Methodology: assessing performance of voice assistant software

The key research question on which we focus in evaluating the performance of voice assistant software for this purpose is: to what extent can we expect voice assistant software performance to be affected by PD symptoms? For this purpose, we focus on the performance of the key speech-to-text component (depicted in fig. 1) – that is, the component that takes arbitrary speech input and attempts to convert it to text in order to extract commands, store information, or take other action as described above.

The other key component that may be affected is the problem of *wakeword detection* – that is, the system's ability to recognise when it is being addressed. This is also an important problem, particularly since wakeword detection often takes place on the edge – which is to say, on small machines such as ESP32 microprocessors, which have severe technical limitations. Wakeword failure means that the system will fail to recognise that it is being addressed, or alternatively that the system will wake inappropriately (false positive). We acknowledge that wakeword detection is also an important area, but it is out of scope for the present study.

For the purpose of evaluating the performance of speech-to-text in this context, we make use of the KCL MDVR-KCL dataset [24], Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. This dataset was chosen on several grounds, including accessibility for reuse, appropriate licensing, compatibility with the broad aims of the original dataset, and clear and relevant data annotation, in that the dataset was annotated by subject matter experts for participant UPDRS scores related to speech characteristics (see discussion of voice self-assessment in section 2). The MDVR-KCL dataset is recorded using a mobile phone rather than

(a) (b)

Figure 2. Example of home deployable voice assistant hardware compatible with Home Assistant (a) Nabu-Casa Home Assistant Voice Preview Edition and (b) Espressif ESP-32 S3 Box-3

high-quality microphones, which is somewhat analogous to the likely conditions for use of home-deployed voice assistant hardware (see fig. 2).

This dataset contains voice recordings from sixteen participants with PD, and 21 healthy control (HC) participants. The dataset contains both scripted speech and unscripted spontaneous dialogue: for ease of comparison within this initial study, we focus on the scripted data. Each participant is asked to read at least one of two readings, 'The North Wind and the Sun' and 'Tech. Engin. Computer applications in geography snippet'. Each file is annotated with a pseudonym, a health status label (PD or control), a Hoehn and Yahr (H&Y) scale rating, a UPDRS II-5 (expert peer-reviewed) score and a UPDRS III-18 (expert assessed score). To interpret the latter two [41], the UPDRS II-5 assesses speech within 'activities of daily living', and rates speech between 0 (Normal), 1 (Mildly affected, no difficulty being understood), 2 (Moderately affected. Sometimes asked to repeat statements), 3 (Severely affected. Frequently asked to repeat statements) and 4 (Unintelligible most of the time). The UPDRS III-18 score assesses speech within the motor examination, and assesses speech on the following scale: 0 (Normal), 1 (Slight loss of expression, diction and/or volume), 2 (Monotone, slurred but understandable; moderately impaired), 3 (Marked impairment, difficult to understand) and 4 (Unintelligible). The H&Y scale is a system for grading severity of PD symptoms, as follows: [31]: 1) minimal or no functional disability, 2) bilateral involvement, without impairment of balance, 3) mild to moderate bilateral disease, with some postural instability, 4) severely disabling, still able to walk or stand unassisted, 5) wheelchair bound or bedridden unless aided.

In this section, we test the performance of speech-to-text transcription on these files, as this is a key component of voice assistants in general. We develop a ground truth for these audio files. We calculate word error rate and processing time based on a standard setup. We compare several models chosen from OpenAI's Whisper. In this manner we aim to assess the technology readiness of this component to this use case.

4.1 Developing a ground truth

Since the audio files contain extraneous audio before and after the participants' readings of the texts, the first author made note of the times in which individuals other than the participant were speaking and stored these for automated cropping. Following this process, the audio files were manually transcribed by the first author. Attention was paid to accurate transcription of the participants' speech. The second author then reviewed each of these transcriptions for accuracy, resulting in a consensus ground truth. These timings and transcriptions are intended for eventual publication, as we feel that they form a useful resource alongside the existing dataset.

4.2 Scoring speech-to-text output with word, character and match error rates

Word error rate (WER) and character error rate (CER) are standard metrics by which to measure the performance of speech recognition systems. Similarly to the Levenshtein edit distance algorithm, the goal of WER and CER are to establish the minimum edit distance between the ground truth string and the hypothesis (e.g. the speech to text system's 'guess' at the correct answer). This can be understood by the following equation:

$$\text{error\_rate} = \frac{I + D + S}{N_{ER}} \tag{1}$$

where I is the number of insertions, D the number of deletions, and S the number of substitutions required to generate the hypothesis from the ground truth string, H is the number of correct matches and N is the overall number of terms: in the case that word error rate is calculated, N is the number of words in the string, while where character error rate is calculated N_{ER} is the number of characters $(H + S + D)$. In other words, word error rate is the ratio of the number of errors to the number of words provided [34].

As an intuitive explanation, comparing a ground truth string of 'This is a fact' to a version, 'This is fact', we note that one word has been omitted. One word must be deleted from the ground truth in order to achieve the hypothesis text. If our hypothesis text has an additional word, e.g. 'This is indeed a fact', one word must be added to generate the hypothesis text. Hence, two identical strings have an edit distance of zero. The larger the *WER* or *CER*, less accurate the system.

$$WIL = 1 - \frac{H}{N} \times \frac{H}{P} \tag{2}$$

The Word Information Lost (WIL) metric attempts to approximate which proportion of word information is unsuccessfully transmitted, quantifying the proportion of information lost. This differs from *WER* in that *WER* calculates the cost of reconstructing the input. In the above equation, P is the number of words in the automated transcript.

There are known shortcomings to the use of these metrics for evaluating voice systems, principally that they do not accurately reflect human perception of the accuracy of such systems [29]. That is to say, technical error and perception of error are likely to differ, according to the area of application in which the technology is used.

4.3 Processing time

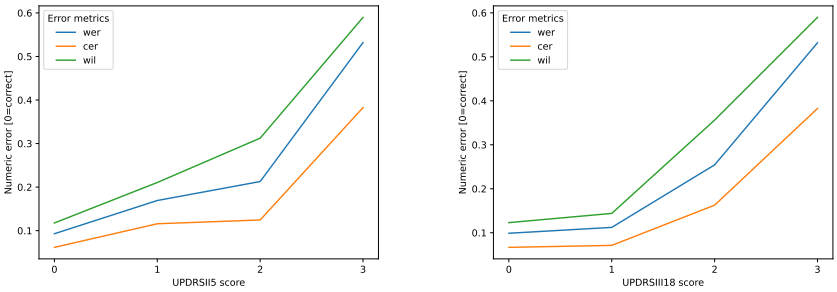
When running these models, we have collected processing time as a loose indicator of relative requirements in terms of processing power and time. We stress that this does not signify that these models are necessarily slow or that their deployment will involve lag. It is clear that, in practice, deployment of a larger model with more parameters will involve making and evaluating suitable hardware choices prior. Rather, we wish to highlight that deployment cost of the larger models is accordingly higher, and potentially that there is less pragmatic likelihood of running these models in a low-power home environment context, such as that indicated in a private federated learning context. We collect this information alongside system performance scoring data in order to establish whether there is clear benefit in deploying the larger models, in this specific user context.

5 Results

5.1 Error in voice annotation by UPDRS score

Unsurprisingly, direct comparison of error metrics between the PD and HC groupings demonstrates that average performance of models is notably worse in the former group ($\overline{WER_{PD}} = 0.16$, $\overline{WER_{HC}} = 0.1$; $\overline{CER_{PD}} = 0.11$, $\overline{CER_{HC}} = 0.07$; $\overline{WIL_{PD}} = 0.2$, and $\overline{WIL_{HC}} = 0.12$).

A more nuanced picture can be seen by examining the performance of the system relative to participant UPDRS scores (see fig. 3). As the figure shows, the speech-to-text system has difficulties dealing with 'moderately affected' (level 2) speech in UPDRS II 5, and significant difficulties with severely affected speech. It also displays increased error on speech scored as 'mildly affected' under UPDRS-II 5. Concretely, WER increases from 0.09 for those with a UPDRS-II 5 score of 0 to 0.17 for those with a score of 1, while WIL almost doubles (0.12 and 0.21 respectively). This is interesting, in comparison with the experience of the annotators participating in this project, who found that as human listeners, speech



(a) (b)

Figure 3. Performance of text-to-speech segmented by Unified Parkinson's Disease Rating Scale (a) UPDRS-II 5 score and (b) UPDRS-III 18 score

scored in this way was as comprehensible as that of control participants. This suggests that the speech-to-text system may be somewhat susceptible to errors as a result of features of mildly affected speech. A UPDRS-III 18 score of 1 has relatively little effect: the WER increases only very slightly, as does CER and WIL. However, a UPDRS-III 18 score of 2 has a significant effect (WER increases from the range of 0.1 to 0.2 in the case of a score of 0 or 1 to 0.25, whilst WIL increases from 0.12 to 0.14 [UPDRS-III 18 score of 0 or 1 respectively] to 0.36. A score of 3 has a similar effect, with WER and WIL exceeding 0.5. Hence, moderate impairment of speech on the motor examination aspect of UPDRS may be expected to significantly impair performance of speech-to-text systems of this kind.

It is worth noting that H&Y scores of 0 and 1 are very similar in performance: indeed, we find that the performance improves very slightly between the two, an effect that we attribute to variation in demographic and selection strategies between the groups. However, samples with a H&Y score of 3 experience approximately double the error rate to the baseline, with samples with H&Y of 4 seeing error rates of around 4 to 5 times the baseline on average depending on metric.

Perhaps the most significant point to take from this analysis is that performance of speech-to-text systems appears to degrade faster than we might expect, given the comprehensibility of voice to the human listener. Informally, it is worth noting that many of the participants in this study, especially in the HC (control) group, have accents that the annotators viewed as regional or international, and while this has not been systematically encoded in the study, it is notable that these accents do not seem to have had similar impact on system performance, potentially implying that the speech-to-text systems used are more resilient to accent variation than to PD symptoms.

Table 1. Performance of Whisper models in control (HC) and PD groups

Model	Parameter Count	Overall Aggregate WIL	Performance by Group					
			PD WER	HC WER	PD CER	HC CER	PD WIL	HC WIL
tiny	39 000 000	0.206	0.197	0.093	0.107	0.042	0.283	0.147
base	74 000 000	0.156	0.142	0.078	0.080	0.038	0.203	0.120
small	244 000 000	**0.104**	**0.105**	**0.048**	**0.065**	0.030	**0.149**	**0.069**
medium	769 000 000	0.108	0.110	0.051	0.073	**0.029**	0.151	0.075
turbo	809 000 000	0.183	0.256	0.151	0.203	0.123	0.234	0.145
large	1 550 000 000	0.194	0.154	0.172	0.111	0.138	0.196	0.193

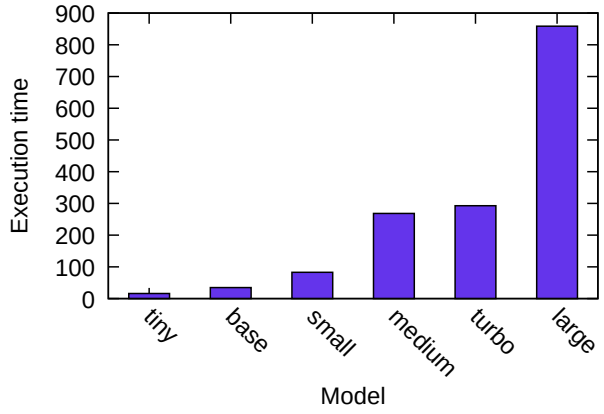


Figure 4. Model average processing time across all segments (mean)

5.2 Model performance in each group

Model performance in each group is summarised in table 1. The best-performing by each metric in the healthy control group (HC) and the PD group are highlighted in bold. See the following subsection for further discussion.

5.3 Processing time

As can be seen in fig. 4, the processing time of the larger models is very significant (e.g. efficient use of these models would require specialised hardware, which likely involves cloud processing). Fortunately from a privacy perspective, there is little apparent benefit of using the larger models in this specific context of use. There is anecdotal evidence that in general voice assistant implementers find the Whisper Base model to be an adequate compromise between execution speed/ hardware requirements and accuracy. However, given the increased error rate and information loss identified in this study for participants with PD, it is likely

that the 'small' model would be the better compromise in this population, as this would provide, for example, a WER only slightly higher than the base model would produce for input from the control group. In effect, the average WIL experienced for individuals with PD where a 'small' model is used is equivalent to that experienced by individuals in the control group using the 'tiny' model. However, there are several important caveats to this, which are discussed in the following section.

6 Discussion

There are some methodological limitations to this study. Principally, the KCL dataset does not provide demographic data, and therefore it is not possible to comment on other factors that may confound the findings. For example, the age range of participants in each cohort is not provided, nor is information provided about screening regarding other diagnoses or confounding factors, other than UPDRS/H&Y scoring of the HC participants. Secondly, we have excluded from the scope of the present paper any detailed examination of the features of the participants' voices other than UPDRS scores, and therefore we are not able to comment within this study on the impact of different features that may be present in individuals – for example, features detectable via software analysis, such as variance in amplitude or pitch [22]. We view this as a potentially fruitful topic for further work. Thirdly, we note that the dataset also has the confounding feature that the participants are engaging in scripted reads of standard texts. It is possible that, depending on the selection criteria of the participants, the control group may be more familiar with this activity, and hence more practised at reading aloud. As one participant [ID29] states, 'I'm not used to reading aloud to people'. There is potential to examine other types of speech – indeed, the KCL dataset itself includes spontaneous speech. However, this is beyond the scope of the present study, and may be explored in future work.

Our first finding is that speech-to-text systems display a higher error rate when processing speech from individuals diagnosed with PD, versus a control group. While this is to be expected with individuals with relatively severe PD, the results suggest that mild to moderate PD has a noticeable effect on transcription error. Our second finding is that, using stock models provided alongside Whisper, the optimal choice of model in terms of accuracy in this particular dataset is the so-called 'small' model. For users with PD, the error rate on smaller models is otherwise quite high. However, since this model has over two hundred million parameters, a device with 2 GB of video RAM is a reasonable minimum requirement for deployment within the home.

Having established that it is reasonable in general, from the evidence available, to expect a loss of accuracy when a speech-to-text system is used by a person with even mild PD versus an individual from a control group, we then question what, other than larger models and better hardware, may help to mitigate this concern. One such approach is that proposed by Zheng, Phukon and Hasegawa-Johnson [44], who demonstrated that fine-tuning may be a beneficial strategy when working

with dysarthria. As discussed in section 3.7, this is facilitated by the logging and retention of diagnostic data, including user utterances captured by the system: however, this significantly impacts participant privacy. Beyond this, an analysis that takes into account the features of individual speech would be helpful in indicating in greater detail how features of speech in PD affect the performance of the speech-to-text system: once their relative significance is understood, it may be possible to use this information to guide efforts to design systems that better mitigate these limitations. For example, an improved microphone system may be an effective mitigation of quiet speech. A further mechanism that may be beneficial is the development of systems that support and actively engage in repair: for example, query of indistinct input and correction of erroneously recorded data or false positives [11].

7 Conclusion

In this study we have discussed design patterns for use of a voice agent designed to support an individual in annotating their actions, responding to system-triggered validation requests, interacting and participating in scripted activities. We have also discussed failure patterns that may occur between the individual and the device. Using an open dataset of spoken passages including participants with PD and control participants, we have benchmarked the performance of a key component of voice agent software, the speech-to-text application Whisper, in several configurations. In so doing, we have shown that according to standard error metrics, even mild PD symptoms affecting speech, which do not appear significant to comprehensibility of the audio to human annotators, appear to have a significant effect on the performance of this tool. We note that this can be partially countered by making use of the 'base' model, although this increases the hardware requirements of home installations of this software.

The question of the suitability of this approach for collecting speech annotation data in the wild still remains to be answered. We would argue that there are several facets to this question: there is the engineering challenge, of building a system that performs as optimally as possible in the intended context of use on both the hardware and software levels. There is the human-interaction challenge of building a system that supports the participant in repairing interactions when error occurs and minimises the frustration of systems that are prone to error. Finally, there is the broader question of the suitability of speech and interaction as a medium for annotating the types of information that are commonly labelled in PD, and the extent to which these actions or activities are susceptible to self-annotation by participants: the use of augmentative and alternative communication methods (AAC) is often recommended, and it is likely that exploring these approaches may be a useful direction for future research.

Acknowledgments. This work was supported by the TORUS Project, which has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC), grant number EP/X036146/1.

References

1. Ai, H., Wang, Y., Yang, Y., Zhang, Q.: An improvement of the degradation of speaker recognition in continuous cold speech for home assistant. In: International Symposium on Cyberspace Safety and Security, pp. 363–373 (2019)
2. Arnold, A., Kolody, S., Comeau, A., Cruz, A.M.: What does the literature say about the use of personal voice assistants in older adults? A scoping review. Disability and Rehabilitation: Assistive Technology **19**(1), 100–111 (2024). https://doi.org/10.1080/17483107.2022.2065369
3. Binda, J., Park, H., Carroll, J.M., Cope, N., Yuan, C.W.(, Choe, E.K.: Intergenerational sharing of health data among family members. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare. PervasiveHealth '17, pp. 468–471. Association for Computing Machinery, Barcelona, Spain (2017). https://doi.org/10.1145/3154862.3154895
4. Borzì, L., Varrecchia, M., Sibille, S., Olmo, G., Artusi, C.A., Fabbri, M., Rizzone, M.G., Romagnolo, A., Zibetti, M., Lopiano, L.: Smartphone-Based Estimation of Item 3.8 of the MDS-UPDRS-III for Assessing Leg Agility in People With Parkinson's Disease. IEEE Open Journal of Engineering in Medicine and Biology **1**, 140–147 (2020). https://doi.org/10.1109/OJEMB.2020.2993463
5. Bot, B.M., Suver, C., Neto, E.C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. *et al.*: The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific data **3**(1), 1–9 (2016)
6. Cheng, P., Roedig, U.: Personal voice assistant security and privacy—a survey. Proceedings of the IEEE **110**(4), 476–507 (2022)
7. Cho, E., Sundar, S.S., Abdullah, S., Motalebi, N.: Will deleting history make alexa more trustworthy? effects of privacy and content customization on user experience of smart speakers. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–13 (2020)
8. Contreras-Ruston, F., Castillo-Allendes, A., Saavedra-Garrido, J., Ochoa-Muñoz, A.F., Hunter, E.J., Kotz, S.A., Navarra, J.: Voice self-assessment in individuals with Parkinson's Disease as compared to general voice disorders. Parkinsonism & Related Disorders **123**, 106944 (2024). https://doi.org/10.1016/j.parkreldis.2024.106944
9. Crotty, B.H., Walker, J., Dierks, M., Lipsitz, L., O'Brien, J., Fischer, S., Slack, W.V., Safran, C.: Information sharing preferences of older patients and their families. JAMA internal medicine **175**(9), 1492–1497 (2015)
10. Cruz-Sandoval, D., Beltran-Marquez, J., Garcia-Constantino, M., Gonzalez-Jasso, L.A., Favela, J., Lopez-Nava, I.H., Cleland, I., Ennis, A., Hernandez-Cruz, N., Rafferty, J., Synnott, J., Nugent, C.: Semi-Automated Data Labeling for Activity Recognition in Pervasive Healthcare. Sensors **19**(14) (2019). https://doi.org/10.3390/s19143035
11. Cuadra, A., Li, S., Lee, H., Cho, J., Ju, W.: My bad! repairing intelligent voice assistant errors improves interaction. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1), 1–24 (2021). https://doi.org/10.1145/3449101
12. Davis, J.T., Ehrhart, A., Trzcinski, B.H., Kille, S., Mount, J.: Variability of experiences for individuals living with Parkinson disease. Journal of Neurologic Physical Therapy **27**(2), 38–45 (2003)
13. Dhiya'Mardhiyyah, A., Latif, J.J.K., Tho, C.: Privacy and Security in the Use of Voice Assistant: An Evaluation of User Awareness and Preferences. In: 2023 International Conference on Information Management and Technology (ICIMTech), pp. 481–486 (2023). https://doi.org/10.1109/ICIMTech59029.2023.10277724

14. Díaz-Oreiro, I., López, G., Quesada, L., Guerrero, L.A.: Conversational Design Patterns for a UX Evaluation Instrument Implemented by Voice. In: Rocha, Á., Ferrás, C., Méndez Porras, A., Jimenez Delgado, E. (eds.) Information Technology and Systems, pp. 530–540. Springer International Publishing, Cham (2022)
15. Distler, M., Schlachetzki, J.C.M., Kohl, Z., Winkler, J., Schenk, T.: Paradoxical kinesia in Parkinson's disease revisited: Anticipation of temporal constraints is critical. Neuropsychologia **86**, 38–44 (2016). https://doi.org/10.1016/j.neuropsychologia.2016.04.012
16. Duffy, O., Synnott, J., McNaney, R., Brito Zambrano, P., Kernohan, W.G.: Attitudes Toward the Use of Voice-Assisted Technologies Among People With Parkinson Disease: Findings From a Web-Based Survey. JMIR Rehabil Assist Technol **8**(1), e23006 (2021). https://doi.org/10.2196/23006
17. Dunbar, J.C., Bascom, E., Boone, A., Hiniker, A.: Is someone listening? audio-related privacy perceptions and design recommendations from guardians, pragmatists, and cynics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **5**(3), 1–23 (2021)
18. Garcia-Constantino, M., Beltran-Marquez, J., Cruz-Sandoval, D., Lopez-Nava, I.H., Favela, J., Ennis, A., Nugent, C., Rafferty, J., Cleland, I., Synnott, J., Hernandez-Cruz, N.: Semi-Automated Annotation of Audible Home Activities. In: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 40–45 (2019). https://doi.org/10.1109/PERCOMW.2019.8730729
19. Germanos, G., Kavallieros, D., Kolokotronis, N., Georgiou, N.: Privacy Issues in Voice Assistant Ecosystems. In: 2020 IEEE World Congress on Services (SERVICES), pp. 205–212 (2020). https://doi.org/10.1109/SERVICES48979.2020.00050
20. Gillivan-Murphy, P., Miller, N., Carding, P.: Voice Tremor in Parkinson's Disease: An Acoustic Study. Journal of Voice **33**(4), 526–535 (2019). https://doi.org/10.1016/j.jvoice.2017.12.010
21. Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R. *et al.*: Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement disorders: official journal of the Movement Disorder Society **23**(15), 2129–2170 (2008)
22. Holmes, R.J., Oates, J.M., Phyland, D.J., Hughes, A.J.: Voice characteristics in the progression of Parkinson's disease. International Journal of Language & Communication Disorders **35**(3), 407–418 (2000). https://doi.org/10.1080/136828200410654
23. Jaber, R., Zhong, S., Kuoppamäki, S., Hosseini, A., Gessinger, I., Brumby, D.P., Cowan, B.R., Mcmillan, D.: Cooking With Agents: Designing Context-aware Voice Interaction. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2024)
24. Jaeger, H., Trivedi, D., Stadtschnitzer, M.: *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls*. Zenodo, May 2019. https://doi.org/10.5281/zenodo.2867216.
25. Jamshed, H., Nurain, N., Brewer, R.N.: Designing Accessible Audio Nudges for Voice Interfaces. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3706598.3713563
26. Little, M.A.: Smartphones for remote symptom monitoring of Parkinson's disease. Journal of Parkinson's Disease **11**(s1), S49–S53 (2021)

27. Mahmood, A., Wang, J., Yao, B., Wang, D., Huang, C.-M.: User interaction patterns and breakdowns in conversing with LLM-powered voice assistants. International Journal of Human-Computer Studies **195**, 103406 (2025)
28. McDonald, L.M., Griffin, H.J., Angeli, A., Torkamani, M., Georgiev, D., Jahanshahi, M.: Motivational modulation of self-initiated and externally triggered movement speed induced by threat of shock: experimental evidence for paradoxical kinesis in Parkinson's disease. PLoS One **10**(8), e0135149 (2015)
29. Mishra, T., Ljolje, A., Gilbert, M.: Predicting Human Perceived Accuracy of ASR Systems. In: INTERSPEECH, pp. 1945–1948 (2011)
30. Mitra, V., Huang, Z., Lea, C., Tooley, L., Wu, S., Botten, D., Palekar, A., Thelapurath, S., Georgiou, P., Kajarekar, S. *et al.*: Analysis and tuning of a voice assistant system for dysfluent speech. arXiv preprint arXiv:2106.11759 (2021)
31. Modestino, E.J., Reinhofer, A., Blum, K., Amenechi, C., O'Toole, P.: Hoehn and Yahr staging of Parkinson's disease in relation to neuropsychological measures. Front Biosci (Landmark Ed) **23**(7), 1370–1379 (2018)
32. Morgan, C., Heidarivincheh, F., Craddock, I., McConville, R., Perello Nieto, M., Tonkin, E.L., Masullo, A., Vafeas, A., Kim, M., McNaney, R., Tourte, G.J.L., Whone, A.: Data labelling in the wild: annotating free-living activities and Parkinson's disease symptoms. In: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 471–474 (2021). https://doi.org/10.1109/PerComWorkshops51409.2021.9431017
33. Morgan, C., Jameson, J., Craddock, I., Tonkin, E.L., Oikonomou, G., Isotalus, H.K., Heidarivincheh, F., McConville, R., Tourte, G.J.L., Kinnunen, K.M. *et al.*: Understanding how people with Parkinson's disease turn in gait from a real-world in-home dataset. Parkinsonism & Related Disorders **105**, 114–122 (2022)
34. Morris, A.C., Maier, V., Green, P.D.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: Interspeech, pp. 2765–2768 (2004)
35. Morris, M.E., Matyas, T.A., Iansek, R., Summers, J.J.: Temporal Stability of Gait in Parkinson's Disease. Physical Therapy **76**(7), 763–777 (1996). https://doi.org/10.1093/ptj/76.7.763
36. Pal, D., Babakerkhell, M.D., Roy, P.: How Perceptions of Trust and Intrusiveness Affect the Adoption of Voice Activated Personal Assistants. IEEE Access **10**, 123094–123113 (2022). https://doi.org/10.1109/ACCESS.2022.3224236
37. Pfeiffer, R.F.: Non-motor symptoms in Parkinson's disease. Parkinsonism & Related Disorders **22**, S119–S122 (2016). https://doi.org/10.1016/j.parkreldis.2015.09.004
38. Politis, M., Wu, K., Molloy, S., G. Bain, P., Chaudhuri, K.R., Piccini, P.: Parkinson's disease symptoms: The patient's perspective. Movement Disorders **25**(11), 1646–1651 (2010). https://doi.org/10.1002/mds.23135
39. Quinn, N.P.: Classification of fluctuations in patients with Parkinson's disease. Neurology **51**(2_suppl_2), S25–S29 (1998). https://doi.org/10.1212/WNL.51.2_Suppl_2.S25
40. Sveinbjornsdottir, S.: The clinical symptoms of Parkinson's disease. Journal of Neurochemistry **139**(S1), 318–324 (2016). https://doi.org/10.1111/jnc.13691
41. Unified Parkinson's Disease rating scale, https://www.parkinsons.va.gov/resources/UPDRS.asp.
42. Woznowski, P., Tonkin L., E., Laskowski, P., Twomey, N., Yordanova, K., Burrows, A.: Talk, text or tag? In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 123–128 (2017)

43. Zargham, N., Reicherts, L., Bonfert, M., Voelkel, S.T., Schoening, J., Malaka, R., Rogers, Y.: Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In: Proceedings of the 4th Conference on Conversational User Interfaces. CUI '22. Association for Computing Machinery, Glasgow, United Kingdom (2022). https://doi.org/10.1145/3543829.3543834
44. Zheng, X., Phukon, B., Hasegawa-Johnson, M.: Fine-Tuning Automatic Speech Recognition for People with Parkinson's: An Effective Strategy for Enhancing Speech Technology Accessibility. arXiv preprint arXiv:2409.19818 (2024)