MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech

Young-Ho Kim  
University of Maryland  
College Park, MD, USA  
yghokim@younghokim.net

Diana Chou  
University of Maryland  
College Park, MD, USA  
dchou4@um.edu

Bongshin Lee  
Microsoft Research  
Redmond, WA, USA  
bongshin@microsoft.com

Margaret Danilovich  
CJE SeniorLife  
Chicago, IL, USA  
margaret-wente@northwestern.edu

Amanda Lazar  
University of Maryland  
College Park, MD, USA  
lazar@umd.edu

David E. Conroy  
Pennsylvania State University  
University Park, PA, USA  
conroy@psu.edu

Hernisa Kacorri  
University of Maryland  
College Park, MD, USA  
hernisa@umd.edu

Eun Kyoung Choe  
University of Maryland  
College Park, MD, USA  
choe@umd.edu

Figure 1: MyMove supports collecting in-situ activity labels using speech on a smartwatch. People can initiate the reporting from the watchface either voluntarily (a) or upon a prompt message (b); describe an activity, time span, and effort level (c); and review & submit the recording (d). MyMove displays a visual confirmation after the submission (e). The example verbal report is from P7. Please refer to our supplementary video which demonstrates the interactions.

ABSTRACT

Current activity tracking technologies are largely trained on younger adults’ data, which can lead to solutions that are not well-suited for older adults. To build activity trackers for older adults, it is crucial to collect training data with them. To this end, we examine the feasibility and challenges with older adults in collecting activity labels by leveraging speech. Specifically, we built MyMove, a speech-based smartwatch app to facilitate the in-situ labeling with a low capture burden. We conducted a 7-day deployment study, where 13 older adults collected their activity labels and smartwatch sensor data, while wearing a thigh-worn activity monitor. Participants were highly engaged, capturing 1,224 verbal reports in total. We extracted 1,885 activities with corresponding effort level and timespan, and examined the usefulness of these reports as activity labels. We discuss the implications of our approach and the collected dataset in supporting older adults through personalized activity tracking technologies.
1 INTRODUCTION

Scarcity of older adults’ activity datasets may lead to biased and inaccurate activity recognition systems. For example, a recent study showed that Fitbit Ultra, a consumer health tracking device, significantly under-reports steps at slow speed of 0.9 m/s, a representative walking speed of older adults [132]. When people walk slowly, with a cane, or a walker, such activity recognition systems have a tendency to not register steps accurately. A recent study looking at older adults’ technology usage for activity tracking shows that more than a half do not trust the accuracy of these devices [102], which are typically trained on younger adults data. To develop activity tracking systems that are inclusive of and beneficial to older adults, it is imperative to collect older adults’ movements and activity data.

Activity tracking technologies can provide meaningful feedback that supports people’s motivations, playing an important role in enhancing physical activity [32, 80, 122]. Like individuals in many age groups, physical activity is important for older adults, favorably influencing their healthy daily routine [27] and active life expectancy [20], chronic health conditions including coronary heart disease, hypertension, and type 2 diabetes [127], psychological health and wellbeing [20], enjoyment [92, 93], and social wellbeing [6]. However, the adoption rate of activity tracking technologies for older adults is relatively low (e.g., 10% for age 55+ whereas 28% for ages 18–34 and 22% for ages 35–54 [78]). Meanwhile, studies continuously report that younger, more affluent, healthier, and more educated groups are more likely to use activity tracking technologies [18, 75, 78, 126].

We suspect that the current activity tracking technologies are designed with little understanding of older adults’ lifestyles and perspectives (e.g., types of activities they engage in and care about) and do not account for heterogeneous physiological characteristics (e.g., gait and locomotion [12]). Our ultimate goal is to support older adults’ agency by designing and developing personalized activity tracking technologies that better match their preferences and patterns. As a first step, we set out to develop an activity labeling tool that older adults can use to collect in-situ activity labels along with their sensor data. These labels could be used to train and fine-tune classifiers based on inertial sensors.

To this end, we conducted a 7-day deployment study with 13 older adult participants (age range: 61–90; average: 71.08), where they collected activity descriptions while wearing a smartwatch and a thigh-worn activity monitor; the thigh-worn activity monitor served as a means for collecting ground-truth sensor data for our analysis and later model development. To facilitate collecting in-situ descriptions with a low data capture burden, we designed and developed an Android Wear reporting app, called MyMove, leveraging speech input, an accessible modality for many older adults [98]. With MyMove on a smartwatch, participants can describe activity type, associated timespan, and perceived effort level. Many smartwatches are equipped with a microphone, which allows people to flexibly describe their activities using speech. As an on-body device, a smartwatch can collect continuous activity sensing data and deliver notifications, which is necessary to collect in-situ data through an experience sampling method (ESM) [64].

Furthermore, prior work co-designing wearable activity trackers with older adults showed that the “watch-like” form factor was mostly preferred due to its ability to tell time, on-body position, and public acceptance [121]. Through our deployment study, with a focus on feasibility, we explore the following questions: (1) How do older adults capture their activities using speech on a smartwatch? and (2) How useful are their verbal reports as an information source for activity labeling?

Our results show that participants were highly engaged in the data collection process, submitting a total of 1,224 verbal reports (avg. 13.45 reports per day per participant) and wearing the smartwatch and monitor throughout the seven-day study period. From these reports, we extracted 1,885 activities with 29 different activity types that comprehensively capture participants’ daily lifestyles. Participants provided time-related information for about a half of the activities but they were more likely to provide complete time information when reporting a single activity or when reporting voluntarily as opposed to being prompted. Participants’ effort level categories were aligned with sensor-based intensity metrics in the corresponding time segments. However, activities that participants evaluated as moderate to high intensity did not meet the standard intensity level according to the sensor-based intensity measurements. All of the 1,224 verbal reports were valid and could be transcribed and understood by a researcher. Furthermore, the word error rates of these reports by two state-of-the-art speech recognition systems were relatively low: 4.93% with Microsoft Cognitive Speech and 8.50% with Google Cloud Speech. Through our study, we demonstrated that by leveraging speech, MyMove can facilitate collecting useful activity labels. We also identified how we can further improve speech-based activity labeling tools for older adults; for example, by leveraging multi-device environments to collect more accurate and fine-grained data and by providing self-monitoring feedback to enhance engagement. The key contributions of this work are:

(1) Design and development of MyMove, an Android Wear reporting app for supporting older adults in collecting their activity descriptions with a low data capture burden by leveraging speech input on a smartwatch.
(2) Empirical results from a deployment study conducted with 13 older adults using MyMove, demonstrating the feasibility of collecting rich in-situ activity descriptions from older adults via speech.
(3) Examining the characteristics and usefulness of the data collected with MyMove, in terms of activity type, time, and effort.
level as well as the quality of the voice recording (automatic speech recognition error).

2 RELATED WORK

In this section, we cover the related work in the areas of (1) understanding older adults’ activities, (2) collecting in-situ behavioral data, and (3) in-situ data labeling for human activity recognition.

2.1 Understanding Older Adults’ Activities

Researchers, healthcare providers, and government officials have been interested in understanding daily activities of older adults because it helps establish and improve health-related guidelines, policies, and interventions [13, 34, 112]. Researchers have defined “activity” differently depending on their research focus. For example, there is a focus on assessing the independence/dependence with functional tasks, as reflected in the concept of ADL (Activities of Daily Living—basic self-maintenance activities, such as eating, dressing, bathing, or toileting) [56] and IADL (Instrumental ADL—higher-level activities that require complex skills and mental load, such as making a phone call, shopping, housekeeping, or financing) [65]. Another subset of research categorizes activities based on the level of energy expenditure (e.g., classification of energy costs of daily activities [1]), as reflected in many physical activity questionnaires they developed to assess older adults’ intensity-specific duration for behavior (e.g., MOST [33], CHAMPS [114], LASA [125]).

Domestic and leisure activities are prevalent in older adults’ daily activities [49, 55, 79, 83]. According to the national time use surveys from 14 countries, older adults (aged 60–75) spent around 6 hours on leisure and ≥2.5 hours on domestic work daily [55]. From the interviews with U.S. older adults, Moss and Lawton found that participants spend about 5 hours a day on obligatory personal care & household activities and more than 6 hours a day on discretionary leisure activities [83]. Another study with Australian older adults reported that participants spend the longest time on solitary leisure (avg. 4.5 hours a day) excluding sleep, followed by IADL (avg. 3.1 hours a day), social leisure (avg. 2.7 hours a day), and ADL (avg. 2.6 hours a day) [79].

Researchers have further examined what kinds of activities older adults engage in during their leisure time [55] grouping them as active (e.g., relaxing, socializing, volunteering, organization work, religion, going out, sports and exercising) and passive (e.g., reading, listening to the radio, watching television, and browsing the internet on a computer) with the latter often involving screen time. Screen time is one of the most prevalent leisure time activities [88]; studies consistently report that older adults spend longer than 2 hours a day watching TV (e.g., avg. 2.5 hours [49], avg. 3.5 hours [83], and over 3 hours for 54.6% of an older population [43]). Screen time is known to be a strong indicator of discretionary sedentary behaviors (i.e., low energy expenditure activities in a seated or reclined posture while awake [110]). Decreased physical activity during leisure time and increased sedentary time is another common characteristic of older adults that may be disproportionately affected by many other factors such as the socioeconomic status of their neighborhood [3]. The U.S. national surveys in 2015–2016 revealed that 64% of older adults aged 65+ reported being inactive (i.e., no moderate or vigorous-intensity activity for 10 minutes per day), and 53% reported that they sit longer than 6 hours a day [131]. In a similar vein, a study using an accelerometer sensor (ActiGraph) found that older adults aged 70+ in the urban UK spend less than 30 minutes on moderate-to-vigorous physical activities and the duration significantly drops with age [23].

This body of knowledge—that is typically based on retrospective recall, surveys, and automated sensing—provides a general understanding of older adults’ activities and time use. In our work, however, the purpose of collecting older adults’ activities is quite different: going beyond understanding how older adults spend their time, we aim to examine the feasibility of creating a training dataset that contains older adults’ activity patterns. To this end, we employ a low-burden, in-situ data collection method that older adults can partake in to collect fine-grained data of their activities.

2.2 Collecting In-Situ Behavioral Data

Methods that rely on retrospective recall, such as interviews or surveys, are subject to recall bias [42], which may be affected by the nature of an event and people’s experiences. For example, in responding to a survey, people were likely to accurately estimate the past duration of intensive physical activities [9, 52, 109], whereas they were likely to underestimate or omit light and sedentary activities [9, 52, 66, 107, 109]. To collect more ecologically valid self-report data, researchers devised Diary Study [8] and Experience Sampling Method (ESM, often interchangeable with ecological momentary assessment or EMA) [64]. Both methods have been employed before the widespread use of smartphones, but smartphones and their notification capability have made it much easier to facilitate these methods. In Diary Studies, people are expected to capture self-report data once (or more) a day using pen and paper or diary apps. Although Diary Studies help researchers collect in-situ self-report data, there can be a delay between when an event happens and when that event is captured. To further reduce recall bias, ESM employs notifications (defined by a certain prompting rule) to signal when to capture data, and people are expected to capture data at the moment of (or shortly after) receiving the notification. Researchers typically employ ESM to collect brief self-report data frequently. Therefore, in an ESM study, it is important to strike the balance between researchers’ data collection needs and participants’ data capture burdens.

To reduce data capture burdens, researchers have explored smartwatches as a new means to facilitate ESM [45, 134]; wearing smartwatches allows for high awareness of and alertness to incoming notifications with glanceable feedback [16, 96]. In terms of the notification delivery, prior work has demonstrated that smartwatch-based ESM can yield shorter response delays [45], higher response rates, and EMA experiences perceived as less distracting [51, 97] when compared to smartphones. On the other hand, an inherent drawback of smartwatches for ESM is their small form factor, which can make it laborious to enter data. Thus, approaches typically employed on smartphones (e.g., entering data via a text box) are inefficient. To ease the data entry, researchers have explored more effective input methods such as the ROAMM [59] and PROMPT [77].
frameworks, which support radial scales and bezel rotation to specify pain level and activity type. Others have combined touch and motion gestures for answering ESM questions [134]. These prior studies predominantly incorporated graphical widgets with touch/hand gestures for structured questions with simple choices (e.g., “yes” or “no”). One input modality on a smartwatch that has not been actively considered for ESM on a smartwatch is speech, which is widely embedded in consumer devices and digital systems [22]. When people speak, they tend to be faster [105] and more expressive [17, 103] than when they type. Speech input requires little to no screen space and researchers found that speech commands can be easier to perform than using graphical widgets on mobile devices (e.g., [60, 113]). Recent work has shown promise for speech input for in-situ data collection on digital devices (e.g., exercise logging on a smart speaker [72], food journaling on a smartphone [71]). For example, Luo and colleagues deployed a speech-based mobile food journal and found that participants provided detailed and elaborate information on their food decisions and meal contexts, with a low perceived capture burden [71]. Using speech input on a smartwatch poses great potential for lowering the data capture burden while enhancing response rate in EMA studies. It allows us to mitigate touch interactions that involve on-screen finger movement, such as scrolling, which may be burdensome for older adults [5]. Given that voice-based interfaces tend to be accessible for many older adults [98] (including those with low technology experience), in this paper, we explore how older adults leverage speech input on a smartwatch to collect in-situ activity data in an open-ended format. This is a novel approach to prior ESM studies that collected responses to structured questions (e.g., multiple choice, Likert scale).

2.3 In-Situ Data Labeling for Human Activity Recognition

Another relevant topic to our work is Human Activity Recognition (HAR), an automated process of relating sensor stream time segments with various human activities (e.g., walking, running, sleeping, eating) [63]. HAR has been extensively applied to a wide range of technologies, from broader consumer fitness trackers to specialized tracking systems for older adults in capturing physical activities and ADLs or detecting falling or frailty [34, 121, 122]. The quality of an HAR model depends on how sensor data (i.e., input) were collected [63]; models trained with the sensor data captured in the lab tend to yield lower accuracy when tested outside [31]. However, gathering both ground-truth activity labels (i.e., the type of activity, the start and end time of an activity) and sensor data in the natural context of daily life is generally challenging because it may not be ethical or feasible for researchers to observe participants’ activity outside the lab [48].

To enable in-situ collection of both the sensor data and the activity labels, the UbiComp and HCI communities have proposed mobile and wearable systems that allow participants to label their own activities, while collecting sensor data in the background (e.g., [82, 108, 119]). For example, VoiSense [108] is a conversational agent on Apple Watch that allows people to capture the physiological or motion sensor data for a designated duration and then specify a label for the session, though, it has not been evaluated yet with users. ExtraSensory App [119] is an in-situ activity labeling system that consists of a mobile and smartwatch app. On the mobile app, people can review their activity history and labels for past or near-future time segments. The smartwatch app complements the mobile app by forwarding notifications or receiving binary confirmations about the current status (e.g., “In the past 2 minutes you were still sitting?”). When labeling on the mobile app, people can select multiple labels from a predefined list (e.g., “Sitting + At work”) that best describes the time segment. Data collected with the ExtraSensory App typically include younger adults (ages 18-42).

Our work extends this line of research on collecting in-situ activity labels in two ways. First, unlike prior systems that primarily target younger adults, we aim to work with older adults with interfaces that are specifically designed for this population (see Design Rationale DR1 in Section 3.1). Second, unlike VoiSense and ExtraSensory App, which collect structured label data through multiple steps of speech or touch inputs, we collect activity information as an unstructured verbal description on the activity type, associated timespan, and perceived level of effort. In doing so, we explore how useful such utterances are as a source of information for activity labeling and discuss the implications of our findings for how to design low-burden in-situ activity labeling systems suitable for older adults.

3 MYMOVE

As a low-burden activity reporting tool intended for older adults, we designed and developed MyMove (Figure 1), a speech-based Android Wear app. MyMove allows people to submit a verbal description of their activities (which we call a verbal report throughout the paper) in two different methods: (i) report voluntarily at any time or (ii) report when they are prompted by ESM notifications. MyMove asks people to include activity, time/duration, and perceived effort level in their verbal report. The activity and associated timespan are the two essential components of labeling sensor data for Human Activity Recognition [63]: activity labels can be extracted from activity descriptions and the timespan connects the activity and the sensor values. Capturing the perceived level of effort is important because it varies from person to person even when they perform the same activity (e.g., the number of repetitions, speed, weight lifted) [10]. In the background, MyMove captures sensor data streams and transmits them to a backend server. In this section, we describe our design rationales and the MyMove system along with the implementation details.

3.1 Design Rationales

DR1: Prioritize Older Adults. Both the form factor and interaction modalities of MyMove are informed by prior work with older adults in support of smartwatches in the context of activity tracking (e.g., [29, 121]), voice as an accessible input modality for many older adults [98], and large target buttons associated with tapping or pressing [15, 84].

We carefully selected hardware (i.e., Fossil Gen 5) that has a relatively big display among other smartwatch options with similar sensing. Interacting with Android Wear’s native notifications requires bezel swiping and scrolling, and we have little control over the text size and layout of a notification. Thus, we designed and
implemented a custom watchface to display our prompt messages (e.g., Figure 1b). We also allowed people to choose either physical or virtual (touchscreen) buttons for most functionalities, considering diverging preferences of older adults on the physical and virtual buttons [77, 128]. We assigned up to two main functions on each screen and placed virtual buttons with a white background (e.g., Figure 1a–1d) near the top-right and the bottom-right physical buttons on the side, with each virtual button matching the corresponding physical button. For consistency, we assigned positive actions (e.g., confirm, launch the reporting) to the top-right button and negative actions (e.g., cancel, dismiss a prompt) to the bottom-right button.

**DR2: Simplify Data Capture Flow.** Considering that data entry is repeated frequently, we streamlined the user interface flow for activity reporting. For example, people can submit an entry by pressing the top-right button twice, first to initiate the recording (Figure 1a or 1b → Figure 1c), and second to end the recording (Figure 1c → Figure 1d). Upon completion of the recording, the review screen (Figure 1d) automatically submits the report so that people do not have to explicitly press the “OK” button. We followed the design of traditional voice recording interfaces, initially allowing pausing and resuming the recording. However, throughout the pilot study we found that pausing/resuming was rarely used but rather made the flow more confusing and therefore removed that functionality.

**DR3: Leverage the Flexibility of Natural Language Speech Input.** It can be challenging to specify activity types or time/duration information using only graphical user interface widgets on a smartwatch. The screen is so small that entering data via a text box can be inefficient. Selecting an activity type from a long list of activities is tedious and prone to error (e.g., ExtraSensory’s smartphone app [119] supports about 50 activity tags on a hierarchical list, but its companion smartwatch app does not support this tagging activity). Furthermore, specifying time/duration using touch is laborious and inflexible; a smartwatch’s small screen does not afford two time pickers (for start and end) in one screen and existing time pickers are not flexible enough to handle the various ways to specify time (i.e., people should specify absolute start and end time) [60]. We also wanted to allow participants to freely describe the effort level to examine what expressions they use to gauge their effort in what situation instead of using the validated scales such as Borg’s CR10 scale [11, 21].

To mitigate these limitations, we leveraged speech input that affords a high level of freedom without requiring much screen space [60]. People can specify multiple information components in a single verbal report (e.g., “I took a 30-minute walk” to specify an activity with duration; “I did gardening, fixing flower beds from 9:00 to 10:30, in moderate intensity” to specify an activity with duration and effort level).

### 3.2 Data Collection

**Verbal Activity Reports.** MyMove collects verbal reports in two different ways: people can submit a report voluntarily at any time or they can submit a report responding to ESM prompts. Each prompt is scheduled to be delivered at random within hourly time blocks while people are wearing the smartwatch. To send the prompts only when people are wearing the watch, we leveraged the smartwatch’s built-in off-body detect sensor. Once a prompt is delivered, the next one is reserved within the next hour window while leaving at least a 30-minute buffer after the previous one. If the user submits a voluntary report, the next prompt is rescheduled based on the submission time following the same rule.

We incorporated custom watchfaces to provide coherent visual interfaces (Figure 1). On the default screen, the watchface displays a clock, the number of reports (logs) that were submitted during the day, and a record button to initiate voluntary reporting (Figure 1a). When a prompt is delivered, the smartwatch notifies the user with two vibrations and displays a message “Describe in detail what you are doing now.” with the record and dismiss buttons on the watchface (Figure 1b). The prompt on the watchface stays for 15 minutes. However, for safety reasons, prompts are skipped if the system recognizes that the user is driving based on the Google Activity Recognition API [36].

When the user starts the recording by tapping on the “Record” or button on the watchface (or corresponding physical button), the watch vibrates three times while displaying the message, “Start after buzz,” to indicate initiation. Then MyMove shows the Record screen (Figure 1c), where people can describe an activity in freeform. The screen displays a message, “Activity, duration, and effort level?” to remind people of the information components to be included. Recordings can be as long as 2 minutes; after which the session is automatically canceled and the audio is discarded. The user completes the recording by pressing the “End” button, after which they are sent to the Review screen (Figure 1d) where they can play back the recorded audio (the Fossil watch had a speaker). The recording is submitted upon pressing the “OK” button or after 8 seconds without any interaction. While recording or reviewing, the user can discard the report using the button.

**Background Sensor Data.** MyMove also collects three behavioral and physiological measurements from the onboard sensors and APIs in the background. First, every minute, MyMove records a 20-second window of inertial sensor measurements—accelerometer, rotation vector, magnetometer, and gravity—in 25 Hz (500 samples each). Second, the system records the step counts in one-minute bins and heart rate samples (BPM) at every minute using the smartwatch’s built-in sensors. Lastly, MyMove collects the classification samples from Google Activity Recognition API, a built-in API that classifies the present locomotion status (e.g., walking, running, still in position, in vehicle, on bicycle) based on the onboard sensors.

### 3.3 Implementation

We implemented the MyMove app in Kotlin [33] on Android Wear OS 2 platform. As a standalone app, it does not require a companion app on the smartphone side. The verbal reports and sensor data are cached in local storage and uploaded to the server when the smartwatch has a stable internet connection. To optimize network traffic and disk space, the MyMove app serializes sensor data using Protocol Buffers [37] and writes them in local files. The server stores the received data in a MySQL database.

---

1Refer to our supplementary video that demonstrates the two reporting methods.

2The Wear OS 2+ watches can be paired with both iPhone and Android.
4 DEPLOYMENT STUDY

In May–July 2021, we conducted a deployment study using MyMove to examine the feasibility of speech-based activity labeling on a smartwatch with older adults and the usefulness of the verbal reports in activity labeling. As part of this study, participants reported their activities using a smartwatch while also wearing an activPAL activity monitor [89] on their thigh; this monitor served to collect ground-truth activity data to complement those captured by the wrist worn smartwatch. Due to the COVID-19 pandemic, all study sessions (introduction, tutorial, and debriefing sessions) were held remotely using Zoom video calls and the study equipment was delivered and picked up by a researcher, complying with COVID-19 prevention guidelines. This study was approved by the Institutional Review Board of the University of Maryland, College Park.

4.1 Pilot Study

We iterated on the MyMove design (e.g., data capture flow) and the study procedure (e.g., tutorials) via piloting with two older adults. In an attempt to balance the power structure between older adult participants and our research team, our first pilot participant was a retired HCI researcher. We asked them to follow the study procedure, interact with MyMove and the thigh-worn sensor for 3 days, and provide feedback on the overall study, not as a representative participant but as someone who is both a member of the intended user group and an expert in human-computer interaction. Their feedback informed our design refinement by significantly simplifying the interaction flows, incorporating icons and labels, as well as adding visual feedback making the consequence of users’ interactions more noticeable. Upon refining the app design and corresponding tutorial materials, we conducted a second pilot session with another older adult (without any HCI background) to ensure that the watch app and tutorial materials are understandable.

4.2 Participants

We recruited 15 older adults (P1–P13; 10 females and three males) through various local senior community mailing lists in the Northeast region of the United States. Since our study required in-person delivery of the study equipment, we recruited participants in the local area. Our inclusion criteria were adults who (1) are aged 60 or older; (2) feel comfortable describing their activity in English; (3) are curious about their activity levels and interested in collecting activity data; (4) have no severe speech, hearing, motor, movement, or cognitive impairments; (5) have stable home Wi-Fi and are able to join Zoom video calls; and (6) are right-handed. We exclusively recruited right-handed people because Fossil Gen 5 is designed to be worn on the left wrist. The physical buttons are on the right side of the display with the fixed orientation, making it difficult to maneuver the buttons with the left hand. This also helped to minimize the effect of handedness on sensor data.

Table 1 shows the demographic information of our study participants and the average daily activities during the data collection period, measured by activPAL monitors. All participants were native English speakers and their ages ranged from 61 to 90 (avg = 71.08). Eight participants were retirees, three were self-employed, and two were full-time employees. Participants had diverse occupational backgrounds and all participants had Bachelor’s or graduate degrees; five had Master’s degrees and one had a Ph.D. All participants were smartphone users; seven used an iPhone and six used an Android phone.

The 7-day activePAL sensor data we collected during the study show our participants’ activity level in more detail: Based on existing conventions for interpreting older adults’ physical activity volume (i.e., step counts), many of the participants were “low active” (46%; 5000–7499 steps/day) or “sedentary” (15%; < 5000 steps/day) [117]. The majority of the participants (77%) did not meet the 150 min/week of moderate-to-vigorous physical activity (MVPA) recommended in the 2018 Physical Activity Guidelines for Americans [94]. The average daily physical activity volume (M = 7246.69, SD = 2302.42 steps/day) was consistent with reduced all-cause mortality risk from previous studies with older women [67]. The mean duration of sedentary behavior was 10 hours and 44 minutes per day (SD = 2 hours and 33 minutes). This high level of sedentary behavior is comparable to device-measured normative values from older adults (10.1 hours/day) [104] and exceeds self-reported normative values from older adults (6.1 hours/day) [135].

In appreciation for their participation, we offered participants up to $150, but we did not tie the activity reporting to the compensation to ensure natural data entry behavior. We provided $25 for completing the adaptation period with the introductory and tutorial sessions, and another $25 for a debriefing interview. During the data collection period, we added $10 for each day of device-wearing compliance (i.e., wear the smartwatch for longer than 4 hours/day), and provided an extra $30 as a bonus for all seven days of compliance. We did not specify a minimum amount of time for wearing the activPAL monitor. Compensation was provided after the debriefing session in the form of an Amazon or Target gift card.

4.3 Study Instrument

We deployed a Fossil Gen 5 Android smartwatch, an activPAL4 device, and a Samsung A21 smartphone to each participant. We chose the Fossil Gen 5 Android smartwatch for its large screen size and extended battery life. The smartwatch has a 1.28-inch AMOLED display with a 416 × 416 resolution (328 PPI). To minimize the effort for the initial set up [91], we deployed smartwatches and Samsung A21 smartphones configured in advance. The phone served as an internet hub for the watch and participants did not have to carry it. While the Bluetooth connection between the watch and the phone was active, the watch periodically uploaded the sensor and verbal reports to our server via the phone’s network connection using the participant’s home Wi-Fi.

To collect the ground-truth activity postures, we also deployed activPAL4 [89], which is a research-grade activity monitor that uses data from three accelerometers to classify fine-grained body posture and locomotion (e.g., stepping, sitting, lying, standing, in vehicle, and biking). The sensor is attached to the midline of the thigh between the knee and hip using hypoallergenic adhesive tape, and the device does not provide feedback to participants. We chose activPAL for three main reasons: First, activPAL can distinguish different stationary postures such as sitting, lying, and standing, more accurately than the wrist-worn or handheld sensors (e.g., Google Activity Recognition API supports only a Still class for a stationary state) [109]. Second, activPAL is pervasive because it has
Table 1: Summary of age and gender of our study participants, their employment status and the latest (or current) occupation, education level, technical proficiency, and the average daily activities measured with an activPAL monitor during the data collection period, including step count, the time spent for moderate-to-vigorous physical activity (MVPA, the total duration at least 100 steps/min), and the time spent sedentary (the time spent sitting and lying while waking).

| Participant | Employment & Latest occupation         | Education        | Tech proficiency | activPAL daily average |
|-------------|----------------------------------------|------------------|------------------|------------------------|
|             |                                        |                  |                  | Steps | MVPA | Sedentary |
| P1 61 (M)   | Retired Senior manager                 | Bachelor’s       | Very confident   | 10,941 | 11h 23m |
| P2 67 (F)   | Self-employed Visual artist            | Bachelor’s       | Enjoy the challenge | 6,192 | 21m | 6h 21m |
| P3 77 (F)   | Retired Qualitative researcher         | Ph.D./M.D.       | Very confident   | 9,655  | 10h 53m |
| P4 70 (M)   | Self-employed Landlord                 | Bachelor’s       | Enjoy the challenge | 7,793 | 32m | 9h 7m |
| P5 81 (F)   | Retired Disability consultant          | Master’s         | A little apprehensive | 8,773 | 23m | 7h 48m |
| P6 79 (F)   | Retired Policy analyst                 | Master’s         | Very confident   | 7,320  | 9h 12m |
| P7 69 (F)   | Full-time Business manager             | Master’s         | Enjoy the challenge | 6,499 | 21m | 12h 5m |
| P8 90 (F)   | Self-employed Piano tutor              | Master-level     | Enjoy the challenge | 6,281 | <1m | 12h 24m |
| P9 62 (F)   | Full-time Communications director      | Master-level     | Very confident   | 5,313  | 13h 50m |
| P10 62 (F)  | Retired Human resource specialist      | Bachelor’s       | Very confident   | 3,430  | 13h 37m |
| P11 67 (F)  | Retired Technical training manager     | Master-level     | Enjoy the challenge | 7,296 | 2m | 7h 19m |
| P12 75 (F)  | Retired Rehabilitation counselor       | Master’s         | Very apprehensive | 4,148 | 9m | 13h 58m |
| P13 64 (M)  | Retired Regulatory specialist          | Master’s         | Enjoy the challenge | 10,566 | 46m | 11h 30m |

a long battery life (longer than 3 weeks). Third, activPAL yields equivalent reliability to Actigraph devices for physical activity [62, 73] and is more accurate than them for capturing slower gait speeds, which are common in older adults [44, 106].

4.4 Study Procedure

The study protocol consisted of four parts: (1) introductory session and four-day adaptation period, (2) tutorial session, (3) seven-day data collection, and (4) debriefing. We iterated on the study procedure and tutorial materials through the pilot sessions with two older adults. The introductory, tutorial, and debriefing sessions were held remotely on Zoom. All sessions were recorded using Zoom’s recording feature.

**Introductory Session & Adaptation Period.** After receiving the study equipment, the participant joined a 45-minute introductory session via Zoom. The researcher shared a presentation slide (refer to our supplementary material) via screen sharing. After explaining the goal of the study, the researcher guided the participant to set up the smartphone by connecting it to the home Wi-Fi, wear the smartwatch on the left hand, and attach the activPAL (waterproofed with a nitrile finger cot and medical bandage) to a thigh. To ensure that the participant felt comfortable handling the smartwatch buttons and the touchscreen elements, we used a custom app in MyMove which can be monitored by the researcher on a web dashboard; the participant went through several trials of pressing a correct button following the message on the screen (e.g., “Tap the button [A] on the screen” or “Push the button [C] on the side”).

We incorporated the adaptation period to familiarize participants with charging and wearing the devices regularly. During this period, which lasted for four days including the day of introductory session, participants were asked to wear the smartwatch during waking hours and the activPAL for as long as possible. The activity reporting feature was disabled and invisible to the participants. At 9:00 PM, an automated text reminder was sent to participants’ own phones to remind them to charge the watch before going to bed.

**Tutorial.** On the final day of the adaptation period, we held a 1-hour tutorial session on Zoom to prepare participants for the data collection period starting the next day. The tutorial mainly covered the activity reporting, including a guide on what to describe in a verbal report and how to perform prompted and voluntary reporting with MyMove on a smartwatch. We instructed that the verbal reports are “free-response descriptions about your current or recently-finished activity” and they can be freely and naturally phrased using one or more sentences. We went through 10 example reports with images of performing the activity in five categories—moving and aerobic exercises, strength exercises, stretching and balance exercises, housekeeping, and stationary activities. All example reports contained the three main information components we are interested in: activity detail, time & duration, and effort level. For each category, we encouraged participants to come up with imaginary reports including those three components.

We covered the activity reporting features by demonstrating example flows using animated presentation slides and asking participants to practice on their own watch. Since the session was remote, we observed the participant’s smartwatch screen via screen sharing feature of MyMove. We gave participants enough time to practice until they felt comfortable interacting with the smartwatch interface. For the rest of the day, participants were also allowed to submit verbal reports as practice; these reports were not included in the analysis.

We also explained the compensation rule (see the Participants section above) in detail using a few example cases. We emphasized that the compensation would not be tied to the number of reports,
but it would depend on the weartime of the smartwatch (i.e., they need to wear the smartwatch at least 4 hours a day.)

**Data Collection.** The day following the tutorial, participants started capturing their activities with MyMove, which lasted for one week. During this data collection period, participants received prompt notifications and the device-wearing compliance guideline was in effect. We also sent charging reminders at night just as during the adaptation period.

**Debriefing.** After the seventh day of the data collection, we conducted a semi-structured debriefing interview with each participant on Zoom for about 40 to 70 minutes. We asked participants to share their general reactions to the interface and smartwatch as well as their experiences with specifying information components, discussing when they would use prompted or voluntary methods, and if they had a preference towards virtual or physical buttons and why. To help participants better recall their experience, we transcribed their verbal reports in advance and shared a summarized table (similar format as Table 3) via screen sharing.

Three researchers participated in the debriefing interview sessions, two of whom led the interviews: following the detailed interview script, each researcher covered about a half of the questions. The third researcher observed nine (out of 13) sessions and filled in one session when the second researcher was not available.

### 4.5 Data Analysis

The study produced a rich dataset including the verbal reports that participants submitted, the sensor data captured from the smartwatch and activPAL, and participants’ feedback from the debriefing interviews. We performed both quantitative and qualitative analysis to examine how older adult participants used MyMove to collect in-situ activity labels and to inspect the characteristics and condition of the collected data. We first examined reporting patterns such as the number of reports collected via two reporting methods as well as audio length and word count of the reports. We analyzed the device usage logs from MyMove and the event logs from activPAL to examine the sensor wearing patterns.

We then analyzed the transcripts from the verbal reports to understand the semantics of activities participants captured. Two authors first independently coded a subset of reports after the data collection of the first four participants was completed (80 out of 354; 23%). We resolved discrepancies and developed the first version of the codebook. As we obtained additional verbal reports from new participants, we iterated multiple sessions of discussions to improve the codebook. After the codebook was finalized, the first author reviewed the entire dataset. Through a separate analysis, we extracted the effort levels from the reports. Two authors separately coded a subset of reports (180; 14.7%) and resolved discrepancies through a series of discussions. After we determined nine categories and how to code data consistently under these categories, the first author coded the remaining data.

We further analyzed the transcribed reports to check how diligently participants reported the time component and how well the self-reported information is aligned with the sensor data. We classified the reports into three categories: (1) **No time cues**: the report does not include any time-related information; (2) **Incomplete time cues**: the report includes time cues that are not enough to identify the activity timespan; and (3) **Complete time cues**: the report includes time cues that are sufficient to identify the activity timespan. For example, one of P8’s prompted reports, “I’m just finished fixing a little dinner,” has time-related information (i.e., end time) but we cannot determine the timespan for this activity without the start time or duration. Therefore, this report is classified into the Incomplete time cues category.

We transcribed the audio recordings of the debriefing interviews. The three researchers who conducted the interviews led the analysis of the debriefing interview data, using NVivo (a qualitative data analysis tool). We grouped the data specific to participants’ usability-related experiences with MyMove according to the following aspects: (1) reactions to MyMove and smartwatch, (2) reactions to specifying information components, (3) reactions to using voluntary and prompted methods, and (4) notions on choosing virtual versus physical buttons. When appropriate, we also referenced this information while interpreting the results from the analyses mentioned above.

### 5 RESULTS

We report the results of our study in six parts, aiming to answer the two research questions—first, to demonstrate the feasibility of collecting the activity reports using speech on a smartwatch; and second, to examine the usefulness of the verbal reports as an information source for activity labeling. In Section 5.1, we provide an overview of the collected dataset, including participants’ engagement in capturing the data. In Section 5.2, we report the types of activities that participants captured. We specifically discuss how participants’ lifestyles and other study contexts affect the reporting patterns and behaviors. In Section 5.3, we report how participants describe the time information in their verbal reports and discuss how the nature of an activity and reporting methods affect the completeness of the time cues. We also explore how the verbally-reported activities are aligned with those detected by sensors on a timeline. In Section 5.4, we explore how participants described their effort level, and assess the validity of the effort level description in relation to the device-based intensity measures. In Section 5.5, we examine the accuracy of automatic speech recognition technologies in recognizing older adult participants’ verbal reports. We further investigate the erroneous instances in detail. Lastly, in Section 5.6, we report on participants’ experience with MyMove, based on the qualitative analysis of debriefing interviews.

#### 5.1 Dataset Overview

While the minimum requirement was to wear the smartwatch and activPAL for at least five days (longer than four hours a day for the smartwatch), all 13 participants wore both devices for the entire seven days. On average, participants wore the smartwatch for 11.6 hours per day ($SD = 1.3$, $min = 9.7$ [P11], $max = 13.6$ [P13]), and activPAL for 23.3 hours per day (10 participants continuously wore activPAL for the entire study period).

We collected 1,224 verbal reports in total, consisting of 617 prompted and 607 voluntary reports: Table 2 shows the verbal reports by participants. Although the reporting was not tied to the compensation, all participants submitted verbal reports every day.
Participants submitted 94.15 reports on average, with a high variance among them ($SD = 52.85$, min $= 51$ [P9], max $= 250$ [P6]). The average audio length of and word count in each report were 18.65 seconds ($SD = 13.65$) and 32.05 words ($SD = 26.15$), respectively. The average audio length per report of each participant ranged from 10.08 [P13] to 32.03 [P12] seconds.

As participants often specified multiple activities in a single report, we extracted activities from each report, a unit of continuous task that can be coded with one or (sometimes) two semantics. For example, the report “Spent the last 12 minutes, eating breakfast, seated in front of the TV. Minimal level of effort.” [P6], specifies two simultaneous activities. We identified 1,885 activities from 1,224 verbal reports, and grouped them into the following four categories:

1. **Singleton**: 760 (62.10%) reports contained a single activity, (avg. 2.50 activities per report),
2. **Sequential**: 303 (24.75%) reports contained a series of activities (avg. 2.50 activities per report),
3. **Multitasking**: 127 (10.38%) had multiple activities performed simultaneously (avg. 2.09 activities per report),
4. **Compound**: 34 (2.78%) were a mix of singleton, sequential, or multitasking (avg. 3.06 activities per report).

### 5.2 Captured Activities

From the 1,885 activities, we identified 29 activity types and grouped them into nine high-level semantics: housekeeping, self-maintenance, non-exercise stepping, screen time, exercise, paperwork/desk work, hobby/leisure, restoring, and social (Table 3). The activity types were generally consistent with prior work in daily activities of older adults [49, 83]. Each participant captured 19.08 unique activity types on average ($SD = 4.35$, min $= 12$ [P11], max $= 26$ [P3]).

Participants frequently captured housekeeping activities such as cleaning, arranging or carrying items. These activities included straightening rooms, vacuuming, washing the dishes, or carrying goods purchased from shopping. Twelve out of 13 participants were living in a house with a yard and 11 of them captured gardening activities. However, specific tasks varied, ranging from light activities (e.g., watering flowers) to heavy activities (e.g., fixing flower beds, planting trees). Participants also frequently captured non-exercise stepping, which involves a lightweight physical activity, mostly brief in nature. For example, these activities included going up & down the stairs, walking around the kitchen at home, and walking to/from a car, as well as pushing a shopping cart in a store. Eleven participants regularly engaged in cardio exercise, which includes walking, biking, and swimming. The most common exercise was taking a walk (including walking the dog) whereas more strenuous exercise such as running was rarely captured. Eight participants engaged in strength and stretching exercises, for example, online yoga classes. Participants also captured brief strength and stretching exercises (e.g., leg lifts) they performed during other stationary activities such as TV watching or artwork. Other types of exercises included online meditation sessions, breathing exercises, and golf.

During debriefing, participants mentioned factors that affected their engagement in specific activities. Gardening was often affected by the season and weather. For example, P1, who participated in the study in mid May, noted that he engaged in gardening more than usual: “This was a high active seven days for me [sic]. Both because of weather and the time of year, we’re trying to transition the garden.” In contrast, P9, who participated in the study in late June, seldom captured gardening and noted, “It was really hot, stinky hot and, you know, not a fun thing to do [gardening] (...) in the earlier in the spring when I planted all my flowers and stuff, that feels more like gardening.” In addition, the COVID-19 lockdown reduced the overall engagement in outdoor physical activities and in-person activities. P4 noted, “I would bike downtown two or three times a week anyhow. Normally if before COVID, I’ve been down maybe four or five times for the last year.” Similarly, P11 remarked, “In pre-COVID, I would have done that [swimming] probably twice, two or three times during the week.” Many participants were involved in one or more community activities and their meetings transitioned to Zoom due to the lockdown, possibly increasing their screen time in place of the face-to-face interactions.

We learned that some activities were inherently easier to capture than others due to the contexts in which they are performed: this may have led to oversampling of those activities. For example, P3 commented on her high number of reports of watching TV: “That [watching TV] had so many times because I was sitting down and it was easy to use the watch. You know, I was taking a break, and the break allowed me to do that.” In addition, common activities were likely to be overlooked, thereby affecting the data capture behavior. For example, P11, who lives with her grandchildren, noted that she did not capture face-to-face interactions with them because such events happened throughout the day, which makes it overwhelming to capture all of them thoroughly: “If I recorded what I do with my grandkids, I would be recording all day [laughs]. A lot of times that I interact with my grandkids is kind of in short verse.”

### 5.3 Reporting Patterns for Time

Table 4 summarizes the time cue categories of activities from Singleton, Sequential, and Multitasking reports. We excluded 34 Compound reports (104 activities) because it was infeasible to reliably extract time cues for each activity. Overall, 984 out of 1781 activities (55.25%) were mapped with time cues, and 770 of them (78.25%) were mapped with Complete time cues. The remaining 796 activities (44.69%) were not mapped with any time cues.
Table 3: Nine activity semantics and 29 activity types, number of reports and participants (Ps), and example snippets from reports. Because the activity semantics and types were multi-coded, the percentages of reports add up to more than 100%.

| Semantics/types        | Reports | Ps | Example snippet                                                                 |
|------------------------|---------|----|---------------------------------------------------------------------------------|
| Housekeeping           | 263     | 13 | I’ve been doing some house cleaning which includes vacuuming. And now I’m polishing and dusting.” – P3 |
| Preparing food         | 123     | 13 | “I’m in the kitchen and I am just preparing breakfast, so I’m standing at the stove and to the toaster.” – P5 |
| Driving/in a vehicle   | 108     | 12 | “Just completed a 30-minute drive, as sitting.” – P1 |
| Gardening              | 99      | 11 | “I’m picking lettuce in my garden, stooping over. It’s not exerting, but it is then a bending and stooping.” – P5 |
| Caring for pets        | 68      | 6  | “Fed dog, bending over to get food and vegetables and reaching to get pills.” – P6 |
| Offline shopping       | 36      | 11 | “At Lowe’s, hardware in the garden section. Walking, pushing a stroller and picking up items and plant, for about 40 minutes.” – P3 |
| Other                  | 12      | 6  | “I have just been doing some light housekeeping chores.” – P7 |
| Self-maintenance       | 186     | 15 | “I ate breakfast from 6:30 until 7:03.” – P13 |
| Eating food            | 36      | 9  | “Process of getting dressed for the day. Pulling my clothes together and getting ready for what I’m going to do today.” – P10 |
| Dressing               | 24      | 2  | “Just completed a shower.” – P6 |
| Personal hygiene       | 27      | 1  | “I’m sitting, looking at a webinar on the phone.” – P4 |
| Treatment              | 17      | 5  | “I am sitting in watching videos on YouTube.” – P10 |
| Non-exercise stepping  | 164     | 11 | “I’ve just returned from a 30 minute walk, fairly easy paced, moderate effort because of the heat and humidity.” – P7 |
| Screen time            | 151     | 12 | “I’m watching TV, just I’ve been watching it for maybe 10 minutes so far.” – P2 |
| Computer               | 27      | 4  | I’m just doing stretching exercises in preparation for my strength training class, which I will be taking and I’ve been stretching for about 10 minutes.” – P10 |
| TV                     | 17      | 5  | “I am sitting at the counter in the kitchen doing a Sudoku.” – P5 |
| Mobile device          | 11      | 8  | “I’ve been seated at a concert for the past two hours.” – P5 |
| Device unspecified     | 10      | 4  | “I’ve just finished an hour and a half long workshop on meditation.” – P3 |
| Exercise               | 118     | 11 | “I’m sitting at my piano, playing the piano.” – P10 |
| Cardio                 | 51      | 8  | “I’ve been working, doing some woodworking in the basement.” – P13 |
| Strength/stretching    | 11      | 3  | “I’ve been seated at a concert for the past two hours.” – P5 |
| Paperwork/desk work    | 17      | 2  | “I am sitting at my piano, playing the piano.” – P10 |
| Hobby/leisure          | 59      | 10 | “I’m lying on my bed, reading a book. I’ve been doing that for half an hour.” – P2 |
| Reading on paper       | 17      | 6  | “I’m sitting at the counter in the kitchen doing a Sudoku.” – P5 |
| Playing puzzle/table game | 10     | 4  | “I’ve been working, doing some woodworking in the basement.” – P13 |
| Crafting/artwork       | 11      | 3  | “I’ve been seated at a concert for the past two hours.” – P5 |
| Seeing at a theater    | 8       | 2  | “I am sitting at my piano, playing the piano.” – P10 |
| Resting                | 54      | 7  | “Since the last ping I took about half an hour nap.” – P7 |
| Face-to-face interaction| 39      | 9  | “I just sat down on my front porch swing and I’m talking to a friend.” – P3 |
| Resting                | 36      | 8  | “I just completed a telephone call, regarding a personal business.” – P6 |

Reports containing a single activity were more likely to include Complete time cues than reports containing multiple activities: 64.87% (493/760) of Singleton activities were mapped with Complete time cues, compared with 20.11% (152/756) for Sequential and 47.17% (125/265) for Multitasking. Of the 319 activities from Sequential activities with time cues, about a half (167) were mapped with Incomplete time cues because participants often specified the start and end time of the entire sequence (i.e., the start time of the first activity and the end time of the last activity). However, this pattern was not consistent across all participants, mainly due to the high individual variance in the number of total reports (See Table 2) and in the portions of Singleton, Multitasking, and Compound activities.

Voluntary reports were more likely to include Complete time cues than prompted reports: 45.60% (409/897) of activities from voluntary reports were mapped with Complete time cues, whereas 40.84% (361/884) from prompted reports. Participants were more likely to omit time cues in prompted reports, especially when reporting simultaneous activities: 61.36% (108/176) of Multitasking activities from prompted reports contained Incomplete or No time cues, in comparison with 35.94% (32/89) of those from voluntary reports. Again, these patterns were not consistent across participants with high individual variance.

Regarding the reports with Complete time cues, we investigated how time segments from verbal reports are aligned with those from activPAL. Figure 2 shows the excerpts of timelines with self-report time segments of selected activities, along with the inferred activities and step counts from activPAL. Time segments from verbal reports for locomotion-based cardio exercises
### Table 4: Number of activities in Singleton, Sequential, and Multitasking reports by reporting method and the time cue category.

| Method     | Singleton reports (=activities) | Sequential activities | Multitasking activities |
|------------|---------------------------------|-----------------------|-------------------------|
|            | With time cue                   |                       |                         |
|            | Complete                        | Incomplete            | No cues                 |
| Prompted   | 226                             | 14                    | 131                     |
| Voluntary  | 267                             | 27                    | 95                      |
| Total      | 493                             | 41                    | 226                     |

such as biking and taking a walk generally corresponded with the bands with an equivalent activPAL class and clusters of peaks in step counts. For example, the red segments in Figure 2a and orange segments in Figure 2b illustrate how they are aligned with activPAL’s Biking and Stepping bands. Other kinds of walking activities from verbal reports, such as walking a dog and moving in a store also corresponded with the activPAL activity patterns, but participants’ movement was more fragmented with the Standing and Stepping classes compared to a pure walking exercise (see the orange segments in Figure 2c and 2d which also overlap with activPAL’s Standing band).

Activities performed while sitting often did not correspond with the momentary changes in the activPAL activities. For example, blue segments in Figure 2e and 2f indicate screen time and desk work activities that participants reported performing while sitting. In all cases, the bands of activPAL’s Sitting class cover a wider region than the self-report time segments.

### 5.4 Reporting Patterns for Effort Level

About a half of reports (644 out of 1,224 reports, 52.61%) contained cues on the effort level (see Table 5), with high variance among participants ($SD = 31.19\%$; $min = 5.26\%$ [P8], $max = 98.04\%$ [P9]). We grouped the effort level cues into seven orderly categories on a spectrum of No effort—Low—Moderate—Strenuous, and two additional categories—Relaxed and Uncategorizable (see Table 6). The most common effort level reported were Low activities (276 reports by 12 participants), followed by Moderate activities (152 reports by 11 participants). The majority of Low activities were stationary activities such as screen time, eating, driving, or desk work, and the Moderate activities included exercises, gardening, or thorough cleaning activities. Strenuous activities were rarely captured (20 reports by 5 participants). The Relaxed category includes responses such as “I’m sitting totally relaxed, reading my phone and watching TV,” and the Uncategorizable category covers responses that conveyed ambiguous level of effort (e.g., “Stretches for my back, knee bends. Nothing too strenuous but just to break up the sitting.”).

To examine how self-report effort level categories are related with device-based intensity measures, we compared intensity measurements across the effort level categories using mixed-effects models because these models can handle unbalanced data with repeated measured from the same participant [95]. For this analysis, we included 480 activities that contained both Complete time cues

![Figure 2: The excerpts of self-report time segments (B) of selected activities, along the timeline with automatically-inferred activities (A) and step counts (C) from activPAL. The colors denote the types of activPAL’s activity classes. The self-report time segments are color-coded as the equivalent activPAL classes.](image-url)
and Effort level cues; we counted two or more activities included in Multitasking reports as one activity because multiple activities (e.g., “watching TV while eating dinner”) were mapped to one effort level (e.g., “it was very low effort”). In this analysis, we excluded the Uncategorizable category. We employed two common indicators of intensity in physical activity research—the percentage of HR_{\text{max}} (the average heart rate during the period expressed as a percentage of age-adjusted maximum heart rate)\(^3\) and walking cadence (steps/min) [2, 118]. We generated a model for each of the three measurements—the percentage of HR_{\text{max}} from smartwatch, walking cadence from activPAL, and walking cadence from smartwatch. We used intercept (participant) as a random effect and effort level category as a fixed effect. From Maximum-likelihood tests with other variables, we found that age, elapsed days, and activity types did not have significant effects on the measurements. Therefore, we excluded them from fixed effects in the models.

We found significant differences among the effort level categories in their intensity measurements across all three metrics: F(7, 407.69) = 12.00, p < .001 for the percentage of HR_{\text{max}}; F(7, 446.69) = 12.00, p < .001 for walking cadence from activPAL; and F(7, 369.96) = 6.19, p < .001 for walking cadence from smartwatch. We conducted post-hoc pairwise comparisons of the least-squared means of intensity measurements among 8 effort level categories using Tukey adjustment in emmeans [69] package in R. Figure 3 visualizes the significance over the 95% confidence intervals of measurements in each category. Across all three metrics, the intensity measurements of the activities specified as Moderate were significantly higher than those of No effort (p < .001) and Low (p < .001). The percentage of HR_{\text{max}} and activPAL-measured walking cadence for Low-to-Moderate activities were also significantly higher than those of No effort activities (p < .005 for the percentage of HR_{\text{max}} and p = .004 for walking cadence). For Moderate-to-Strenuous activities, only the percentage of HR_{\text{max}} was significantly higher than that of No effort (p = .003) and Low (p = .036) activities. The activities specified as No effort and Low did not differ across all metrics.

Participants’ subjective evaluation of the effort level did not match the standard intensity level of physical activity, especially for the activities that are Moderate or above (26.67%; 128/480). Of the 119 Moderate, Moderate-to-Strenuous, and Strenuous activities with the percentage of HR_{\text{max}} measurements, only one activity exceeded the lower bound of standard moderate intensity (64%–76% for moderate-intensity physical activity [2]). Similarly, five (out of 128) and three (out of 113) activities in the same categories exceeded the threshold of moderate intensity walking cadence (100 steps/min

---

**Table 5: Number of reports with and without effort level cues by each participant.**

| Effort level cue | Total | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | P13 |
|------------------|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|
| Included         | 644   | 25 | 75 | 20 | 30 | 175| 42 | 5  | 50 | 84 | 27  | 55  | 25  |
| Not included     | 580   | 44 | 11 | 11 | 11 | 38  | 82 | 75 | 30 | 90 | 1   | 5   | 11  | 14  |
| Total            | 1224  | 69 | 88 | 131| 68 | 112 | 250| 72 | 95 | 51 | 89  | 38  | 69  | 94  |

---

**Table 6: Categories of verbalized effort level cues with the number of reports and participants (Ps), and example phrasings from utterances. The effort level cues are highlighted in bold.**

| Effort level category | Reports | Ps | Example phrasings |
|----------------------|---------|----|-------------------|
| Relaxed              | 43      | 9  | “Lying in bed, watching a retirement seminar life. Super relaxed.” – P4 |
|                      |         |    | “I’m sitting down and the salesperson is helping me try on shoes. Pretty leisurely.” – P3 |
| No effort            | 87      | 8  | “Trying to research something on my computer. No effort.” – P2 |
| No-to-Low            | 5       | 3  | “Standing in the kitchen, preparing lunch. Little to no effort.” – P1 |
| Low                  | 276     | 12 | “I’ve been eating for probably about 20 minutes. And effort level is low.” – P10 |
|                      |         |    | “Had a 15 minute walk with the dog. It was light exertion.” – P9 |
|                      |         |    | “I’ve been in the kitchen, cooking. Minimal effort.” – P7 |
| Low-to-Moderate      | 37      | 5  | “Cutting material for large raised bed garden. Light to moderate activity.” – P6 |
| Moderate             | 132     | 11 | “In the garden again and bending down, digging holes in the ground. Moderate exertion.” – P2 |
|                      |         |    | “Thoroughly wiped down stainless refrigerator and cleaned inner seal of doors, 25 minutes. Medium exertion.” – P6 |
|                      |         |    | “Preparing lunch, heating a bowl of soup up. My activity level is average.” – P10 |
| Moderate-to-Strenuous| 10      | 2  | “Walking through the airport for about a half hour, medium to heavy intensity.” – P9 |
| Strenuous            | 20      | 5  | “I moved boxes and canned goods and so on into the storage area. Expended a great deal of energy doing that. Was tired afterwards.” – P12 |
| Uncategorizable      | 44      | 8  | “Dressing and cleaning up for about 15 minutes total. Not much effort.” – P5 |

---

\(^3\)We used Nes and colleagues’ formula (211 – 0.64 * age) [87] as an estimate of age-adjusted maximum heart rate to reflect the age-related changes.
or higher for moderate activity [118]) with the measurements from activPAL and the smartwatch, respectively.

To examine how predictive the device-based intensity measurements are for the effort level, we conducted a multiple linear regression analysis using MASS [123] package in R. This method initially adds all predictors—the three device-based intensity measurements—to a model and iteratively excludes the predictors that do not make a significant contribution to the prediction, reassessing the contributions of the remaining predictors at each step. We first transformed the seven ordinal categories (No effort − Strenuous) into a continuous effort level scale (1−7, with Low as 3 and Moderate as 5) and used it as a dependent variable. For this analysis, we included 349 activities which contain the values of all three measurements. A significant regression equation (see Table 7) was found ($F(3, 345) = 15.25, p < .0001$), with an adjusted $R^2$ of .11. Although all three measurements collectively contributed to the prediction and were thus included to the final model, only walking cadence from activPAL was statistically significant ($p = .0004$). The $R^2$ value denotes that the model explains only 11% of the variance of the effort level scores. This implies that it may not be feasible to accurately predict the exact effort level score using only the device-based measurements.

### Table 7: Regression model for the effort level score, fitted from the device-based intensity measurements, $F(3, 345) = 15.25, p < .0001$, adjusted $R^2 = .11$. The positive coefficient denotes that the given parameter is positively correlated to the effort level score.

| Parameter                                | Coef. | SE  | t-statistic | p-value |
|-------------------------------------------|-------|-----|-------------|---------|
| Constant                                  | 2.00  | .73 | 2.67        | < .01   |
| Walking cadence (activPAL)                | 0.02  | .01 | 3.55        | < .001  |
| Walking cadence (Smartwatch)              | -0.01 | .02 | -1.43       | .15     |
| % of age-related maximum HR               | 0.03  | .02 | 1.71        | .09     |

* ***$p < .001$; **$p < .01$; *$p < .05$*  

5.5 Quality of Voice Recording

To investigate the potentials of activity labeling with speech input, we assessed how accurately the existing automatic speech recognition (ASR) technologies can recognize participants’ speech inputs, especially since there is prior evidence on disproportionate ASR word error rates for older adults’ voices [19, 124]. Considering the transcribed text of verbal reports by our research team as the ground-truth, we compared it with the output from two commercial ASR services, Microsoft Cognitive Speech [81] and Google Cloud Speech [38]. Using their REST APIs, we retrieved the recognized text from the audio files for each verbal report. We then calculated Word Error Rate (WER) of the recognized text using the human-transcribed text. When calculating WER, we removed punctuation and fixed contractions using NLTK (Natural Language Toolkit) [7] and Contractions Python Library [61]. On average, the Microsoft API recognized reports with an word error rate of 4.93% per report per participant ($N = 13, SD = 2.12$. This is slightly lower than 5.10% that Microsoft had reported in 2018 [133]. The Google API yielded an error rate of 8.50% per report per participant ($N = 13, SD = 2.97$). This is 3.60% higher than 4.90% that Google had officially announced in 2017 [100].

We performed an error analysis to gain insights into the potential effect these errors may have in automating the retrieval of activity labels from free form verbal reports. Specifically, we manually inspected a total of 651 verbal reports where there was a disagreement between our ground truth and the best performing ASR service. Many of the errors (70.97%; 462/651) did not affect the words capturing activity type, time, or effort level, i.e., with the local context of the verbal report someone could correctly infer this information if it was reported. Typically, errors in these reports involved filler words, conjunctions, or other details that participants provided along their activity. For example, misrecognized conjunction in the ASR output of P1’s report, “Eating lunch, Ann [should be and] about to get on a zoom call, seated, viewing on a laptop for an hour,” does not affect the coding of activity type (eating food and screen time). Interestingly, some (9.74%; 45/462) disagreements in these
reports were due to background or irrelevant speech being perhaps correctly captured by ASR but being omitted in the ground truth by our team as they were not intended to be part of the verbal report. For example, this would occur when participants were capturing sedentary activities like watching TV and the voice from the TV was also captured.

Even some of the errors involving words that captured activity type could be recoverable. These cases include errors in the verb tenses (e.g., “just came downstairs and fix [should be fixed] me some coffee...” [P8]) or compound words (e.g., “Walked up stairs [should be upstairs] to second floor...” [P6]). This was also the case for time and effort level. For example, the ASR service often made formatting errors in recognizing time (e.g., “Read a book from 6:15 until 647 [pronounced ‘six forty-seven’; should be 6:47’]” [P13]), which can be fixed referring to the local context. A disagreement in P6’s report, “… Standing, minimal [should be minimum] level of exertion,” does not affect the coding of effort level category.

If we had relied solely on the ASR output for their corresponding reports, 82 (out of 651; 12.60%) would have affected our coded activity type, time, or effort level. For example, it is challenging to extract time from the ASR output of P1’s report, “Since about 132 frozen 245 [should be 1:30 to present, 2:45…” without listening to the audio record. In addition, verbs were sometimes recognized as a totally different one, changing the original meanings in text (e.g., “I am just resting [should be dressing] after taking a shower…” [P3]). We anticipate that automated solutions may be more susceptible to some of these errors.

5.6 Participants’ Experience with MyMove
Following the week-long data collection period, we conducted debriefing interviews and guided participants to reflect on their experiences. Their responses helped us understand both strengths and challenges in using MyMove to create verbal activity reports. Participants provided feedback on their experience using MyMove interface and the smartwatch device, specifying information components for reporting, when they used prompted or voluntary methods, and preferences in using virtual vs. physical buttons. At the end of the debriefing interview, all participants agreed to be contacted for a future follow-up session in the project, acknowledging their interest in contributing to this project.

5.6.1 Reactions to the MyMove interface and smartwatch. Participants seemed to have a generally positive experience with MyMove on the smartwatch. Ten participants noted that both the interface and the smartwatch contained features that made reporting easy. For instance, P1 commented the flexibility in having multiple reporting methods (“I think it was easy enough to report, because I was allowed to, you know, report it in various ways”), and explained physical features of the smartwatch that were favorable (“the size of the screen is good for my age group, and as well as the buttons were relatively, easily to access”). P5 mentioned how the multiple modalities helped with the reporting process (“It was very efficient watch. It was nice that you could just either touch [the screen] or the [physical] buttons.”). Participants also appreciated the text on the screen, indicating the type of information components to include when recording their activity reports (“I’d remember what information I had to give you so that was very helpful for me.” [P10]).

On the other hand, participants faced challenges when interacting with the system. At the debriefing interview, six participants mentioned that the watch occasionally did not respond to their touch and that they had to click on the Record button a couple of times to start the recording. For example, P3 mentioned, “There were times when I thought I’d recorded something... it seemed like the watch was telling me I hadn’t recorded it. So I recorded it again.” We reflect on this challenge and an alternative design in Section 5.6. Some participants expressed concerns with wearing the smartwatch long-term (“I don’t know if I would want to wear this watch all the time to do it” [P6]), and with the smartwatch’s battery life (“I didn’t have any challenges with the watch, except for the fact that it ran out of battery.” [P9]).

5.6.2 Reactions to specifying information components. When reflecting on their experience with specifying the activity, timespan, and effort, many participants reacted positively. Several found that describing their activities was relatively easy: (“It was easier than I thought it would be” [P10]), (“I didn’t really have any problems with it. It was pretty straightforward” [P7]), (“If you just wanted what I was doing at the moment, I didn’t find that difficult” [P2]).

Participants also expressed challenges in specifying the level of effort required and the time taken. Seven participants reported having difficulty describing their effort level since it was hard to determine, especially for activities involving multiple tasks (“In the midst of that activity, I did something else that may have changed the amount of energy required... the effort was the hard one for me to actually document that piece of it” [P1]). To help determine effort, some participants would use physiological indicators such as breathing, muscle strain, tiredness, or even their exercise performance (“I can just look at my Strava recording and give you a time and a speed, which sort of gives you an intensity” [P4]). Six participants found specifying time components to be challenging, including recalling the amount of time taken or just remembering to add time components to the activity description. Eight participants utilized different strategies to assist with tracking activity timespan. Methods included using their memory, a device, or even writing the time down in order to remember the time (“Whenever I started an activity, I would just look at the watch before I started, and then try to record it right afterwards, so I had it right there” [P13]).

5.6.3 Situations for using voluntary and prompted methods. In addition, participants described situations in which they would use voluntary and prompted reporting. We found that each method had unique advantages, including having the freedom to report voluntarily at any time and being aware about reporting activities due to prompted notifications. Some participants appreciated the pings because they served as a reminder to record the activity right away or after finishing the activity. P8 commented how “it was good to have it, because it reminded you that maybe you hadn’t recorded what you were doing,” and P1 stated, “Had it not been for the watch’s alert, I’m not quite sure that I would have captured that information as well.” When asked what participants disliked about receiving ping notifications, six participants said that pings were delivered during inopportune moments (“There were a couple of times when I just couldn’t answer the ping. That was in a meeting or something” [P7]). Participants also reported that pings seemed unnecessary for redundant activities (“I was... reporting the same thing all the time”)
We were pleasantly surprised by the high adherence and engagement with older adults using speech on a smartwatch. Going forward, we aim to examine the feasibility of collecting in-situ activity labels using a wake-up functionality associated with using the virtual buttons.

13.45 reports per day on average even though the compensation limit the scope of what activities are report-worthy because our goal was to examine the feasibility of collecting in-situ activity labels with older adults using speech on a smartwatch. Going forward, such comprehensive data capture may not be necessary; we expect that the burden of capturing activity labels would be reduced when we have a fixed set of targeted activities (e.g., walk, gardening, golf, yoga) that require labeling and better mechanisms for estimating activity timespan (e.g., by automatically detecting abrupt changes in sensor data) and effort level (e.g., by leveraging heart rate data).

6 DISCUSSION

In this section, we first reflect on several aspects that are related to the feasibility of MyMove in facilitating data collection with older adults, such as engaging older adults in activity labeling, using the verbal reports as an information source for activity labeling, and capturing older adults’ activities in a comprehensive manner. We also discuss how our findings and the data older adults collected using MyMove can be used toward creating personalized activity trackers that attune to idiosyncratic characteristics of individual users and their unique needs. We then discuss limitations in our study that may affect the generalizability of our findings as well as future work.

6.1 Engaging Older Adults in Activity Labeling

We were pleasantly surprised by the high adherence and engagement of our participants in the one-week data collection. On average, they wore the smartwatch for 11.6 hours and activPAL for nearly 24 hours every day. Furthermore, our participants submitted 13.45 reports per day on average even though the compensation was not tied to the number of reports. Given that the most challenging aspect of an EMA study is its high data capture burden and frequent interruptions [120], we believe that this is a promising outcome. Participants’ positive feedback on MyMove indicates that the system itself may have contributed to this high adherence; it provided flexible ways to capture data using speech, including voluntary and prompted reporting methods as well as simplified data capture flow and UI. Even though all but one experienced the smartwatch for the first time through our study, all participants could use MyMove without much trouble.

In debriefing, however, most participants stated that collecting data in this manner would not be sustainable, and the one-week duration would probably be the maximum they could continuously engage at this level. As is common with other ESM studies, our study imposed a high burden on the participants, for example, having them consciously think of activity type, start/end time, and effort level when they receive hourly notifications. In our study, we did not limit the scope of what activities are report-worthy because our goal was to examine the feasibility of collecting in-situ activity labels with older adults using speech on a smartwatch. Going forward, such comprehensive data capture may not be necessary; we expect that the burden of capturing activity labels would be reduced when we have a fixed set of targeted activities (e.g., walk, gardening, golf, yoga) that require labeling and better mechanisms for estimating activity timespan (e.g., by automatically detecting abrupt changes in sensor data) and effort level (e.g., by leveraging heart rate data).

6.2 Leveraging Verbal Reports as an Information Source for Activity Labeling

It was encouraging to see that all of the 1,224 verbal reports are valid and that researchers could transcribe and understand all of them. Although some participants, in the debriefing interview, mentioned that they accidentally triggered the recording, it seemed that they were able to cancel it or the recording was timed out and thus erased. This demonstrates that, despite the challenges older adult participants faced with the unfamiliar technologies, they still could successfully submit valid reports using our novel data collection approach. Furthermore, the word error rates by two state-of-the-art automatic speech recognition systems were relatively low: 4.93% with Microsoft Cognitive Speech and 8.50% with Google Cloud Speech. We were reassured that Microsoft Cognitive Speech’s error rate on our older adult participants is lower than what Microsoft had reported in 2018 (5.0%). Nonetheless, these numbers serve merely as anecdotal evidence. Our small sample is not representative of older adults; all participants were native English speakers in the US and none of them identified as disabled. Of course, speech as an input modality, can be advantageous for many disabled people such as blind individuals [4, 47, 136] and those with upper limb motor impairments [50, 76]. However, it is still limited for dysarthric [24, 76], deaf [35], and accented [99] speech as well as for low resource languages and noisy environments, all being active areas of research.

With advances in speech recognition, we believe that we could leverage for many older adults their verbal reports as a reliable data source in an automated manner. This opens up an opportunity for automatically extracting user-generated activity labels (type and semantics).

That said, inferring quantitative or ordinal data from free-form text is not a trivial problem, as in the case of the effort level coding. As such, data such as effort level would be better off if they are collected in a structured way (e.g., have people select from a scale or predefined categories). This would require new UI & interaction designs e.g., leveraging a simple touch interaction on a smartwatch or predefined voice commands.

6.3 Comprehensive Capturing of Older Adults’ Activities

Even though understanding daily activities of older adults was not the main goal of our study, the collected data seem to cover more classical types of activities older adults perform while reflecting recent trends. We consider this as additional evidence to demonstrate the feasibility of in-situ data collection with older adults.

The types of activities emerging from our participants’ reports overlap with most of the activities reported in prior literature, though their naming/grouping may not be fully aligned. For example, an interview study with 516 German older adults conducted in 1996 identified 44 types of activities and grouped them into...
eight categories—Personal Maintenance, Instrumental ADLs, Reading, Television, Other Leisure, Social Activities, Paid Work, and Resting [49]. While these activities appeared in our dataset, we categorized them differently, for example, the Reading activity type under Hobby/leisure and many of the Instrumental ADL activities under Housekeeping. We note that there is not a clear consensus on how to group activities, so it will be important to preserve both raw labels and coded categories for future reference.

New activities have also emerged as digital technologies have advanced over time. For example, our Screen time category includes “Computer,” “Mobile device,” and “TV,” whereas prior studies conducted in 1982 [83] and 1998 [49] had the “Watching TV” as the top-level category (without the notion of screen time). Screen time has absorbed other activities that would have been categorized differently in the past. For example, many of our participants read news on the internet, rather than a printed paper. Screen time also commonly appeared with other activity types, such as social activity (e.g., a video call) and exercise (e.g., online yoga session), which may have transitioned from in person to remote during the pandemic. Labeling systems are bound to evolve as technology changes the way people achieve different functions. Reflecting these changes in designing activity labeling systems, there may be value in having multiple dimensions, such as device type, posture, semantics, to better characterize the captured activities.

6.4 Capturing Non-Exercise Stepping as a Meaningful Activity for Older Adults

Researchers have advocated the importance of promoting non-exercise physical activities (free living activities that involve light and moderate physical activities, such as gardening, laundry, cleaning, or casual walking) for older adults [107, 111, 112]. Recent evidence shows that non-exercise physical activities are positively associated with longevity [26, 41] and cardiovascular health [26], suggesting that any activity is better than no activity. However, it is challenging to capture non-exercise physical activities using recall-based methods or sensor-only methods [111]. For example, the interview study by Horgas and colleagues identified a generic “Walking” category [49]. Studies that leveraged accelerometer sensors often use statistically determined thresholds for sensor movements and bout duration to categorize activities by intensity level, treating sporadic and light activities as non-exercise (e.g., [14, 66]). However, applying uniform thresholds may ignore individualized characteristics [101, 109] and there are no standardized thresholds validated for older adults [39]. In addition, most of such studies do not capture user-generated context, potentially important details to understand what people did.

In our study, 14% of verbal reports (171/1224) contained non-exercise stepping with context. We assert that our in-situ data collection method enabled participants to capture more subtle, light-intensity lifestyle activities (e.g., walking around home) that would have not been captured otherwise. Older adults’ non-exercise stepping activities, which tend to be in slow gaits, are difficult to detect accurately with current waist and wrist-worn accelerometers. The context information captured in verbal reports may be a valuable supplement for device-based monitoring to support the training of person-specific classifiers for non-exercise physical activities.

6.5 Personalizing Activity Trackers

Even though our participant sample was relatively homogeneous and geographically constrained to the same area, we observed high variation in the types of activities captured depending both on the participant and on other factors, such as the time of the year. We identified many implications from our findings for the design of personalized activity tracking systems with older adults. When automating the tracking of these activities from the sensor data, researchers can potentially leverage some of the existing datasets from younger adults (e.g., WISDM [129, 130], UCI-HHAR [115], and ExtraSensory [119]) to pre-train models for higher level activities, such as sitting, standing, and walking that tend to be common among people regardless of their age. However, preliminary results indicate that model adaptation is necessary as models trained on younger adults’ sensor data tend to perform worse on older adults [28]. In addition, the diversity in the activity types and semantics among the older adults in our study calls for model personalization beyond adaptation, as one-to-one mapping between older adults’ activities and those available in the datasets (from younger adults) may not be possible. We could employ novel model personalization methods like teachable machines [25, 46, 54, 68], which leverage advances in transfer learning [90] and meta learning [30, 70]. Systems like MyMove could play a critical role in facilitating this personalization process by supporting older adults and other underrepresented populations in fine-tuning the models in activity tracking applications with their own data, so that the applications can reflect their idiosyncratic characteristics.

6.6 Usability Challenges with Smartwatch’s Low-power Mode

While our participants had generally positive experiences with MyMove and the smartwatch, six participants occasionally experienced the watch being unresponsive (Section 5.6.1). We suspect that this was caused during Wear OS’s low-power mode (also known as ambient mode). When a user is not interacting with their watch for a while, Wear OS automatically enters into the low-power mode, dimming the watch display to save the battery. In this low-power mode, MyMove’s button icons and labels were still visible in low contrast. To make the virtual buttons interactive, participants needed to “wake up” the screen by tapping anywhere on the watch display. Alternatively, they could push the physical button to start the recording without needing to wake up the screen.

We showed the buttons and icons in low contrast during the lower-power mode because we wanted to use them as a visual reminder to encourage data capture. However, the low-power mode was, in hindsight, an unfamiliar concept to participants, especially those who are new to a smartwatch: some participants thought that they could interact with the visible button in the low-power mode. Instead of showing the button icons and labels in low contrast, hiding them completely might have been a better design to avoid confusion, which is an interesting design tradeoff we learned from this study.

6.7 Limitations and Future Work

In this section, we discuss the limitations of our study that could impact the generalizability of our findings. Although we aimed to
recruit participants with diverse backgrounds, our participants are not representative samples of older adults. They were all highly educated (e.g., having a college degree or above), had high-baseline technical proficiency (e.g., being able to use a Zoom video call), and did not have speech, hearing, motor, movement, or cognitive impairments. While this work is just a first step toward designing and developing inclusive activity tracking systems, we believe it is important to conduct a follow-up study with older adults with different educational backgrounds, health conditions, and technical proficiencies. This would help us extend our understanding of the strengths and limitations of in situ data collection with speech on a smartwatch. As discussed above, we anticipate that this modality can be advantageous for people that were not captured by our small sample such as those who are blind or have low vision, as speech input can be more efficient for this user group [4, 136]. However, it may not be inclusive of dysarthric, deaf, and accented speech, especially if the goal is automatic extraction of the activity labels. Recent speech recognition personalization efforts like Google’s Project Euphonia [40, 74] are promising. Similarly, efforts that attempt cross-lingual knowledge transfer in speech recognition from high- to low-resource languages (e.g., [58, 116]) can make speech input more inclusive. Even then, the challenge of automatically extracting activity labels, timing, and effort levels from verbal reports remains. Information extraction from unstructured reports is an active area of research in natural language processing (e.g., processing medical verbal or written reports [57, 85]). Similar to the healthcare context, we could leverage transfer- and meta-learning techniques to deal with the lack of training data. More so, in contrast to healthcare, we also have an opportunity to shape (i.e., via design and personalization) the user interactions with the activity trackers. Thus, we can influence the structure and vocabulary in the reports to meet the algorithmic capabilities halfway e.g., by optimizing across flexibility, efficiency, and effectiveness for both users and algorithms.

Our study preparation (e.g., dropping off & picking up study equipment) and study design provided more face-to-face time with the participants than a typical remote deployment study. This provided a chance for older adult participants to ask questions and troubleshoot issues. Thus, these repeated interactions may have contributed to forming rapport between participants and researchers, which in turn, could have contributed to the high engagement. We had two onboarding sessions with the 4-day adaptation period in between. During the 4-day adaptation period, participants became used to wearing and maintaining (e.g., charging) the devices, and were ready to collect data on Day 5 of the study. Some participants explicitly mentioned that the tutorial and the adaptation period were critical for their engagement. For example, P5 commented, “Giving me a few days to get used to the equipment and how it worked... Making sure I had plugged in and make sure that I was charging, how to record, and I thought that was really good. And just how to give the reports, I think the orientation was very helpful as well.” While we believe that giving a good tutorial before the actual experiment is important, we acknowledge that our particular approach may not scale. In addition, the study compensation and participants’ interest in contributing to a research project may also have affected participants’ engagement, although it is common to incentivise participants in ESM studies.

We note that we did not collect information on medication use of participants. Medications, such as \( \beta \)-blockers, can influence heart rate and may blunt the response to higher intensity exercise, resulting in lower heart rate measurements. As such, the intensities we recorded during Strenuous activities may be not accurately reflect the degree of vigor with which the participant was being active, resulting in the percentage of \( HR_{max} \) that were closer to the low intensity activities. Future work should consider the incorporation of participants’ medication information to further validate heart rate intensities, especially for high-intensity activities.

We chose a smartwatch as an only means to collect verbal activity reports and to deliver notifications. In the future, we can leverage other “smart” devices for more comprehensive and accurate data collection. For example, when the TV is on and the person is nearby (without much movement), we can infer that the person is watching TV. In addition, we can leverage the speech input capability of other devices. For example, a person can report their activities using a smart speaker that is becoming more prevalent (the speaker can even play the recording back to the person). Similarly, a person can record their activities from their smartphone, tablet, laptop, or desktop as all these devices are equipped with a microphone. Since these devices have larger display than a smartwatch, people can view or edit data they captured elsewhere (e.g., a smartwatch or smart speaker) from these devices.

In our study, we did not provide feedback other than the number of reports participants submitted on a given day. However, in the debriefing interview, half of our participants reported that they became more aware of the activities they performed and how they spent the time. Also known as the “reactivity effect,” this is a well-known phenomenon in behavioral psychology [86]. We believe that our approach to collecting in situ data can serve a dual purpose of activity labeling and self-monitoring; the latter can be further augmented through providing informative and engaging feedback—for example, showing how much they have been sitting, working out, gardening—and people may be more motivated to engage in desirable activities while capturing data (labels).

7 CONCLUSION

In this work, we examined the feasibility of collecting in situ activity reports with older adults, with the ultimate goal of developing personalized activity tracking technologies that better match their preferences and patterns. We built MyMove, an Android Wear reporting app. Considering older adults as the main user group, we streamlined the data capture flow and leveraged the flexible speech input on a smartwatch. Through a 7-day deployment study with 13 older adults, we collected a rich dataset including older adult participants’ verbal reports, the sensor data from a smartwatch and a thigh-worn activity monitor, and participants’ feedback from the debriefing interviews. Our results showed that participants were highly engaged in the data collection. They submitted a total of 1,224 verbal reports. Additionally, the wear time of the smartwatch (11.6 hours/day) and thigh-worn activity monitor (23.3 hours/day) was very high. Examining the verbal reports further, we found that all of them were valid, that is, a researcher could understand and transcribe them. Moreover, verbal reports could be transcribed with state-of-the-art automatic speech recognition systems with
acceptable error rates (e.g., 4.93% with Microsoft Cognitive Speech). These results, taken together, indicate that our novel data collection approach, realized in MyMove, can facilitate older adults to collect useful in-situ activity labels. Going forward, we are excited to continue our endeavors towards building personalized activity tracking technologies that further capture meaningful activities for older adults.

ACKNOWLEDGMENTS
We thank our study participants for their efforts, time, and feedback. We also thank Catherine Plaisant, Yuhan Luo, and Rachael Zehrung for helping us improve the tutorial protocol. We are also grateful for Bonnie McClellan and Explorations On Aging for helping us recruit study participants. This work was supported by National Science Foundation awards #1955568 and #1955590.

REFERENCES
[1] Barbara E. Ainsworth, William L. Haskell, Arthur S. Leon, David R. Jacobs, Henry J. Montoye, James F. Sallis, and Ralph S. Paffenbarger. 1993. Compendium of Physical Activities: Classification of Energy Costs of Human Physical Activities. Medicine & Science in Sports & Exercise 25, 1 (1993), 71–80. https://journals.lww.com/acsmsmfulltext/1993/01000/Compendium_of_Physical_Activities_classification.11.aspx
[2] American College of Sports Medicine. 2021. ACSM’s Guidelines for Exercise Testing and Prescription (11 ed.). Lippincott Williams & Wilkins, Philadelphia, PA, USA, 548 pages.
[3] Michael J. Annear, Grant Cushman, and Bob Gilfold. 2009. Leisure time physical activity differences among older adults from diverse socioeconomic neighborhoods. Health & Place 15, 2 (2009), 482–490. https://doi.org/10.1016/j.healthplace.2008.09.005
[4] Shiri Azenkot and Nicole B. Lee. 2013. Exploring the Use of Speech Input by Blind People on Mobile Devices. In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (Bellevue, Washington, USA). (ASSETS ’13) Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/2513385.2513440
[5] Maxim Bakaev. 2008. Fitts’ Law for Older Adults: Considering a Factor of Age. In Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems (Porto Alegre, RS, Brazil) (IHC ’08) Sociedade Brasileira de Computação, BRA, 260–263.
[6] Paul B. Baltes and Margret M. Baltes. 1990. Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, Cambridge, United Kingdom.
[7] Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python (1st ed.). O’Reilly Media, Inc., Sebastopol, CA, USA.
[8] Niall Bolger, Angelina Davis, and Edkol Ramos. 2003. Diary Methods: Capturing Life as It Is Lived. Annual Review of Psychology 54, 1 (2003), 579–616. https://doi.org/10.1146/annurev.psych.54.101601.140300
[9] Marc Bonnefoy, Sylvie Normand, Christiane Pachiaudi, Jean René, and Marie-Louise Bélanger, and Janice L. Thompson. 2011. Objectively Measured Physical Activity in a Diverse Sample of Older Urban UK Adults. Medicine & Science in Sports & Exercise 43, 4 (April 2011), 647–654. https://doi.org/10.1249/MSS.0b013e3181f36196
[10] Luigi De Russis and Fulvio Carmeli. 2009. On the impact of dysarthric speech on contemporary ASR cloud platforms. Journal of Reliable Intelligent Environments 5, 3 (2019), 163–172. https://doi.org/10.1007/s12880-018-00085-y
[11] Cathie Finocchio, Jodie Forlizzi, and Anind Dey. 2012. Considerations for Technology Contingent ASR. In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS ’12 ACM Press, New York, New York, USA, 33. https://doi.org/10.1145/2384916.2384923
[12] Wijakvadi Dwiwedi, Jaina Gandhi, Raj Parikh, Merijke Coenraad, Elizabeth Bon, and Hernisa Kacorri. 2021. Exploring Machine Teaching with Children. In 2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, IEEE, Washington, DC, USA, 11 pages.
[13] Paul A. Gardiner, Bronwyn K. Clark, Genevieve N. Healy, Elizabeth G. Eakin, and Janine L. Thompson. 2011. Objectively Measured Physical Activity in Older Adults: A Doubly Labeled Water Study. Journal of the American Geriatrics Society 49, 1 (Jan. 2001), 28–35. https://doi.org/10.1111/j.1532-5415.2001.49006.x
[14] Gunnar Borg. 1990. Psychophysical Scaling with Applications in Physical Work and the Perception of Exertion. Scandinavian Journal of Work, Environment and Health 16, SUPPL. 1 (1990), 55–58. https://doi.org/10.5271/swehe.1815
[15] Gunnar Borg. 1998. Borg’s Perceived Exertion and Pain Scales. Human Kinetics, USA.
[16] Gregory R. Boss and J. Edwin Seegmiller. 1981. Age-Related Physiological Changes and Their Clinical Significance. The Western Journal of Medicine 135, 6 (Dec. 1981), 434–40. http://www.ncbi.nlm.nih.gov/pubmed/7336713http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1273316
[17] Lawrence R. Brawley, W. Jack Rejeski, and Abby C. King. 2003. Promoting Physical Activity for Older Adults: the Challenges for Changing Behavior. American Journal of Preventive Medicine 25, 3 Suppl 2 (Oct. 2003), 172–83. https://doi.org/10.1016/S0741-8485(03)00162-8
[18] Matthew P. Buman, Eric B. Helker, William L. Haskell, Leslie Pruitt, Terry L. Conway, Kelli L. Cain, James F. Sallis, Brian E. Saelens, Lawrence D. Frank, and Abby C. King. 2010. Objective Light-Intensity Physical Activity Associations With Rated Health in Older Adults. American Journal of Epidemiology 172, 10 (Nov. 2010), 1155–1165. https://doi.org/10.1093/aje/kwp249
[19] Eli Carmeli, Hagar Patish, and Raymond Coleman. 2003. The Aging Hand. The Journals of Gerontology: Series A 58, 2 (Feb. 2003), M146–M152. https://doi.org/10.1093/gerona/58.2.M146
[20] Marta E. Cecchinato, Anna L. Cox, and Jon Bird. 2017. Always Online? User Experience of Smartwatch and Their Role within Multi-Device Ecologies. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 3557–3568. https://doi.org/10.1145/3025453.3025558
[21] Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. 1991. Expressive Richness: A Comparison of Speech and Text as Media for Revision. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’91) Association for Computing Machinery, New York, NY, USA, 21–26. https://doi.org/10.1145/108844.108848
