Evaluating the use of voice-enabled technologies for ground-truthing activity data

Przemyslaw Woznowski, Alison Burrows, Pawel Laskowski, Emma Tonkin, and Ian Craddock
Faculty of Engineering, University of Bristol, Bristol, BS8 1UB, UK

***Abstract*—Reliably discerning human activity from sensor data is a nontrivial task in ubiquitous computing, which is central to enabling smart environments. Ground-truth acquisition techniques for such environments can be broadly divided into observational and self-reporting approaches. In this paper we explore one self-reporting approach, using speech-enabled logging to generate ground-truth data. We report the results of a user study in which participants (N=12) used both a smart-watch and a smart-phone app to record their activities of daily living using primarily voice, then answered questionnaires comprising the System Usability Scale (SUS) as well as open ended questions about their experiences. Our findings indicate that even though user satisfaction with the voice-enabled activity logging apps was relatively high, this approach presented significant challenges regarding compliance, effectiveness, and privacy. We discuss the implications of these findings with a view to offering new insights and recommendations for designing systems for ground-truth acquisition 'in the wild'.**

***Index Terms*—Ground-truth acquisition; self-annotation; voice-logging; smart-watch app; smart-phone app; activity logging.**

I. Introduction

Recent advances in smart home technologies have fuelled an interest in providing desirable context-aware services to end users. There are several key areas of application for these technologies such as healthcare (e.g. [1]–[3]), where correctly recognising and responding to human activity can be critical. Monitoring people's functional status and providing appropriate interventions requires streamed sensor data to be processed in real-time and with a high level of certainty, using machine learning techniques [4]. To ensure this is the case, activity recognition (AR) algorithms have to be trained and validated. AR techniques to date have successfully worked on scripted or pre-segmented activity sequences. However, the challenge of acquiring ground-truth or benchmark data from deployed smart home technology remains unsolved.

Prior efforts to properly and consistently label naturalistic activity data have relied on observational techniques, including direct (e.g. researchers following participants [5]) and indirect approaches (e.g. video cameras [6]–[8]). However, these techniques can be intrusive for the participants and time-consuming for the researchers. Self-reporting approaches (see for example [9], [10]) present more acceptable solutions for the users but may incur a trade-off in terms of accuracy and reliability. Overall, these approaches are not scalable given that the ground-truth data is annotated manually and offline. It is imperative that ground-truth acquisition solutions provide information that is useful and usable for machine learning, while also meeting essential user acceptance and compliance criteria for real-world deployment.

Fig. 1. Participants logging their activities via smart-phone (left) and smart-watch (right).

In many respects, self-reporting approaches to capturing ground-truth data can draw on lessons learned from research into various forms of self-tracking [11]–[15]. However, if we consider that there are a number of barriers to the wider adoption of self-tracking technologies [16], the challenge to engage users in self-reporting activity data merely for ground-truth purposes can seem even greater. The range of people ideally involved in ground-truthing exercises would necessarily mean different and perhaps insufficient motivation to consistently and reliably log activity data. For example, a survey of people's health tracking practises showed that people with chronic conditions are significantly more likely to track a health indicator or symptom [17]. Also, from the perspective of acquiring rich ground-truth data, people should report activities and locations frequently and in as much detail as possible. This places considerable capture burden on the users, which is simply not feasible for long periods of time.

In this paper, we sought to understand if off-the-shelf speech recognition technology could be suitable for self-reporting ground-truth data. We were particularly interested in understanding the natural frequency with which people logged activities, as well as the granularity of logged activities. This paper offers new insights for developing systems for ground-truth acquisition 'in the wild', using voice-enabled technologies.

II. Method

A. Ground-truth apps

Our decision to investigate the use of speech recognition was based on participant feedback from pilot studies that investigated the use of other self-report (e.g. pen and paper diaries, ontology-driven smart-phone apps) and observational (e.g. head-mounted cameras) approaches. We felt a voice-based approach would be acceptable and reduce capture burden on the users. We built a voice-based smart-phone and similar smart-watch version of an Android app, following guidelines from machine learning colleagues to incorporate functions of logging activities, their duration and location. Activity and location were logged using speech recognition, and duration was obtained from recording start and end time of an activity. In each version of the app, users had three menu choices: Log (log the start of a new activity); Ongoing (look up and terminate or delete ongoing activities); and Finished (look up with an option to delete terminated activities).

B. Participants

Participants were recruited from our institution via an email invitation to participate. We recruited 12 participants (3 female), aged 18-44. Of these participants, none had previous experience of using a smart-watch, but all used a smart-phone; 5 used life-logging apps. Five participants were native English speakers.

C. Data collection

Participants were asked to use either the smart-watch or the smart-phone version of the app for a period of two days, followed by the other version of the app for a further two days. During this time, participants were encouraged to log their activities and location of those activities while going about their everyday lives. At the end of each two day evaluation period, participants completed the same questionnaire to provide feedback on the tested app. This questionnaire contained the System Usability Scale (SUS) [18], as well as two open-ended questions to list the three best and three worst things about using that version of the app. We selected this instrument to assess each version of the app because, in addition to being easy to administer, it has been shown to have good reliability and validity measures with a minimum sample size of 12 participants [19].

After participants had evaluated both versions of the app, they were asked to fill out a final questionnaire. This final questionnaire asked participants about their preferred version of the app, whether they felt their logged activities and locations reflected reality, and if there had been activities or locations that they chose not to log. Participants were also given the opportunity to make suggestions for improving the apps and for improving the logging of activities generally. At the end of the study, all participants had used both versions of the app and completed three questionnaires each.

III. Quantitative results and analysis

Data collected during the experiments were stored on a database and subsequently exported to a collection of CSV files for detailed analysis. These comprised the following information: activity name, location, start and end time.

A. Number of logs

A total of of 296 activities were logged with the smart-watch, which corresponds to an average of just under 25 activities per user across two days. In reality, the most prolific participant logged 52 activities across two days, while the least prolific participant logged only 11. For the total number of days across all participants (24 days), there was only a single day in which no activity was logged with the smart-watch and only three activities had no location information specified.

For the smart-phone, 102 activities were recorded in total across all participants, averaging 8.5 activities per participant over a two day period. The number of logs ranged from a maximum of 21 to only 3 activities logged in 48 hours. The majority of participants stopped logging activities after the initial introduction to the app, resulting in a third of the first days not containing a single activity logged. Compliance continued to decrease in the second day, when 7 out of 12 participants did not log a single activity. Only one activity had no location information specified.

B. Activity data

There were no clear differences in type of activities logged with ‘Working’, ‘Eating’ and ‘Walking’ being top three for both devices (see Table I). In Table I, several of the ‘Uncategorised’ activities are entries that were incorrectly translated from speech (29 activities for the smart-watch). This can be partly explained by the fact that over half the participants were non-native English speakers. However, there are no ‘Uncategorised’ entries in the smart-phone data, which is most likely due to the better quality of smart-phone microphones and the fact that the process to delete erroneous entries was easier given the larger interface of the smart-phone.

Participants were much more likely to forget to log activities via smart-phones – only 5 out of 12 participants logged data on both days using the smart-phone, compared to 11 using the smart-watch. Yet, analysis of logged activities showed that participants tended to report high-level activities (e.g. ‘Cooking’) rather than granular activities (e.g. ‘Peeling vegetables’, ‘Cutting vegetables’) for both devices. The majority of the most frequently reported activities are the same across both devices, which shows that the type of logging device used did not influence the type of activities logged.

C. Location data

As with activities, participants attributed their own labels to log location. For the purpose of this analysis, we grouped locations into six most commonly logged categories. Smart-watch distribution of logs was as follows: Kitchen (15.2%), Bedroom (9.8%), Home (8.78%), Bathroom (5.74%), Out of Home (23.31%), Work (21.62%), Unrecognised (14.53%), and no location given (1.01%). Smart-phone distribution of logs was as follows: Kitchen (14.71%), Bedroom (10.78%), Home (30.39%), Bathroom (1.96%), Out of Home (14.71%),

Activity	SW	SP	Activity	SW	SP
Shower	0	5 (~5%)	Getting dressed	9 (3%)	0
Relaxing	4 (~1%)	2 (~2%)	Watching TV	10 (~3%)	5 (~5%)
Cleaning	2 (~1%)	0	Sitting	11 (~4%)	0
Cycling	3 (1%)	0	Drinking coffee/tea	13 (~4%)	6 (~6%)
Washing up	3 (1%)	1 (~1%)	Sleeping	13 (~4%)	13 (~13%)
Food prep.	7 (~2%)	0	Cooking	14 (~5%)	8 (~8%)
Shopping	7 (~2%)	2 (~2%)	Going out	14 (~5%)	7 (~7%)
Toilet	8 (~3%)	0	Walking	33 (~11%)	15 (~15%)
Meetings	8 (~3%)	3 (~3%)	Eating	37 (~12%)	15 (~15%)
Out of bed	8 (~3%)	4 (~4%)	Uncategorised	41 (~14%)	0
Brushing teeth	8 (~3%)	2 (~2%)	Working	43 (~15%)	14 (~14%)

TABLE I
TOTAL NUMBER OF ACTIVITIES LOGGED VIA SMART-WATCH (SW) AND SMART-PHONE (SP).

Work (13.73%), Unrecognised (12.75%), and no location given (0.98%). Although percentage of locations logged was similar for both devices, there were some notable differences between certain locations. Logs for Home were significantly higher with the smart-phone, while logs for Bathroom, Out of Home, and Work were comparatively higher with the smart-watch. Bathroom was the location where the fewest activities were logged, though it was interesting to note that there were roughly three times more logs here using the smart-watch than the smart-phone. Only three participants consistently preferred to use coarse location names such as ‘Home’ or ‘Work’ over detailed, room-level information.

IV. QUALITATIVE RESULTS AND ANALYSIS

The results and analysis presented in this section are derived from the app feedback questionnaires, as well as the final questionnaire. The following subsections address key themes that emerged from analysis of this qualitative data.

A. *Usability of the apps*

We calculated the SUS scores using the method described in [18], obtaining an average score of 76.68 (good usability) for the smart-phone app and 68.75 (average usability) for the smart-watch app. However, there was much more variation in individual SUS scores for the smart-watch app, with a standard deviation of 17.01. In contrast, standard deviation for the smart-phone app was just under half that value at 9.62. Interestingly, although the highest individual scores were similar for the smart-phone and smart-watch versions (90 and 87.5, respectively), the lowest score for the smart-watch app was significantly lower at 32.5 compared to 60 for the smart-watch. Reasons for this discrepancy are presented below.

B. *User preferences*

When asked which version of the app they preferred, 6 participants chose the smart-phone version and 5 participants chose the smart-watch version; 1 participant had no preference. Participants who preferred using the smart-phone cited the larger interface, better Internet connection, more accurate voice recognition, and the need to carry fewer devices as reasons for their choice. One reported drawback of using the smart-phone version was that people sometimes forgot to carry their phone, especially when they were at home. Participants also said they forgot to report their activities when using the smart-phone, because of the familiarity of the device. Conversely, the majority of participants who preferred using the smart-watch (4 out of 5 participants) said that *the device itself served as a reminder for them to log activities*. These participants valued having a dedicated device, which facilitated hands-free activity logging. Although wearing the speech-recording device seemed to have several advantages, we observed that for some participants the wrists were already a busy area where they already wore watches, jewellery, and activity trackers (see Fig. 2). Negative feedback specific to using the smart-watch app pertained to the fact that the smart-watch did not have a built-in stand alone WiFi module and, therefore, required the smart-phone to be within close range.

Overall, participants felt both apps were easy to use and liked the simple interaction afforded by the speech recognition. However, the speech recognition was fallible, in particular for non-native English speakers. When speech recognition recorded incorrect entries, participants felt frustrated at the absence of an option to edit the data manually. This was especially problematic for location data, which was recorded at the end of an activity, because the only available option was to delete the entire activity log.

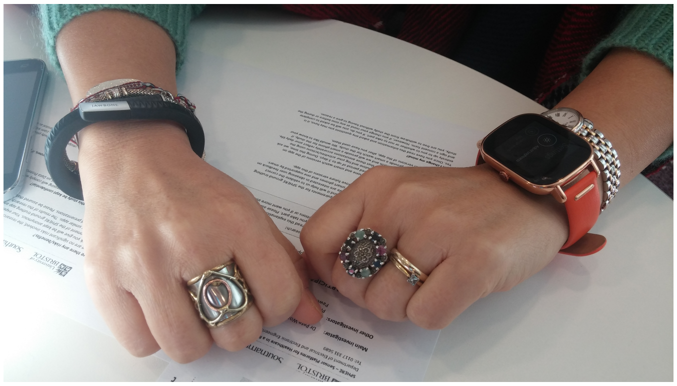

Fig. 2. Participant wearing their Jawbone activity tracker and Asus ZenWatch 2 smart-watch.

C. Feasibility of voice-enabled ground-truth acquisition

Of the 12 participants, 7 felt that the activities and/or locations recorded did not reflect reality. Reasons given for this included voice recording errors, forgetting to start or finish recording an activity, and ambiguous descriptions of activity or location data. Participants reported difficulty recording data in noisy environments and around children. Participants were unable to log activities when their devices were not connected to the Internet as this is a requirement for the speech recognition software. A few participants commented on the granularity of the logged data; one participant explained that *"the rate at which the activities are being carried out is faster than the data logging procedure — for example, drinking a glass of water requires less time than logging it — thus, some of the short activities are not logged"*. Participants tended to group short activities under a high-level label for the activity.

Nine participants noted that there were activities and locations they chose not to log. Using speech recognition in communal places was perceived as embarrassing, either because they were typically quiet places or because of the presence of others. Some participants felt *there were activities that were too personal or confidential to share* and chose not to log them.

D. Suggested improvements

Most participants felt the logging process would be improved by allowing entries to be edited, even though this would make the interaction more complex. In the case of the smart-phone version, by far the most popular suggestion was *to include a text entry option*, which would be more discrete to use in communal spaces and would provide added privacy. Participants mentioned that a future development of the apps could be to allow people to select from most common activities and locations, as well as *to pull location data from Google Maps so the system could be tailored to familiar locations* such as home or work. Some participants also requested the option to add supplementary details about logged activities. Incorporating reminders to log new activities and to update ongoing activities was the second most frequently proposed improvement to both apps. A few participants reported that they would like *to see their logged data in daily or weekly charts, perhaps using additional strategies such as leaderboards* to encourage people to log data more often. Given that the smart-watch had to be used in close proximity to the smart-phone for WiFi connectivity, one participant actually reported using both devices concurrently and suggested that the option to use multiple devices should be available in future iterations.

V. Recommendations

Based on our findings, we offer recommendations to researchers and developers wishing to exploit the potential of voice-enabled technologies to support ground-truth data acquisition.

1) *Use a dedicated device whenever possible and consider providing complementary devices.*
   We found that having a dedicated device served as a reminder for people to log their activities, given that the number of activities logged was greater for the smart-watch than for the smart-phone. However, not everyone will accept wearing a device on their wrist, particularly since some people already wear other accessories here. One possible solution is to provide alternative logging devices that can be used alone or in combination with other devices.
2) *Do not expect users to log fine-grained or personal activities.*
   We found that logging short activities could take more time than performing those activities, which resulted in participants not bothering to log them. In addition, some activities were considered too personal to log and, therefore, it may not be feasible to expect all activities to be logged. While there is a clear need to be realistic and respectful of user's privacy, there is also an opportunity to investigate alternative strategies to obtain ground-truth for such situations.
3) *Voice-enabled logging should be used in combination with other logging modalities.*
   We found that voice-enabled technologies were generally considered suitable for private spaces such as the home, but could be embarrassing to use in public areas. Also, the speech-to-text recognition was prone to error, which resulted in incorrect entries. Enabling users to log activities through speech recognition and giving them an option to edit entries via written text is one way to remove these barriers and provide more control to the users.
4) *Users expect more intelligence from smart technologies to capture location data.*
   We found that participants were sometimes frustrated by the need to log location, given that this information is frequently available through GPS-enabled services. However, indoor location information is harder to capture accurately and remains unresolved from a ground-truth perspective. There is an opportunity here for machine learning techniques to be applied to learn and eventually predict locations in which certain activities occur.
5) *Consider using motivational strategies to encourage people to log activities more frequently.*
   We found that the average number of logged activities for both devices was relatively low and participants' compliance dropped over time. There are a number of motivational strategies available, some of which were mentioned by participants in our study (e.g. prompts, visualising logged data, leaderboards), which could be explored as mechanisms to sustain or increase compliance with capturing ground-truth activity data.

VI. Conclusion and future work

We investigated the use of emerging (smart-watch) and widespread (smart-phone) voice-enabled technologies to provide more user-friendly solutions to the challenge of ground-truth acquisition for activity data. In a study with 12 participants, we found that participants only logged on average 4 and 12 activities per day using the smart-phone and smart-watch apps respectively. While these numbers are extremely low in comparison to the total number of activities humans perform on a daily basis, this self-report approach to ground-truth logging was highly acceptable to participants in this study. Our findings also revealed that logging activities with a smart-watch resulted, on average, in three times more logged activities. This suggests that the ease of use of smart and wearable technologies may enhance user compliance with logging ground-truth activity data, which may be further boosted if coupled with known motivational strategies such as gamification. While we acknowledge that the findings reported in this paper may not be generalisable owing to the study sample and duration (two days of use for each version of the app), we argue that as an exploratory study our findings provide early guidelines on using speech-enabled technologies for ground-truth data acquisition. There is scope for future research in this area, focusing on larger and more diverse samples. For both apps, there were technical limitations such as the need for constant Internet connectivity to reach servers, which was essential for the speech recognition. Another mitigating factor was that speech recognition did not always capture speech correctly, due to diverse accents, background noise, and other factors. We anticipate that future developments in smart devices and speech recognition will afford improved interactions with these technologies and thus lead to better voice-enabled solutions for capturing ground-truth.

Acknowledgements

This work was performed under the SPHERE IRC, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1. We thank our collaborators and the participants who took part in this study for their time and insights.

References

[1] P. N. Dawadi, D. J. Cook, M. Schmitter-Edgecombe, and C. Parsey, “Automated assessment of cognitive health using smart home technologies,” *Technology and health care*, vol. 21, no. 4, pp. 323–343, 2013.

[2] S. S. Intille, K. Larson, E. M. Tapia, J. S. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson, “Using a live-in laboratory for ubiquitous computing research,” in *Pervasive Computing*. Springer, 2006, pp. 349–365.

[3] N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock, “Bridging e-health and the internet of things: The sphere project,” *Intelligent Systems, IEEE*, vol. 30, no. 4, pp. 39–46, 2015.

[4] N. C. Krishnan and D. J. Cook, “Activity recognition on streaming sensor data,” *Pervasive and mobile computing*, vol. 10, pp. 138–154, 2014.

[5] J. Pärkkä, M. Ermes, P. Korpipää, J. Mäntyjärvi, J. Peltola, and I. Korhonen, “Activity classification using realistic data from wearable sensors,” *IEEE Transactions on Information Technology in Biomedicine*, 2006.

[6] L. Atallah, B. Lo, R. Ali, R. King, and G.-Z. Yang, “Real-time activity classification using ambient and wearable sensors.” *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 13, no. 6, pp. 1031–9, Nov 2009.

[7] M. G. Tsipouras, A. T. Tzallas, G. Rigas, S. Tsouli, D. I. Fotiadis, and S. Konitsiotis, “An automated methodology for levodopa-induced dyskinesia: assessment based on gyroscope and accelerometer signals.” *Artificial intelligence in medicine*, vol. 55, no. 2, pp. 127–35, Jun. 2012.

[8] P. Woznowski, R. King, W. Harwin, and I. Craddock, “A human activity recognition framework for healthcare applications: ontology, labelling strategies, and best practice,” in *2016 International Conference on Internet of Things and Big Data (IoTBD)*. Rome, Italy: INSTICC, April 2016.

[9] L. Bao and S. S. Intille, “Activity Recognition from User-Annotated Acceleration Data,” *Proceedings of PERVASIVE 2004*, pp. 1–17, 2004.

[10] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, “Accurate activity recognition in a home setting,” in *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp ’08*. New York, New York, USA: ACM Press, 2008.

[11] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz, “Understanding quantified-selfers’ practices in collecting and exploring personal data,” in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 1143–1152.

[12] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby *et al.*, “Activity sensing in the wild: a field trial of ubifit garden,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 1797–1806.

[13] M. Crete-Nishihata, R. M. Baecker, M. Massimi, D. Ptak, R. Campigotto, L. D. Kaufman, A. M. Brickman, G. R. Turner, J. R. Steinerman, and S. E. Black, “Reconstructing the past: personal memory technologies are not just personal and not just for memory,” *Human–Computer Interaction*, vol. 27, no. 1-2, pp. 92–123, 2012.

[14] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia, “Privacy behaviors of lifeloggers using wearable cameras,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 571–582.

[15] I. Li, A. K. Dey, and J. Forlizzi, “Understanding my data, myself: supporting self-reflection with ubicomp technologies,” in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 405–414.

[16] I. Li, A. Dey, and J. Forlizzi, “A stage-based model of personal informatics systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 557–566.

[17] S. Fox and M. Duggan, “Tracking for health. pew internet & american life project. january 2013,” 2013.

[18] J. Brooke, “Sus: A quick and dirty usability scale,” in *Usability evaluation in industry*, P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds. London: Taylor & Francis, 1996, pp. 189–194.

[19] T. S. Tullis and J. N. Stetson, “A comparison of questionnaires for assessing website usability,” in *Usability Professional Association Conference*, 2004, pp. 1–12.