A data sharing platform for earables research

Jovan Powar
jsp50@cam.ac.uk
Computer Laboratory, University of Cambridge
Cambridge, UK

Alastair R. Beresford
arb33@cam.ac.uk
Computer Laboratory, University of Cambridge
Cambridge, UK

ABSTRACT

Ear-worn wearable devices, or earables, are a rapidly emerging sensor platform, with unique opportunities to collect a wide variety of sensor data, and build systems with novel human-computer interaction components. At this point in the development of the field, with projects such as eSense putting hardware in researchers' hands but being limited in reach, the sharing of datasets collected by researchers with the wider community would bring a number of benefits. A central data sharing platform would enable wider participation in earables research and improve the quality of projects, as well as being a vehicle for better data quality and data protection practices. We discuss the considerations behind building such a platform, and propose an architecture that would achieve better privacy-utility tradeoffs than existing research data sharing efforts.

KEYWORDS

datasets, earables, wearable computing, pervasive computing, data sharing, privacy-preserving data sharing

ACM Reference Format:
Jovan Powar and Alastair R. Beresford. 2019. A data sharing platform for earables research. In *EarComp '19, September 07–09, 2019, London, UK*. ACM, New York, NY, USA, 6 pages. https://doi.org/xx

1 INTRODUCTION

The recent emergence of ear-worn wearable devices, or earables, presents novel opportunities for innovation and research in personalised computing, both through new modalities of human-computer interaction and as a sensor platform. Thanks to recent advances in earable computing power and commercial availability, we find ourselves at the beginning of a new branch of wearables research. In this work we propose a tool for this new community to share data and results from studies conducted with earables, with the aim to foster collaboration and lower the barrier to entry for new research, by establishing and facilitating a set of standards for data quality and protection.

With this tool we aim to contribute a novel design for data sharing platforms, which enables sharing of rich and potentially sensitive datasets, integrating datasets gathered from across heterogeneous set of contributors and methodologies. We see the potential to introduce an approach to data sharing that strikes a better balance between utility and data protection than has been achieved in previous data sharing platforms.

In this paper we note the types of data that can be collected with an earable platform, and discuss the potential breadth and implicit data protection burdens of data collection studies. Through a discussion of the fit of existing data sharing approaches to the earable case, and a wider range of considerations for a data sharing exercise, we propose a sharing platform design that accommodates heterogeneous studies with an adaptive, rule-based approach to data protection.

As the earables research community is in its infancy, there are many tradeoffs to be made between the incentives of contributors and third parties, as well as assumptions to be validated around the nature of future research. Therefore, we also present an appraisal of the potential benefits and drawbacks inherent in the data sharing problem, and invite members of the community to comment on these problems, many of which we leave open, in order to better shape the design of the system.

2 RESEARCH WITH EARABLES DATA

The position of earables—at the head, at the focus of the human sensory experiences of sound and balance, in close contact with the skin—makes them a rich platform for collecting a variety of data. Motion or vibration sensors can contribute data on activity, gait, speech, breathing patterns, or even facial expression. Sensors in contact with the skin can contribute continuous data on the internal state of the human body: optical sensors can measure heart rate and blood oxygenation, while electrodes measuring galvanic skin response can provide an indicator for stress. For now, we focus on the eSense project [3, 4], whose wireless earphone hardware

EarComp '19, September 07–09, 2019, London, UK

ACM ISBN ?…$00.00
https://doi.org/xx

incorporates a microphone and an inertial measurement unit (IMU).

The potential breadth of datasets collected by earables is exceptionally large, and will only continue to grow as new or improved sensor hardware is introduced. This phenomenon presents opportunities and challenges alike for a research community. For a data source of such breadth, we must anticipate that the applications of this data will be similarly broad, and cannot be predicted. As such, it is prudent to make available to future researchers as much and as varied a corpus of datasets as possible. This is especially important at the initial stages of earables research, where hardware is scarce. However, this also presents problems of data quality—when data is contributed by a highly heterogeneous set of initial collection studies, we must establish a baseline of documentation and quality standards.

Collection study design

We cannot anticipate all datatypes collected in studies with earables, and since the goal of the data sharing platform is to foster collaboration on novel research the architecture of the platform must not—as far as is reasonable—impose restrictions on what datasets can be contributed. Instead, we propose that the system be designed to receive broad classes of data (discussed below in the context of privacy risks), upon which sharing and data protection policies can be designed. If a contributing researcher uploads a dataset with a datatype that is not included in the existing policies, it can be flagged for review—at which point the moderators of the system can perform a data protection analysis, leading to a new policy.

Studies will also differ in the modalities of collection. In terms of data collection 'episodes', being one continuous recording of earable sensor data from one participant, we should expect a wide variance both in temporal scope and environmental scope. Temporally, we expect datasets comprised both of short- and long-lived episodes of data collection, and datasets comprised of many one-off episodes or repeated episodes from a user.

We expect datasets to span a range of environments, which can be most helpfully parametrised by the level to which they are controlled. This ranges from highly controlled—a lab environment where the participant performs specified tasks—to minimally controlled, or 'in the wild'—episodes that take place in unspecified public spaces, with no preordained activity being undertaken.

Parametrising this space of episode types will be useful both for ensuring utility—the third-party researcher using the platform will be able to easily find and compare similar datasets—and aid in data protection, as the sensitivity of a dataset can be highly dependent on temporal and environmental scope.

Data protection

With the richness of earables data comes a range of implicit data protection burdens. A central challenge for any data sharing system will be to support and manage the *greatest* burdens associated with current or future datatypes, while *minimising the friction to third party researchers* who wish to make use of the datasets. Below we outline the data protection burdens implicit to a range of earable-collected datatype.

We propose that a data sharing platform for earable data should be capable of receiving datasets containing any or all of these datatypes, and tailor its data protection processes adaptively to which datatypes are included—either at contribution-time or at query-time. This way, greater procedural friction associated with one datatype, such as audio, does not need to be applied to a user who wishes only to access short-term IMU logs.

IMU and mobility. Privacy risks from mobility data and IMU (Inertial Measurement Unit) traces are usually minimal in the average case for research, where short-lived traces are collected. Risks arise mainly from the collection of long-term traces, which leak information about activity such as commute timings, leisure activities, or working schedules.

Some privacy risk may come from the uniqueness of certain mobility characteristics such as gait, which has been used to fingerprint individuals. It is unclear how much entropy can be derived from ear-collected gait analysis, but it is unlikely that it could be used to 'blindly' identify a subject—that is, if one does not already have a gait fingerprint of the subject, and does not already know that they contributed to the dataset. Under a conservative evaluation of the reidentification power of IMU data, it will nonetheless be necessary for researchers to assume that if a study participant contributes IMU data to their dataset, third parties will be able to reidentify that participant in other datasets available through the platform. This must be considered in any consent agreements made between researchers and their participants.

Audio. As earables are usually marketed to users as wireless headphones, they invariably include a microphone. Audio recordings are a potentially highly sensitive type of data to collect on users, especially if the study takes place 'in the wild'. This problem is compounded by the fact that it cannot be said for sure that an audio recording does not contain sensitive information (whether it picked up someone else's conversation, or if it captured a highly personal moment for the subject), without listening through it completely.

Therefore, a data sharing platform must allow contributing researchers to tag their datasets with information that describes the risk of this sensitivity. If the dataset is tagged as containing only short audio snippets from a controlled environment, minimal data protection mechanisms would need to be employed; if the recordings were made on the street, or at the subject's home, the system must consider each of those cases as progressively more sensitive, and apply greater protections. These might come in the form of stronger licensing agreements, stronger consent requirements to contribute data at all, or transformation of audio into representations with lower fidelity.

Future sensors (e.g. electrodes). While we can speculate about the usage and collection modalities of novel sensor datatypes, such as galvanic skin response, we know that the uses of those datatypes will evolve as they become available to researchers. Therefore, it is important not to prescribe data protection policies for these datatypes but to continuously evaluate the tradeoffs between their sensitivity and their utility.

While the system's policies must be incrementally formed as more datatypes are added to the corpus, it is important that this early lack of strategy be properly presented to participants at the point of consent. Consent documentation must clearly explain the open-ended nature of the usage of the participants' data, and the users' right to have the data minimised or better protected as soon as technical means become available should be communicated.

Metadata. The majority of datasets contributed to a public repository of earable datasets will not share a collection methodology, and even in the initial case where we focus solely on the eSense platform, may have been pre-processed by different applications. Therefore it is necessary to provide detailed metadata on the activity captured by each dataset. This must include, but not be limited to the time-frame of collection events, their frequency, and the degree to which the environment was controlled. As noted above, each of these will also contribute information on the sensitivity of the dataset.

Other non-sensor information will usually be collected in an earables study. This might be information on participants, such as age, gender, or level of physical activity; or tagging of the sensor data collected, such as by activity or location. Both to ensure utility to third-parties and for data protection, it is important to establish a baseline of data quality for these sorts of information. The question of how strictly to draw the specification of this metadata should be agreed with the community.

3 EXISTING APPROACHES TO SHARING DATASETS FOR RESEARCH

There are a number of other data collection projects which have made their data available to third-party researchers, as well as systems designed to handle that sharing. These studies range in scope from collection studies where the researchers have collected a dataset and roll their own sharing platform, to collection and sharing tools created for third parties to integrate into their own studies, to platforms that simply serve as repositories for datasets (employing varying levels of mediation).

Two notable collection studies are the Device Analyzer [6] and Haystack [1] projects. Each of these studies publicly released an Android app to capture various data about an individual's smartphone activity. These participants are members of the public who are incentivised by reports about their smartphone activity (as well as a desire to contribute to research, of course).

Both studies took a different approach to sharing their data. From Haystack, the researchers published a dataset containing a subset of the data collected—anonymised TLS handshakes for 1378 devices over the course of two years. This anonymised data was published under a Creative Commons Attribution 4.0 International license, on a the public dataset sharing website Zenodo.

Device Analyzer, on the other hand, makes available their entire dataset, after post-processing and pseudonymisation of certain fields [7]. As this was deemed to be highly sensitive data (continuous data for thousands of users on all aspects of smartphone usage) even when pseudonymised, the sharing mechanism chosen employed significantly more friction, resulting in months passing between initial application and access being granted. The project's administrators require an application to include a research proposal, be intended to result in published research, and require a license agreement to be signed between the institutions.

Both of these approaches are instructive for the design of an earables sharing platform, but neither can be followed closely. Haystack's approach allowed for a low barrier to entry, but only for a very limited subset of their data, while Device Analyzer provides a very rich dataset but behind a significant barrier to entry—by design. While these projects illustrate the two poles of the privacy vs. utility tradeoff, they do not match our case exactly as they deal with only one collection study. A platform for earables, which could host data from zero to high sensitivity, would be of little use to the community if it were to address only one instance of the privacy-utility tradeoff. Instead, we advocate a system that can adapt to the sensitivity of datasets with progressively stronger or more suitable protections.

The AWARE framework [2] provides a platform for researchers to have study participants contribute data from their smartphone, such as ambient noise or location. Researchers can write their own plugins to collect specific data for their purposes. The data is uploaded to the servers (hosted by the project or by the researchers themselves), which host a web dashboard for researchers to access the data. Collection tools such as AWARE are able to provide a single interface to sensors which can be used by different researchers, delivering a uniformity of data format and quality that is useful for inspecting data from multiple studies.

There are a number of systems proposed for hosting data securely and privately. One notable recent work is ScrambleDB [5], a database designed to store multiple datasets, as is our goal. Its 'pseudonymisation-as-a-service' allows a dataset to be decoupled into constituent datasets, providing non-linkability guarantees for the decoupled outputs. In addition to enforcing access control, this principle allows a more granular, targeted enforcement of data protection policies. Approaches such as these are central to our proposed system, as is detailed in Section 5.

4 PARAMETERS ON DATA SHARING

There are many parameters for a data sharing platform that must be set at the time of design. Here we present a range of parameters relevant to the earables case; we present our reasoning around each, but for the most part we leave the parameters open to input from the wider earables community.

Third party qualification

Who? As noted in the comparison between Device Analyzer and Haystack in Section 3, there are a range of options for who should be allowed to browse and download datasets from the platform. The four options we consider, in increasing order of restrictiveness are:

- Anyone, via a public website
- Anyone who makes an account on the platform, agreeing to Terms of Service
- Only with researchers intending to pursue work for public academic publication (including industrial researchers)
- Only with academic researchers at known institutions

Each increased level of restriction serves two purposes: first, 'locking in' researchers to the platform aids community building, as it increases the likelihood that they will comply with terms of service or license agreements, which may include stipulations to cite the contributors of datasets used in their research; second, increasing the likelihood of compliance serves as a data protection mechanism, which can alleviate the need to perform more restrictive technical transformations—put simply, the more likely the third party is to behave well, the more of the raw data you can give them.

We believe that a third party should need an account to use functionality beyond browsing metadata. A generic license agreement, covering citation rights and guidelines on ethical usage of the data, should be agreed to before a user is allowed to check out any datasets.

From a legal perspective, it may be the case that for certain datatypes sharing will be covered by the GDPR. Here we have three options: prohibit upload of those datatypes, ensure the data cannot be accessed by anyone not in a GDPR-compliant country, or simply to limit third parties to users in GDPR-compliant countries. Until we clarify this legal situation, it would be prudent to take the third option.

With a community resource, there is always a danger that someone may take without giving; the more often third parties withdraw data without depositing new data, the more likely the system is to suffer abuse (i.e. that a third party breaks the terms of service). This could be combatted by only allowing users to check out datasets if they have already shared their own. However, this policy is also vulnerable to abuse, as it may encourage users to contribute fake or incomplete data as a workaround, reducing the average data quality on the platform. We leave open the question of how better to encourage users to be good community actors.

Datatypes

As noted in Section 2, the datatypes collected as part of earables research are diverse and cannot be anticipated. Therefore, it is important that any sharing system aiming to facilitate future research is able to host arbitrary datatypes. Currently, we will focus on the eSense platform, which has an IMU onboard, as well as a microphone; even in this case, it would be overly prohibitive to allow only IMU traces and audio recordings to be uploaded. Studies will likely generate much more varied data, such as information about the collection environment, participants, location, etc.

Expanding the scope of data handled by the platform significantly complicates the problem of data protection; we discuss our approach to this problem, by constructing generic classes of non-sensor data and applying data protection rules thereupon in Section 5.

An open question is how to ensure data quality. By specifying data representations, or accepting only aggregate data, we impose a degree of homogeneity on contributors' data, which may limit the design of contributors' studies and may turn them off sharing data altogether. Conversely, if we were to also provide libraries to be used in data collection (similar to the AWARE framework), we might lower the barrier to entry for new research. Answering this question will come down to consensus from the community on the demand for shared data collection tooling.

Tooling complexity

If successful, the platform would be well placed to do some heavy lifting on the behalf of researchers. This might be as simple as producing aggregate data for export, or as complex as hosting and running analysis code. Both of these tooling options provide utility while also improving data protection; contributors can specify that only aggregate data may be used by third parties rather than allowing downloads of raw data, or the third party could never be given the data at all, instead having their code 'come to the data'. However, this might be unnecessary complexity—it remains an open question what degree of use such a system would garner.

The obverse approach would be to not host the data on the platform at all—it may well be the case that even preparing the data for upload is excessive work for a contributing researcher, and they would prefer instead a catalogue, where third parties can find their details and request the data in whatever way the contributor sees fit. This would, of course, negate much of the data protection benefits we have discussed, along with the secondary benefits to the community. We discuss the questions of user incentives further in Section 6.

5 A PROPOSED ARCHITECTURE

We propose a system in which contributing researchers upload their data as a collection of linked datasets, one collection per study. Datasets are split into two classes: **core** sensor data and metadata and **accessory** non-sensor data or additional metadata.

Core datasets

The initial classes of core sensor dataset will be IMU data and audio recordings. For each, we will require that each collection episode is tagged with a pseudonym for the participant, a timestamp, and appropriate metadata such as sampling frequency.

A contributor must also describe the details of the study methodology—the temporal and environmental scope of collection, as described in Section 2.

Accessory datasets

Accessory datasets are any further information collected as part of a study. Examples might include detailed participant information, location traces, or ground truth data collected from other sensors.

Checkout

A third-party researcher wishing to check data out of the system will be presented with an interface to browse studies and their datasets, inspect metadata, and choose which datasets they wish to download using a 'checkout' model. At this point, any procedural or legal data protection mechanisms can be applied—agreeing to a license agreement, verifying institutional affiliation etc. The third-party will also be given a metadata summary of the dataset they will be supplied—statistics such as sampling frequencies and included columns. This summary will give an overview of the output, including the level of availability of particular fields if the output has been composed from multiple studies. It also serves to highlight where transformations have been applied to the dataset due to data protection policies.

Protective transformations

Depending on the permissions granted to a third party, and the data that they have checked out (or are requesting to check out), transformations may be applied to core datasets to minimise data protection risks. Which protections are enforced and when is specified by global system policies and any options explicitly chosen by the dataset's contributor. Some example rules might be:

- the dataset's contributor has specified that third parties with a different institutional affiliation to theirs must receive a lower IMU sampling rate
- due to the consent form given to participants, you must have agreed to a license agreement with the contributor to receive column X in this table
- you have previously checked out a location trace dataset from this study, and so cannot be provided with this audio trace dataset

Policies made up of rules such as these will be drawn up for each expected class of core and accessory dataset.

6 COMMUNITY INCENTIVES

As the sharing platform we propose is intended to facilitate the growth of a research community around earables, we must consider the balance of incentives of our target users, as well as validate the assumptions we have made about their intents. We assume, for example, that the majority of researchers would be happy to share their data were it a simple enough task, and that there is significant utility to be gained from increasing discoverability and allowing composition of different datasets. We also assume that an earables-collected dataset will have utility to future studies. We expect that the 'first wave' of eSense research will help us validate these assumptions.

We here present a rundown of the incentives we anticipate to influence uptake and usage of an earables data sharing platform. We discuss each incentive and disincentive to the best of our ability, but we require input from the research community to validate them and understand the true balance between them.

Motivations to contribute to a data sharing platform

Community kudos. Contributors who have produced a novel or high quality dataset would like recognition from peers, and a higher chance of being cited

Allow participants access to their own data. The platform could alleviate the need for contributors to build an access portal where their study's participants can view their own data (or this could be a 'free' incentive to participants if it was not previously considered).

Easy compliance. As the platform would have a generic license, and systems for managing data protection, contributors could use a slick pre-prepared consent process that we provide. This would reduce the work they need to do, and assure easy compliance with the GDPR.

Reproducibility. Making your data available means that others can verify your analysis, or closely match your dataset in one they collect independently, to validate your research.

Motivations not to contribute data

Perceived additional work. If the contributor wishes to collect data in a specific way, or wishes to build their dataset ad-hoc, it may be perceived as too much effort to make their dataset compliant with whatever standards the platform expects. Simpler still, the prospect of having to spend time tagging and uploading may seem overly onerous.

Existing non-compliant practices. A researcher may already have collected a significant portion of their dataset without having obtained sufficient consent from participants for further sharing. If the consent process they have already established does not cover use cases such as publishing the data to our platform, it is unlikely the contributor will discard that data. Similarly, if the platform's data protection rules and consent requirements are stricter than those of the prospective contributor's institution, the extra work may be perceived as overly restrictive.

Institutional wariness. As the platform would be hosted by the University of Cambridge, contributing researchers may have reservations about providing data—either because they do not want to share with our institution, or do not trust our processes to manage sharing with others. This wariness could be enough for some researchers that they would prefer to share directly with a third party.

Reverse kudos. If the contributor's study ended inconclusively or if the data was of poor quality, the contributor may choose not to publicise their data for fear of judgment (even though that dataset might still be useful for other purposes).

Motivations to check data out from the platform

Data availability. If successful, the platform would host a wide diversity of samples, and a larger volume of data than even a well-resourced researcher could easily collect.

Easier than running collection. If the researcher wishes to quickly test an assumption, or has limited resources, it would often be easier to check out a dataset from the platform, rather than collecting it independently.

Unavailable accessory data. Similarly, a researcher could check out a dataset that includes accessory data that they would not have been able to collect themselves, such as high-fidelity ground truth IMU, heart rate, or room temperature.

7 CONCLUSION

In this paper we have proposed a data sharing platform for earables research, and discussed the considerations involved in ensuring it provides utility to the emerging earables research community. The approach we advocate would allow for wider participation in earables research, and simplifying many aspects of running studies—such as providing a streamlined consent process and handling legal compliance. We believe our proposed architecture allows for better balancing of collaborative utility and data protection than previous efforts. We have left a number of questions open, most notably:

- Should the platform include standardised data collection tooling?
- How strictly should data format and quality standards be specified?
- Do the perceived benefits of a complex sharing system outweigh the drawbacks?
- Does the community actually see utility in this approach?
- How do we encourage users to be good community actors?

We aim to discuss these questions further with the community as the first results from the eSense project are presented.

REFERENCES

[1] ICSI UC Berkeley. last accessed 2019-08-15. ICSI Haystack Project. *https://haystack.mobi/.*

[2] AWARE framework. last accessed 2019-08-15. AWARE - Open-source Context Instrumentation Framework For Everyone. *http://www.awareframework.com/.*

[3] F. Kawsar, C. Min, A. Mathur, and A. Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (Jul 2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[4] Nokia Bell Labs. last accessed 2019-08-15. eSense Earable Computing Research. *http://www.esense.io/.*

[5] Anja Lehmann. 2019. ScrambleDB: Oblivious (Chameleon) Pseudonymization-as-a-Service. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 289–309.

[6] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. 2014. Device Analyzer: Large-scale mobile data collection. *ACM SIGMETRICS Performance Evaluation Review* 41, 4 (2014), 53–56.

[7] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. last accessed 2019-08-15. Device Analyzer for Android. *https://deviceanalyzer.cl.cam.ac.uk/.*