HHS Public Access

Author manuscript

Proc ACM Interact Mob Wearable Ubiquitous Technol. Author manuscript; available in PMC 2018 August 20.

Published in final edited form as:

Proc ACM Interact Mob Wearable Ubiquitous Technol. 2017 September ; 1(3): . doi:10.1145/3130902.

EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments

ABDELKAREEM BEDRI,
Carnegie Mellon University, Carnegie Mellon, 5000 Forbes Avenue, Pittsburgh, PA 15213, US

RICHARD LI,
Georgia Institute of Technology, 85 5th St NW, Atlanta, GA 30308, US

MALCOLM HAYNES,
United States Military Academy, Thayer Hall, West Point, NY 10996, US

RAJ PRATEEK KOSARAJU,
Stanford University, 450 Serra Mall, Stanford, CA 94305, US

ISHAAN GROVER,
Massachusetts Institute of Technology, 20 Ames St, Cambridge, MA 02139, US

TEMILOLUWA PRIOLEAU,
Rice University, 6100 Main St. Houston TX 77005 US

MIN YAN BEH,
Carnegie Mellon University, Carnegie Mellon, 5000 Forbes Avenue, Pittsburgh, PA 15213, US

MAYANK GOEL,
Carnegie Mellon University, Carnegie Mellon, 5000 Forbes Avenue, Pittsburgh, PA 15213, US

THAD STARNER, and
Georgia Institute of Technology, 85 5th St NW, Atlanta, GA 30308, US

GREGORY ABOWD
Georgia Institute of Technology, 85 5th St NW, Atlanta, GA 30308, US

Abstract

Chronic and widespread diseases such as obesity, diabetes, and hypercholesterolemia require patients to monitor their food intake, and food journaling is currently the most common method for doing so. However, food journaling is subject to self-bias and recall errors, and is poorly adhered to by patients. In this paper, we propose an alternative by introducing EarBit, a wearable system that detects eating moments. We evaluate the performance of inertial, optical, and acoustic sensing modalities and focus on inertial sensing, by virtue of its recognition and usability performance. Using data collected in a simulated home setting with minimum restrictions on

participants' behavior, we build our models and evaluate them with an unconstrained outside-the-lab study. For both studies, we obtained video footage as ground truth for participants activities. Using leave-one-user-out validation, EarBit recognized all the eating episodes in the semi-controlled lab study, and achieved an accuracy of 90.1% and an F_1-score of 90.9% in detecting chewing instances. In the unconstrained, outside-the-lab evaluation, EarBit obtained an accuracy of 93% and an F_1-score of 80.1% in detecting chewing instances. It also accurately recognized all but one recorded eating episodes. These episodes ranged from a 2 minute snack to a 30 minute meal.

Additional Key Words and Phrases

Wearable computing; activity recognition; automatic dietary monitoring; earables; chewing detection; unconstraint environment

CCS Concepts

Human-centered computing →Ubiquitous and mobile computing design and evaluation methods; Computing methodologies →Supervised learning by classification; Hardware →*Sensor devices and platforms*; Applied computing →Health informatics

1 INTRODUCTION

Fueled by the ubiquity of sensing and computation, the interest in automatic food intake monitoring is flourishing. There are three main aspects to food intake monitoring: (1) *when* does eating take place [13, 14]; (2) *what* is eaten [25, 42]; and (3) *how much* is eaten [11, 37]. To answer these questions, nutritionists typically rely on self-reports. While these tools suffer from several limitations, bootstrapping the user with semi-automated tracking of how often, at what time, and for how long they ate holds significant promise. Similar to step counting applications, where even a coarse estimate of activity levels can motivate users to improve their behavior [27], monitoring *when* food is consumed can perhaps help in modifying and improving unhealthy behaviors. Over the last decade, a significant amount of research has explored various approaches to fully automate food intake monitoring. Devices ranging from a microphone on the neck [38] to EMG-measuring eyeglasses [40] to in-ear microphones [16] have been explored. Since an important first step in research is to achieve reasonable lab-controlled performance, most work so far has thus occurred in laboratory settings with reasonable results [3, 17, 36]. In the real world, however, eating is often accompanied by and occasionally occurs simultaneously with other activities, such as speaking and walking, and such behavior is rarely captured in laboratory settings. The leap from a lab's controlled environment to the real world is too big; replicating laboratory performance in the real world, therefore, turns out to be extremely difficult [14].

In this paper, we present *EarBit* (Figure 1), an experimental, head-mounted wearable system that monitors a user's eating activities while remaining resilient to the unpredictability of a real world setting.

EarBit uses chewing behavior as a proxy for eating, resulting in instrumentation of the head. As an experimental platform, EarBit's design allows for the collection of data from a number of sensing modalities (optical, inertial, and acoustic). We use these sensors to determine the combination of sensing modalities that is most effective for detecting the moment of eating. To reduce the gap between results from a controlled laboratory setting and the real world, the algorithms for these sensors (shown in Figure 1) were developed and evaluated in a semi-controlled home environment that acts as a living lab space. The results of this study indicated that an inertial sensor behind the ear (measuring jaw motion) in tandem with an inertial sensor behind the neck (monitoring body movement) produced good results in detecting eating activity, and was also the form factor considered most comfortable by the participants; particularly since the function of the inertial sensor behind the neck is used to detect activities like walking and could be replaced by a user's smartphone or wrist-mounted activity tracker.

Eating detection models trained on data from the semi-controlled study were then tested on a new dataset collected in a relatively relaxed "outside the lab" environment. We recruited a new set of 10 participants, and instead of asking them to come to our study location, we gave them the EarBit prototype and asked them to use in their own environments. We collected data for a total of 45 hours. EarBit's IMU is essentially a chewing sensor, and at a 1-second resolution, EarBit correctly recognized chewing activity with an accuracy of 93% and an F_1 score of 80.1%. When these Outside-the-Lab chewing inferences are aggregated into separate eating episodes, EarBit accurately recognized *all but one* recorded eating episodes (delay = 1 minute). These events ranged from 2 minutes snacks to 30 minutes meals.

The main contribution of this paper is a demonstration of the experimental EarBit system that successfully recognizes eating episodes in a real world setting. This contribution comes in three parts:

1. An evaluation of a wearable setup for eating detection based on off-the-shelve form factors.
2. A novel, semi-controlled laboratory protocol used to judge the effectiveness of combinations of three sensing modalities for eating detection.
3. A machine learning model that uses inertial data collected in the semi-controlled environment to reliably recognize eating episodes in a real world setting.

2 RELATED WORK

A wide variety of sensing approaches have been used to solve the complex problem of automatic food intake monitoring. The choice in sensing modality as well as sensing location is difficult because it must be effective in detecting eating instances and be robust against natural behaviors. Additionally, the sensing modality should be practical in a wearable form factor (i.e., portable and compact, energy efficient, and aesthetically pleasing). A comprehensive review of sensor-based methods and systems for automatic dietary monitoring can be found in [31]. This section will discuss some of the large body of work with an emphasis on two aspects: the sensing modality (acoustic, motion, and

multimodal) utilized and its potential for detecting eating events in unconstrained environments.

2.1 Acoustic Sensing

Acoustic-based systems have primarily been placed on the neck-region or in-the-ear for sensor-based dietary monitoring. Amft et. al. [1] used a microphone placed inside the ear canal to detect eating and to classify between four food types. To validate their method, researchers ran a lab study with four participants. Their system was able to recognize eating from other non-eating samples with up to 99% accuracy using C4.5 decision tree classifier. Using isolated chewing instances, the system achieved 80%–100% accuracy in classifying between eating chips, apple, pasta, and lettuce.

Amft and Troster [2] investigated the recognition of swallowing activity during food intake. Using gel electrodes and electret condenser microphone, the researchers recorded the surface EMG and sound from the neck after integrating the sensors in a collar like fabric. 868 annotated swallows were collected in the lab from 5 participants. The study included different activities of eating and drinking. Signal intensity thresholding was used to evaluate the system performance in detecting swallowing form data streams. While achieving a recall of 70%, the precision remained low due to the large number of false positives. Feature similarity search was then used as a second method of evaluation, and it was able to improve the overall precision.

A neck-worn system was used for real-time swallowing detection in [28] and a recall performance of 79.9% and precision of 67.7% was achieved. Using a similar neck-worn system, [39] demonstrated the ability to identify 12 different activities; precision of detecting eating was 81.3%. While [28] conducted their study in the lab, [39] evaluated their system both in the lab and with a small-scale study in unconstrained environments.

Acoustic sensing has also been used for detecting food types. In these cases, the results of detecting when the user is eating is conglomerated with the results of identifying the food types being eaten. [1] compared six wearable microphone locations for recording chewing sounds and observed that the inner ear location provides the highest acoustic signal intensity. [9] presented the AutoDietary system for food type recognition and obtained an 84.9% accuracy in identifying food types between 7 types. Meanwhile, in [30], the authors modified a hearing aid to include two microphones (one in-ear and one for ambient noise), also in an attempt for food type classification. They achieved a user-dependent accuracy of 79% and a user-independent accuracy of 66%. Rahman et. al. [34] developed a wearable system (BodyBeat) that place a piezoelectric microphone on the neck to detect non-speech body sounds. One of the activities they investigated is food intake. The BodyBeat system was able to recognize chewing and swallowing sounds to categorize the current activity as eating or drinking. Performing leave-one-user-out cross-validation, BodyBeat was able to recognize eating with 70.35% recall and 73.29% precision.

In general, a primary drawback of acoustic-based systems is that it can be significantly affected by environmental noise. This concern is indicated by the trend that the majority of the aforementioned studies use data collected solely in the lab or under controlled settings.

2.2 Motion-based Sensing

Inertial sensors have been placed on the body in a number of different locations. In [35], the authors used the 3-axis accelerometer in a commercial smartwatch to detect the motion of bringing food to the mouth using the arm. Using a model trained from data collected in a laboratory setting, they achieved F-scores between 71.3% and 76.1% on data collected from 8 participants in unconstrained environments. The metrics were produced looking at hour long segments of time, and these unconstrained environments studies lasted a week, with an additional study lasting an entire month. Similarly, the authors in [13] used a wrist-worn system consisting of an accelerometer and gyroscope for detecting eating periods in free-living. Data was collected from 43 participants over a 12-hour period and results showed an accuracy of 81% at 1 second resolution.

GlasSense is a wearable system developed to recognize different facial activities by monitoring the movement of the temple [12]. GlasSense has two load cells embedded in the hinges of a 3D printed eyeglasses to measure the temporalis muscle activity, This signal is used to recognize facial activities such as: chewing, talking, head movement, and winking. A pattern recognition pipeline based on SVM classifier was used to classify between natural head movement, left chewing, right chewing, left wink, right wink, and talking. Using leave-one-user-out cross validation on a dataset of 10 users collected in the lab, GlasSense was able to classify between the six activities with an average F1 score of 94%.

2.3 Multimodal Sensing

Multimodal systems are expected to be more robust than unimodal systems if sensors are combined in a way that each sensing modality contributes unique information toward the goal of dietary monitoring [31]. Examples of sensor combinations that have been evaluated in literature include vibration sensors located behind-the-ear combined with a nose-bridge electromyography sensor in an eyeglass form factor [41] as well as a camera and microphone combination in a headset form factor [22]. A headset form-factor has also been explored by [7], wherein they demonstrated a prototype of proximity sensor embedded inside an ear-bud that detected the deformation of the ear canal cause by jaw movements. Combined with an on-body gyroscope to detect body motion, the system was able to detect eating activities with 95.3% accuracy on data collected in the lab.

The technique of using a dedicated inertial sensor for detecting body motion was also explored by the Automatic Ingestion Monitor (AIM) system presented in [14], in which an accelerometer was used to better inform a novel artificial neural network algorithm. Combined with other sensing modalities (a jaw motion sensor and a hand gesture sensor) in a multi-unit system, the system was able to obtain an average accuracy 89.8% in detecting eating episodes of 30 seconds. The dataset comprised of 12 participants in unconstrained environments over a 24 hour period.

This paper is different from other works in the literature because we introduce a multi-modal wearable system to detect eating events in relatively unconstrained environments. We started with evaluating three sensing modalities: optical, inertial and acoustic, and evaluated the performance of the system in a semi-controlled lab study and outside-the-lab study. Our

evaluation assessed the performance of the system in detecting eating both at a 1 second resolution and at the event level.

Generally, it is hard to formally compare performance with previous work in the domain of automatic food monitoring. The primary reason is because systems are usually evaluated with different protocols for data collection (e.g, constraints, types of activities). However, if we made a general comparison between EarBit and some of the systems discussed above, which were evaluated in unconstrained environments [13, 14, 35], EarBit achieved a 93% accuracy in detecting chewing at a 1-second resolution and obtained an F1 score of 90.5% in detecting eating events with 72.2% coverage accuracy. These results show that EarBit mostly outperforms these systems on both the frame and event levels. However, these are only informal and preliminary comparisons, and even with EarBit's relatively better performance, the problem of eating recognition is far from solved.

3 DATA COLLECTION

McGrath identified three key factors when conducting a study: precision, generalizability, and realism [24]. However, it is difficult to collect data that has all three elements. At one extreme, laboratory experiments allow researchers to accurately measure behavior because the researcher can control when and where behaviors of interest occur [10], but this data often lacks realism. At the other extreme, *in situ* observations allow researchers to capture real life behavior. However, this data often lacks precision due to the lack of proper instrumentation or control, resulting in poor ground truth data. Consequently, the leap from a controlled study to the *in-situ* study often becomes intractable for machine learning models.

3.1 Semi-controlled Lab Study

In order to bridge the gap between controlled and real life studies, we collected our training data in a simulated natural environment. We observed participants interacting in a sensor-instrumented home (the Aware Home at the Georgia Institute of Technology) especially designed to support ubiquitous computing research [20]. This 3-storied building spans over 5,000 sq. ft., and is embedded with various sensors to support data collection.

3.1.1 Scenario—The participants were invited to the Aware Home for dinner. Once at the house, a researcher facilitated the group's activities over a 75-minute session. There were 3 to 4 different participants in each session. In an attempt to catalyze conversation, participants were chosen such that each participant was familiar with at least one other participant at the dinner. In total, sixteen participants (19–25 years, 9 female & 7 male) participated in a total of 5 sessions.

After completing a brief demographic survey, the participants were asked to wear the multi-sensor setup shown in Figure 1. Once the participants felt comfortable with the hardware, they either ate dinner, took a tour of the home, or engaged in free-flowing conversation while watching TV. Although the group had the freedom to chose the order of activities, all participants performed all activities in each session. Also, there was no restriction on the duration of each activity.

The tour required walking through the home, including walking up and down a flight of stairs. The group decided whether to eat their dinner either in the dining area or the living room. The participants chose their dinner entree from local restaurants with different cuisines. While watching TV, participants were also offered snacks, such as potato chips, chocolate candy, peanuts, apples, and bananas. Additionally, participants were provided bottled water and assorted sodas to drink. Since participants already knew each other, they were comfortable with spontaneous, free-flowing, natural conversations that rarely required any host facilitation. Additionally, familiarity allowed the participants to eat in a natural manner without being self-conscious about their manners. For example, participants often talked and ate simultaneously.

Of the sixteen participants involved in the study, only ten participants provided usable data. Four participants had to leave prematurely due to a personal emergency, and two participants had corrupted or missing sensor data. Nevertheless, our semi-controlled dataset had 12.5 hours of annotated data with almost 26% labeled as *chewing*.

3.1.2 Ground-truth—We used four video cameras to record participants' activity. Three stationary video cameras were located in the dining area and living room, and a handheld camera was handled by a researcher. This camera followed participants when they went to areas outside the range of the stationary cameras (e.g., the stairs, kitchen). In order to sync the devices' data with the cameras, each participant was asked to perform a gesture of tilting their heads from side to side. To sync the video cameras, we switched the house lights on and off three times at the beginning of each session.

Over the course of the scenario, user behaviors included walking, standing, sitting, talking, eating, laughing, watching TV, etc. At the conclusion of the scenario, participants completed a post-study survey. The survey covered: (1) comfort ratings for different hardware components of the experimental device; (2) comfort ratings for different combinations of components; and (3) an open question about their experience with the experimental device. Following the post-study survey, we engaged the participants in an informal focus group and discussed usability, comfort, and practicality.

3.2 Outside-the-lab Study

The semi-controlled Aware Home study put the participants in a social group and aimed to collect the data in a realistic setting. While we largely succeeded in collecting realistic behavior, the participants were still aware of multiple cameras and the data recording focused on capturing eating events. For example, it would be uncommon for a user to spend 26% of their day eating. While a high percentage of eating episodes are an optimal approach to collect training data, it is not an ideal evaluation scenario. Hence, we decided to evaluate our algorithms in a slightly more relaxed and naturalistic environment. We outfitted 10 new participants (3 female and 7 male, aged 18 to 51) with EarBit and asked them to take it out of the lab and use in their natural environments. In this study, participants recorded data in diverse environments including houses, offices, cars, restaurants, prototyping workshops, streets and public transport. None of these participants were part of the previous study and the participants were advised to engage in at least one eating activity. We recorded up to two

3 hours sessions with each participant. The session length was limited by our groundtruth collection device: GoPro Hero 3.

Considering participants were going to use EarBit outside a controlled environment, groundtruth collection becomes hard. Traditionally, self-reporting any eating activity is a standard practice for determining ground truth for eating studies in unconstrained environments. However, a number of previous studies (e.g., [5]) and our own pilot study showed that self-reporting is not reliable. In an initial version of our study, several participants indicated that they forgot to write down eating times while they were eating. Instead, they wrote best guesses of time and duration. In other instances, participants did not remember to write down eating times until after the study was over. Hence, we revised the study to obtain ground truth via a chest-mounted GoPro Hero 3 camera. The camera faced upward towards the participant's face and continuously recorded activities around the participant's head (Figure 2a). Apart from asking participants to try and not occlude their mouth while eating (Figure 2c), there was no change to the instructions given to participants. They were told to conduct their normal, daily activities, and to self-report eating via manual logging. The GoPro sessions lasted for 3 hours, due to the battery constraints of the camera. In order to collect sufficient per-person data, participants were asked to complete two sessions. However, 5 of the participants were unavailable for a second session, and we had a total of 15 outside-the-lab sessions (3 hours each). 11% of the recorded data was identified as *chewing* and is representative of an average user's daily life [8]

3.3 Video Annotation

To acquire ground truth for each user's activities in both studies, we hand-annotated the video recordings from both studies. We used Chronoviz [15] to synchronize video and sensor data. Four coders annotated the data by manually inspecting the recorded audio and video. The annotations included six labels divided into two categories: body movement (*moving* or *stationary*) and jaw activity (*chewing*, *drinking*, *talking*, or *other*). Any labeling window can have one annotation from each of the two categories, but not two from the same. For example, a user could be walking (body movement) and eating (jaw activity), but cannot chew and talk (both jaw activities) simultaneously.

Moving included a wide variety of actions, like walking, body rocking, etc. *Stationary*, *chewing*, *talking*, and *drinking* are self explanatory actions. We used the *other* label for relatively infrequent but significant jaw actions such as laughing, coughing, and yawning. We did not annotate portions of the video when the participant could not be seen; though that was rare. We performed the annotation by considering non-overlapping 1 second window of video and labeling it as the activity that lasted the longest within the window. High granularity annotations allow us to learn from small, quick transitions. For example, Figure 3 shows a user having a meal over a 10 minute period. The user transitions through a number of activities while having his or her meal, and we are able to annotate small and sporadic periods of silence in addition to the main activity of chewing.

Additionally, a section of video can have more than one label, one from each of the two categories. For example, a person that is walking while eating simultaneously can have both of these labels for the same segment of data. A similar example is depicted at the end of the

expanded subplot in Figure 3, in which the subject was labeled to be both *moving* (from the body movement category) and *talking* (from the jaw activity category). However, when selecting frames for training and testing we resolve the confusion in mixed activities labels by giving different priority level for each class. The *moving* label has the highest priority, followed by *other*, *drinking*, *chewing*, *talking*, and finally *stationary*, respectively. The activity with the highest priority becomes the dominant label for the frame. Mixed activities that had eating overlapped with labels with higher priority represented only 2.2% in the semi-controlled lab (Aware Home) data set.

To annotate the video for the outside-the-lab study, we chose to provide only two labels (*chewing* and *not-chewing*). Therefore, it is important that the multi-class machine learning models trained using the semi-controlled study be ultimately converted into binary classification models for the outside-the-lab study. We will discuss this in detail in Section 4.2.

For annotating the recorded videos, we employed 4 coders and then used Cohen's Kappa to compute inter-rater reliability [21]. Kappa (K) was computed using a 15-minute video sample from the Aware Home dataset. This video was chosen so as to encompass a wide range of activities. Because any subset of activities could take place simultaneously or individually, the annotations are not conditionally-independent. Hence, we computed the inter-rater reliability for each activity separately, where $0.60 < K \leq 0.80$ represents satisfactory agreement and $K > 0.80$ represents near-perfect agreement. Our worst inter-rater reliability was $K = 0.69$ (for *stationary*) and our best was $K = 0.99$ (for *other*); average agreement across all activity labels was 0.84.

4 SYSTEM DESCRIPTION

In this section, we first describe the initial set of sensors identified to be suitable for detecting chewing/eating through instrumentation on/near the head. We then discuss the process of choosing an optimal subset of sensors leading to a revised design of EarBit and its machine learning algorithms.

4.1 Choosing the Right Sensor(s)

Our goal is to design a system that accurately detects the chewing activity as a proxy for food intake. We aim to achieve this using an optimal number of sensors, while considering the social acceptability and comfort of the form factor. To this end, we investigated a number of sensors and compared their performance and usability.

4.1.1 Sensor Selection—Previous research in food intake monitoring has focused on tracking various actions that occur during an eating activity, so-called proxies for eating. These include hand-to-mouth gestures, chewing, and swallowing. Although hand motion is involved in most eating activities, and has the advantage of leveraging common commercial sensing platforms that people already have (e.g., smartwatches or activity trackers), it has limitations (i.e., usually only one hand is instrumented) and we felt that detecting chewing and swallowing is more directly associated with eating and, therefore, sufficient to infer

eating episodes. Fontana et al. support our claim, indicating that in a naturalistic environment, jaw motion can be more indicative of eating activities than hand gestures [14].

To detect chewing, we exploit two sensing modalities: optical and inertial. The optical sensor is the VCNL4020 fully-integrated **proximity sensor** with an infrared emitter. Bedri et al. have used this sensor to track jaw motion for detecting chewing in a controlled environment [6, 7]. The sensor is placed at the entrance of the ear canal and measures the degree of deformation at the canal caused by the movement of the mandibular bone. The sensor is fixed inside a Bose IE2 ear-bud, and it features a wing to tuck under the outer-ear flap. The system does not require any calibration for different users and we evaluated its adaptability to different users in the prototype testing phase.

Apart from the in-ear proximity sensor, we augmented the outer-ear flap of the ear-bud with a **9 Degree-of-Freedom IMU** (LSM9DS0). Rahman et al. used a similar sensor to detect eating events in a controlled setting [33]. The flap helps in coupling the IMU to the temporalis muscle. This is one of the four mastication muscles and links the lower jaw to the side of the skull covering a wide area around the ear. During chewing, the muscle continuously contracts and relaxes, and this movement can be picked up by the IMU. Figure 4 shows an example of sensor stream of the behind-the-ear IMU while the user was talking, eating, and then walking.

The system also includes a **microphone** around the neck; a HBS-760 Rymemo Bluetooth headset (Figure 1). A similar microphone-based approach has been used to detect swallowing [38]. These works recommended placing a microphone coupled to the throat with some level of acoustic shielding. With the aim to increase comfort, we modified the type and placement of the sensor to be slightly more socially-acceptable. It leads to slightly degraded signal-to-noise ratio, but we accept it as a reasonable compromise. In addition to these sensors, we also placed a 9-DOF IMU behind the user's neck (Back IMU in Figure 1). This IMU is used to measure large body motions, such as walking. In the future, such information could alternatively be extracted from a wrist-worn fitness device or a smartphone.

Data from the two IMUs and proximity sensor is sampled at 50 Hz using a Teensy 3.2 microcontroller, which stores the received data on an SD card. The microcontroller, back IMU, and battery are housed in a casing and attached to the back of the Bluetooth headset, as shown in Figure 1. Audio from the wireless Bluetooth microphone is recorded at 22.05 KHz and sent to an Android phone. We developed four copies of the prototype for instrumenting multiple users simultaneously for the semi-controlled lab study.

4.1.2 Sensors Comparison—Using only the dataset from the Aware Home semi-controlled study, we compared different sensing modalities on the basis of their recognition performance and usability.

The activity recognition processing pipeline was based on prior literature and compared the performance of different sensors and all combinations of sensors using leave-one-user-out (LOUO) user-independent testing. We used the approach suggested by Bedri et al. to

develop the processing pipeline for the IMU and proximity sensor ([6], see Figure 5). Bedri et al. also recommended using Hidden Markov Models (HMMs) with 10 minute segments for the final classification. For the neck microphone, past work suggested using Mel-Frequency Cepstral Coefficients (MFCCs) to differentiate between speech and non-speech activities [23, 29]. Such a capability can be valuable to differentiate between talking and other activities. Therefore, we calculated 26 MFCCs from the microphone data (100 ms using 20-filter bank channels) before calculating further features from the audio.

Figure 6 shows a preliminary comparison across the sensing modalities. The IMU placed behind the neck (back-IMU) was used in all sensor conditions because it helped to filter out movement based on more gross body activities (e.g., walking). The behind-the-ear IMU (E) performs better than other combinations. The combination of behind-the-ear gyroscope and proximity sensor (E+P) has comparable results to E, but there are no clear benefits of using the additional sensor. Beyond this preliminary performance evaluation, we decided to focus our attention only on the behind-the-ear gyroscope. While it had marginally better performance than other sensors, more importantly it was the most preferred sensor by the users.

Our post-session survey highlighted that the participants did not prefer using the in-ear proximity sensor. Respondents rated comfort and usability on a five point Likert scale. Wilcoxon Signed Rank Test showed that the users found back-of-the-ear IMU more comfortable than the in-ear proximity sensor ($Md = 4$ *vs.* $Md = 3.5$, p ¡ 0.05). In the informal focus group session as well, multiple users complained about the in-ear earbud.

> ”The [in] ear piece was uncomfortable. It felt piercing and itchy.”
>
> ”The Bose headphones felt uncomfortable after extended periods of use.”
>
> ”I’m not used to having something in my ear when I’m eating”

Thus, we decided to limit the evaluation of the Outside-the-lab study to the behind-the-ear IMU and used the back-IMU to cancel large body motions.

4.2 Redesign of recognition pipeline

The processing pipeline described in Section 4.1.2 was based on prior literature and we used it to do a preliminary comparison of performance of various sensing modalities. Instead of opting to continue to optimize our Hidden Markov Models, we decided to switch to a different machine learning approach. In general, HMMs are more suited for discovering patterns and transitions in temporal data sequences. They are ideal when the model needs to develop an understanding of the *shape of the signal*. However, Figure 4 shows that the behind-the-ear IMU acts as a very direct sensor that captures the oscillation patterns of the temporalis muscle when a user is chewing. The behind-the-ear IMU simplifies the machine learning problem to primarily differentiate between magnitude and periodicity of motion from different activities. For this problem, we believe summary statistical features and an algorithm like Random Forests should suffice.

In the rest of this section, we provide full details of our machine learning pipeline, and provide explanations for various design decisions. Figure 7 shows the whole processing pipeline.

4.2.1 Signal Conditioning—The new pipeline starts with a **preprocessing** step to condition the raw signals. This step includes smoothing the 50 Hz gyroscope data using a Butterworth filter of order 5 (cut-off frequency = 20 Hz). Data is then segmented using 30 second windows sliding at 1 second.

4.2.2 Feature Extraction—Our feature set aims to encode the relevant information about the motion of the temporalis muscle when the user's jaw moves. For each 30 second window, we compute 78 features to characterize jaw movement while chewing. These features are essentially 13 features computed for each axes of the gyroscopes placed on the ear and back (i.e., 13 features × 3 axes × 2 sensors = **78 total features**).

When a user chews, the jaw moves, and the back-of-the-ear IMU picks up the motion. In an ideal case, energy or magnitude alone will be very high for such motions and low when the user is doing some other activity. However, a user performs many activities that can generate significant motion that gets recorded on the behind-the-ear gyroscope; walking and talking are common examples. Figure 4 shows example data from the *y*-axis of the gyroscope when the user was talking, then transitioned to eating, and then walking. One valuable insight captured by Figure 4 is that **chewing motion is more periodic than many other activities, such as talking**. On the other hand, walking and some other large motions (e.g., exercises) are also periodic. Though in some cases the overall magnitude of motion while walking is significantly larger, it won't always be true. For such cases, a separate IMU on the body (in our experiments behind-the-neck IMU, but in practice a wrist-worn or pocket-held device) can be used to detect these large motions, as shown in other research related to activity recognition [4, 18, 19]. Next, we list our 13 features that capture information about the magnitude and periodicity of motion for different axes and sensor locations. These features include time and frequency domain features that are commonly used in recognizing human activities from inertial data. Size of the FFT is same as the size of the feature calculation window (i.e, 30 seconds = 1500 samples). In [26]. Morris et. al. introduced a set of 5 features based on signal auto-correlation to reliably recognize repetitive strength-training exercises using inertial sensor. In general, the auto-correlation of any periodic signal with frequency *f* will produce another periodic signal with peaks at lag 1/f, while a signal that has no periodic component will produce no peaks when it's auto-correlated. Just like strength-training exercises, chewing produces repetitive motion that can be captured using same features. Hence, our features set also includes auto-correlation features, and were computed using the same methods as applied in [26].

1. **Magnitude of motion**.
    a. **Root Mean** Square encodes the amount of energy in the signal.
    b. **Variance** is square of RMS and encodes similar information. Having both RMS and variance can provide flexibility if there is non-linearity in some axes.

c. **Entropy** reflects the amount of information (or conversely noise) in the signal. Entropy tends to be a strong feature in detecting silent and noisy activities, such as silence and speech. The normal formula for Shannon's entropy was used to compute the entropy feature, but the bins are predefined in increments of 10, ranging from −50 to 50. The outliers were assigned to a separate bin.

d. **Peak Power** is the magnitude of the dominant frequency of the signal. If a signal is fairly repetitive (e.g., eating and walking in Figure 4), the magnitude of the main frequency can indicate the intensity of motion, and can help in differentiating between facial and whole-body motions.

e. **Power Spectral Density** is magnitude of power spectrum in logarithmic scale.

2. **Periodicity of motion**.

a. **Zero Crossing** captures the rough estimate of the frequency of the signal.

b. **Variance of Zero Crossing**. Zero crossing is going to be high for any high-frequency data, and can be severely affected by noise. We calculate the variance in the times at which signal crosses zero, to record the periodicity of zero crossings.

c. **Peak Frequency** is the dominant frequency of the signal, calculated through a frequency transformation.

d. **Number of Auto-correlation Peaks**. Abnormally high or low number of peaks here indicate noisy signal.

e. **Prominent Peaks** are the number of peaks that are larger than their neighboring peaks by a threshold (0.25). Higher number of prominent peaks suggest a repetitive signal.

f. **Weak Peaks** are the number of peaks that are smaller than their neighboring peaks by the same threshold (0.25) as Prominent Peaks.

g. **Maximum Auto-correlation Value** is the value of the highest auto-correlation peak. A higher value suggests very repetitive motion.

h. **First Peak** is the height of the first auto-correlation peak after a zero crossing.

4.2.3 Feature Selection—Given the large number of computed features, we introduced a feature selection step in our pipeline. This step helps in avoiding the curse of dimensionality and enhances the generalizability of our eating detection models by reducing overfitting.

We implemented the feature selection process using the sequential forward floating selection algorithm (SFFS), which is proven to be very effective in searching for optimal feature set [32]. For feature evaluation, we used random forest classifiers to build models using out

semi-controlled lab dataset. A leave-one-user-out cross validation was performed at each step, and the exclusion and inclusion criteria for features was based on the F1 score of chewing detection.

The SFFS algorithm selected 34 out of 78 features as most effective for eating detection. These 34 features came from all 13 feature types across different axes. The most common selected feature types are entropy, peak frequency, the number of auto-correlation peaks, and first peak after a zero crossing.

4.2.4 Recognition—We use Random Forests (implemented with the *Scikit-learn* toolkit in Python) and leave-one-user-out validation to avoid overfitting. Furthermore, we keep all Random Forest-specific parameters at their default values to avoid any manual overfitting. This is where Random Forests are especially useful because they do not need much manual tuning and the only major parameter is the size of the trees. However, with separate feature selection phase, we do not need to control the size of the trees as well in most cases. Therefore, we only optimize some of our windowing parameters and we will discuss those in detail later in this section.

Detecting Chewing: The labels in the Aware Home dataset included: chewing, walking, talking, stationary, drinking, and other. Due to the very low number of occurrences in the dataset, the latter two labels, which represented 5.3% and 1.2% of the dataset respectively, were removed from training and classification tasks. Completely removing these instances from the dataset would skew the timeline. Therefore, the algorithm simply skips these instances during training and classification tasks, but still uses the sensor information to calculate features for other instances (remember that the features are calculated over 30 second windows). In our dataset with 26% of data points labelled as chewing. This happens because our training data was collected in a social setting when the group of participants were socializing and a significant amount of time was spent eating. While this is not representative of an average day in a user's life, it provides us with some robust training data.

In contrast to the Aware Home dataset, the Outside-the-lab dataset only had two labels: *chewing* and *not-chewing*. However our machine learning models made a four-class classification: chewing, walking, talking, and stationary. Instead of changing the classifier's output classes to match the labels used in the Outside-the-lab dataset, we simply treat all non-chewing predictions as "not-chewing". Therefore, when we report results in Section 5, we convert our performance metrics to reflect the performance of a binary classifier. In the interest of uniformity, we do this conversion to binary classification for both the semi-controlled lab (Aware Home) data set and the Outside-the-lab datasets.

The machine learning model produces recognition results every 1 second (recall that we used 30 second windows sliding by 1 second). Since, there is seldom any need for 1 second resolution for chewing inference, we apply a moving average on the confidence value returned by the Random Forests. Consecutive values were averaged together to produce the new confidence value for each second. The moving average window was centered on the

value to be predicted. The size of moving average window (optimal value = 35 samples) is tuned using the Aware Home dataset.

The output of the filter is converted into a binary decision by using a simple threshold of 0.5. An example of this post-processing is shown in Figure 8. The result of this tuning procedure will be discussion in Section 5.

Detecting Eating: Aggregating Chewing Inferences: Although EarBit acts as a chewing sensor, most users will be interested in identifying eating events. We aggregate individual chewing inferences into eating event inferences through a two-step process (shown as the last step in Figure 8): *merging of events* and *filtering short events*.

Merging of events helps in removing sporadic discontinuities in eating recognition. This is based on an assumption that a user won't have two meals within 10 minutes of each other. Therefore, we merge all labeled and recognized eating events that occur within 10 minutes of each other. Here, we understand that time cannot be the only factor in segmenting meals. For example, a user might start eating an apple, leave for an urgent meeting, and then come back to continue eating the fruit. Perhaps a richer understanding of the user's activities and intent would be necessary, but that is not the focus of this paper.

In addition to the merging step, we added a second layer of filtering to remove small isolated events that are less than 2 minutes in duration. This filtering step comes at the cost of skipping very short snacks, which is a compromise we made to improve precision in detecting full meals and snacks that are longer than 2 minutes.

Overall, we minimize the number of tunable parameters in our approach; Random Forests also implicitly minimize the need of tuning parameters (as discussed earlier). Therefore, the only tuning parameter for EarBit is the size of the moving average filter. All other parameters were based on domain knowledge and assumptions about the user's behavior. For example, for merging events, we assume that a user won't have two separate meals within 10 minutes of each other. This assumption was also confirmed when we analyzed the video recordings. None of the tunable and human-set parameters were optimized using the outside-the-lab dataset. That dataset was collected to evaluate EarBit's performance and we made sure that none of EarBit's parameters were optimized on it.

5 RESULTS

In this section we will discuss EarBit's performance in detecting eating in our two studies. We started by developing and validating our algorithm on the Semi-Controlled Lab dataset and then we used those models to evaluate performance of the Outside-the-lab dataset. We completely sequestered the data from the Outside-the-lab dataset and analyzed it only after the algorithm was ”frozen”, that is, after satisfactory validation on the Aware Home dataset. This was done to avoid any unintentional and manual overfitting on the test data.

For evaluation, we test the algorithm's performance on both frame-level (chewing detection) and event-level (eating episode detection). The main performance measures are F_1 score, precision, recall and accuracy. For the event level analysis, we also reported *delay*, which

measures the time from the beginning of an eating event till EarBit starts recognizing it. Additionally, we also measure *coverage*, i.e., what percentage of actual event was recognized. For example, if a user spends 15 minutes having dinner, but EarBit predicts a 12 minute eating event, then Coverage is 80%. In cases where the predicted event starts before or ends after the actual event, *Coverage* can give artificially high results. However, we did not have any case where the predicted event exceeded the time-bounds of the actual eating episode.

The main difference between *coverage* and *recall* as metrics in our evaluation is, recall is computed directly on prediction values produced by Random Forest. While coverage is computed after applying the filtering steps on the prediction results as shown in figure 8

5.1 Validation on Semi-controlled lab dataset

To validate the performance of EarBit's algorithm, we used leave-one-user-out cross validation. We used these validations to tune our only tunable parameter: size of the moving average window.

Figure 9 shows chewing recognition results for semi-controlled lab study as a function of the moving average window size. The results stabilize at 35 seconds mark. EarBit's cross-validation accuracy is 90.1%, F1 score is 90.9%, precision is 86.2%, and recall is 96.1%.

For the event-level performance, with a 35 seconds moving average window, EarBit captured all 15 eating events in the dataset, and falsely recognized one non-eating episode as eating. It achieved 89.6% coverage and the average delay in event recognition is 21.3 second. Once the moving average size and the machine learning models were final, we evaluated its performance on the Outside-the-lab dataset.

5.2 Outside-the-lab Study

For the Outside-the-lab data, with a 35 seconds moving average window, EarBit detects chewing with an accuracy of 93% (F1 score = 80.1%, Precision = 81.2%, Recall = 79%). When converted into eating episodes, EarBit successfully recognized 15 out of 16 eating episodes, and it only falsely recognized 2 additional eating episodes. The average delay is 65.4 seconds and the mean coverage is 72.2%. After reviewing the dataset we found that during the 2 falsely recognized events the participants were talking, and for the single miss-classified eating event the participant was eating a frozen yogurt. Since our models was trained on chewing instances, this explains why events that don't contain regular chewing such as eating ice cream or soup cannot be fully recognized.

As we discussed earlier, the filtering step was added to help reduce the number of false recognized eating events. To evaluate the effect of this filtering step, we also ran our analysis after excluding it from the pipeline. As expected, the number of false positives increased to 10 for the semi-controlled lab dataset and 20 for outside-the-lab dataset.

6 DISCUSSION

The overall results from the semi-controlled lab study and outside-the-lab study show that EarBit was successful in detecting eating with high accuracy outside the lab. EarBit was able to recognize accurately almost all eating events in both environments we tested it on. The sole falsely recognized eating event was eating frozen yogurt, which doesn't contain the regular chewing activity that our model is trained on. The high event coverage values (89.6% in-the-lab and 72.2% outside-the-lab) indicate EarBit capability in automating the food journaling process with a precise logging of meals and snack duration's. EarBit also requires about a minute to recognize an eating episode. This low delay values allows EarBit to be used in applications that require just-in-time interventions.

6.1 What it means for the end user?

The outside-the-lab study has 45 hours of recorded data. In this duration, EarBit had only 2 falsely recognized eating events. If we assumed that a typical user sleeps for 8 hours a day, our dataset has approximately 3 days worth of daily activities. That means that EarBit generates 0.7 false positives per day. For a typical user who eats 3 to 6 means and snacks daily, the false positives do not pose a significant usability challenge. Although this extrapolation would not always be accurate, it provides a reasonable trend of the results.

By reviewing our outside-the-lab dataset, we found that the falsely recognized events are mostly due to talking activities. After visualizing the entire dataset, we found a total of 26 talking events. EarBit has only classified 7.6% of them as eating. We believe the features set we used helped in correctly recognizing most of these events as non-eating, but using EarBit with a modified user interface can improve its precision by incorporating more data from the user. For example, as soon as EarBit detects an eating event it can prompt the user with a question ”Are you eating?”, if the user's response was positive the system carries on with the food journaling process, but if it was negative the system can ask the user for a label ”So what are you doing?” and then utilize this instance to generate a better user adaptive model.

6.2 Study design

Eating detection in most laboratory settings lacks ecological validity. At the same time it is often hard to collect accurate data in unconstrained environments. Our study design aimed to solve both problems. Researchers equipped the Aware Home for recording and monitoring various eating scenarios. At the same time, the nature of a house facilitates normal interactions and eating behaviors. Thus, the researcher is able to control the environment while the participant behaves in a more natural manner. However, it was obvious to the participants that they were video recorded and the researchers were present as well. These factors meant that the setting wasn't entirely natural. Moreover, the proportion of eating events was higher than an average day in a user's life. We addressed some of these issues in the outside-the-lab study. As the participants used the system in their own environments, the proportion of eating events was more natural in this study, but they had a chest-mounted camera for groundtruth. Hence, the data collection was not entirely naturalistic here as well. We believe our fine-grained labeling of activities, and the protocol of training and evaluating the model on data from significantly different settings produced repeatable and generalizable

results. However, the quest for a true evaluation of eating activity in unconstrained environments remains unfinished.

We believe our study can serve as a good starting point for future studies on eating detection, and we hope other researchers use and improve our pipeline to detect activities - like eating - in unconstrained environments.

6.3 Self Reporting

Self-reported eating is the predominant method used to record eating in unconstrained environments. However, this method of reporting is known to be inaccurate. For example, in a laboratory setting [5] found that both people with and without eating disorders under-reported eating.

During our study, we found multiple issues with self reporting. When comparing ground truth between video footage and self report obtained from collecting data in unconstrained environments, we found that some participants forgot to report eating episodes, reported best guess eating times, and/or reported best guess eating durations. One participant reported the following:

> "*1:00 A.M.: Snacking some during movie*
>
> 19:32 snacking some more
>
> *(There was probably more but I don't remember how long it went)*"

Another participant said, "*I forgot I was wearing the device and got caught up in a conversation we were having over lunch, so I totally forgot to write down what time I started eating. I think I ate for about 30 minutes*". Participants in the study were provided monetarily incentives to report eating activity, yet on occasion they still forgot to report. From this discussion, it is probable that many studies involving self-reported eating suffer from inaccurate and incomplete data. Since our evaluation tests the system's performance on how accurately it recognizes chewing instances and eating events, we had to obtain more reliable ground truth. To overcome this issue, we decided to equip participants with a wearable camera to record their activities outside the lab. This condition imposed some limitation on the session duration due to the short battery life of the camera. The camera also can impose some restriction on the user behavior, but we believe this is a reasonable compromise for obtaining a reliable ground truth in unconstrained environments.

6.4 Form Factors

During our pilot study, we realized that in some cases the behind-the-ear IMU was not placed properly and was floating. Almost half of the earpiece was above the pinna, instead of being behind it. This issue meant that the sensor was not coupled to the temporalis muscle. We solved this issue by demonstrating the correct way to put the device to our participants and giving clear instructions to make sure that the sensor is placed properly. We largely succeeded in making sure there were no placement issues and a review of the video footage showed that there were no visible placement issues with the sensors. However, when a device like EarBit is used in the real world, it would be important for the system to be resilient and adaptive to placement issues. In our future prototypes, we are experimenting

with embedding the sensor in eye-glasses and using firmer silicone mounts in case of earbuds.

7 CONCLUSION

In this paper, we introduced EarBit, a wearable system that detects chewing instances and eating episodes in unconstrained environments. We started by evaluating three sensing modalities: optical, inertial and acoustic, and ended up settling on a behind-the-ear inertial sensor.

To assess the performance of EarBit, we conducted studies both in a semi controlled lab environment and outside-the-lab studies. In the former environment, participants engaged in a variety of prescribed activities, including eating, talking, walking, etc. Data from this study was used to train a supervised machine learning model. Next, we tested the model against data collected from our outside-the-lab study, and the trained model was found to detect chewing at the frame level with an accuracy of 92% and an F_1 score of 80%. At the event level evaluation in unconstrained environments, EarBit accurately recognized all but one recorded eating episodes, which ranged from 2 minute snacks to 30 minute meals. EarBit brings us one step closer to automatically monitoring food intake, which can ultimately aid in preventing and controlling many diet-related diseases.

Acknowledgments

This work is supported by the Center of Excellence for Mobile Sensor Data-to-Knowledge MD2K.

References

1. Amft OliverStäger MathiasLukowicz PaulTröster Gerhard. Proceedings of the 7th International Conference on Ubiquitous Computing (UbiComp'05). Springer-Verlag; Berlin, Heidelberg: 2005. Analysis of Chewing Sounds for Dietary Monitoring; 56–72.
2. Amft O, Troster G. Methods for Detection and Classification of Normal Swallowing from Muscle Activation and Sound; 2006 Pervasive Health Conference and Workshops. 2006. 1–10.
3. Amft OliverTröster Gerhard. Recognition of dietary activity events using on-body sensors. Artificial intelligence in medicine. 2008; 42(2):121–136. 2008. [PubMed: 18242066]
4. Bao LingIntille Stephen S. International Conference on Pervasive Computing. Springer; 2004. Activity recognition from user-annotated acceleration data; 1–17.
5. Bartholome Lindsay T, Peterson Roseann E, Raatz Susan K, Raymond Nancy C. A comparison of the accuracy of self-reported intake with measured intake of a laboratory overeating episode in overweight and obese women with and without binge eating disorder. European journal of nutrition. 2013; 52(1):193–202. 2013. [PubMed: 22302613]
6. Bedri AbdelkareemVerlekar ApoorvaThomaz EdisonAvva ValerieStarner Thad. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). ACM; New York, NY, USA: 2015. Detecting Mastication: A Wearable Approach; 247–250.
7. Bedri AbdelkareemVerlekar ApoorvaThomaz EdisonAvva ValerieStarner Thad. Proceedings of the 2015 ACM International Symposium on Wearable Computers (ISWC '15). ACM; New York, NY, USA: 2015. A Wearable System for Detecting Eating Activities with Proximity Sensors in the Outer Ear; 91–92.
8. Bertrand MarianneSchanzenbach Diane Whitmore. Time use and food consumption. The American Economic Review. 2009; 99(2):170–176. 2009.

9. Bi Y, Lv M, Song C, Xu W, Guan N, Yi W. AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life. IEEE Sensors Journal. 2016 Feb; 16(3):806–816. 2016. DOI: 10.1109/JSEN.2015.2469095
10. Carpendale Sheelagh. Information Visualization. Springer; 2008. Evaluating information visualizations; 19–45.
11. Chae JunghoonWoo InsooKim SungYeMaciejewski RossZhu FengqingDelp Edward J, Boushey Carol J, Ebert David S. Volume estimation using food specific shape templates in mobile image-based dietary assessment. Proc SPIE Int Soc Opt Eng. 2011.
12. Chung JungmanChung JungminOh WonjunYoo YongkyuLee Won GuBang Hyunwoo. A glasses-type wearable device for monitoring the patterns of food intake and facial activity. Scientific Reports. 2017; 7(2017):41690. [PubMed: 28134303]
13. Dong Y, Scisco J, Wilson M, Muth E, Hoover A. Detecting Periods of Eating During Free-Living by Tracking Wrist Motion. IEEE Journal of Biomedical and Health Informatics. 2014 Jul; 18(4): 1253–1260. 2014. DOI: 10.1109/JBHI.2013.2282471 [PubMed: 24058042]
14. Fontana JM, Farooq M, Sazonov E. Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior. IEEE Transactions on Biomedical Engineering. 2014 Jun; 61(6):1772–1779. 2014. DOI: 10.1109/TBME.2014.2306773 [PubMed: 24845288]
15. Fouse AdamWeibel NadirHutchins EdwinHollan James D. CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11). ACM; New York, NY, USA: 2011. ChronoViz: A System for Supporting Navigation of Time-coded Data; 299–304.
16. Gao Y, Zhang N, Wang H, Ding X, Ye X, Chen G, Cao Y. iHear Food: Eating Detection Using Commodity Bluetooth Headsets; 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). 2016. 163–172.
17. Kalantarian H, Alshurafa N, Sarrafzadeh M. A Survey of Diet Monitoring Technology. IEEE Pervasive Computing. 2017 Jan; 16(1):57–65. 2017. DOI: 10.1109/MPRV.2017.1
18. Kern NickySchiele BerntSchmidt Albrecht. European Symposium on Ambient Intelligence. Springer; 2003. Multi-sensor activity context detection for wearable computing; 220–232.
19. Khan AM, Lee YK, Lee SY, Kim TS. A Triaxial Accelerometer-Based Physical-Activity Recognition via Augmented-Signal Features and a Hierarchical Recognizer. IEEE Transactions on Information Technology in Biomedicine. 2010 Sep; 14(5):1166–1172. 2010. DOI: 10.1109/TITB. 2010.2051955 [PubMed: 20529753]
20. Kientz Julie A, Patel Shwetak N, Jones BrianPrice EdMynatt Elizabeth D, Abowd Gregory D. CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08). ACM; New York, NY, USA: 2008. The Georgia Tech Aware Home; 3675–3680.
21. Lazar JonathanFeng Jinjuan HeidiHochheiser Harry. Research methods in human-computer interaction. John Wiley & Sons; 2010.
22. Liu J, Johns E, Atallah L, Pettitt C, Lo B, Frost G, Yang GZ. An Intelligent Food-Intake Monitoring System Using Wearable Sensors; 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks. 2012. 154–160.
23. Matos S, Birring SS, Pavord ID, Evans H. Detection of cough signals in continuous audio recordings using hidden Markov models. IEEE Transactions on Biomedical Engineering. 2006 Jun; 53(6):1078–1083. 2006. DOI: 10.1109/TBME.2006.873548 [PubMed: 16761835]
24. Mcgrath E. Readings in Human-Computer Interaction: Toward the Year 2000. 2nd. Citeseer; 1995. Methodology matters: Doing research in the behavioral and social sciences.
25. Mirtchouk MarkMerck ChristopherKleinberg Samantha. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2016. Automated estimation of food type and amount consumed from body-worn audio and motion sensors; 451–462.
26. Morris DanSaponas T ScottGuillory AndrewKelner Ilya. Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM; 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises; 3225–3234.
27. Munson Sean A, Consolvo Sunny. 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops. IEEE; 2012. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity; 25–32.

28. Olubanjo TemiloluwaGhovanloo Maysam. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE; 2014. Real-time swallowing detection based on tracheal acoustics; 4384–4388.
29. Olubanjo T, Ghovanloo M. Tracheal activity recognition based on acoustic signals; 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014. 1436–1439.
30. Päßler SebastianWolff MatthiasFischer Wolf-Joachim. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. Physiological measurement. 2012; 33(6):1073. 2012. [PubMed: 22621915]
31. Prioleau TemiloluwaMoore ElliotGhovanloo Maysam. Unobtrusive and Wearable Systems for Automatic Dietary Monitoring. IEEE Transactions on Biomedical Engineering. 2017 2017.
32. Pudil PavelNovovičová JanaKittler Josef. Floating search methods in feature selection. Pattern recognition letters. 1994; 15(11):1119–1125. 1994.
33. Rahman Shah AtiqurMerck ChristopherHuang YuxiaoKleinberg Samantha. Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015 9th International Conference on. IEEE; 2015. Unintrusive eating recognition using Google Glass; 108–111.
34. Rahman TauhidurAdams Alexander TravisZhang MiCherry ErinZhou BobbyPeng HuaishuChoudhury Tanzeem. BodyBeat: a mobile system for sensing non-speech body sounds. MobiSys. 2014; 14:2–13.
35. Thomaz EdisonEssa IrfanAbowd Gregory D. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ’15). ACM; New York, NY, USA: 2015. A Practical Approach for Recognizing Eating Moments with Wrist-mounted Inertial Sensing; 1029–1040.
36. Vu TriLin FengAlshurafa NabilXu Wenyao. Wearable Food Intake Monitoring Technologies: A Comprehensive Review. Computers. 2017; 6(1):4. 2017.
37. Xu ChangHe YeKhannan NitinParra AlbertBoushey CarolDelp Edward. Image-based food volume estimation; Int Workshop on Multimedia for Cooking & Eating Activities (CEA). 2013. 75–80. 2013
38. Yatani KojiTruong Khai N. Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM; 2012. Bodyscope: a wearable acoustic sensor for activity recognition; 341–350.
39. Yatani KojiTruong Khai N. BodyScope: a wearable acoustic sensor for activity recognition. UbiComp ’12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 2012:341–350. 2012.
40. Zhang RuiAmft Oliver. Proceedings of the 2016 ACM International Symposium on Wearable Computers (ISWC ’16). ACM; New York, NY, USA: 2016. Bite Glasses: Measuring Chewing Using Emg and Bone Vibration in Smart Eyeglasses; 50–52.
41. Zhang RuiAmft Oliver. Regular-look eyeglasses can monitor chewing. Int Joint Conf on Pervasive & Ubiquitous Computing. 2016:389–392.
42. Zhang RuiBernhart SeverinAmft Oliver. Wearable and Implantable Body Sensor Networks (BSN), 2016 IEEE 13th International Conference on. IEEE; 2016. Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses; 7–12.

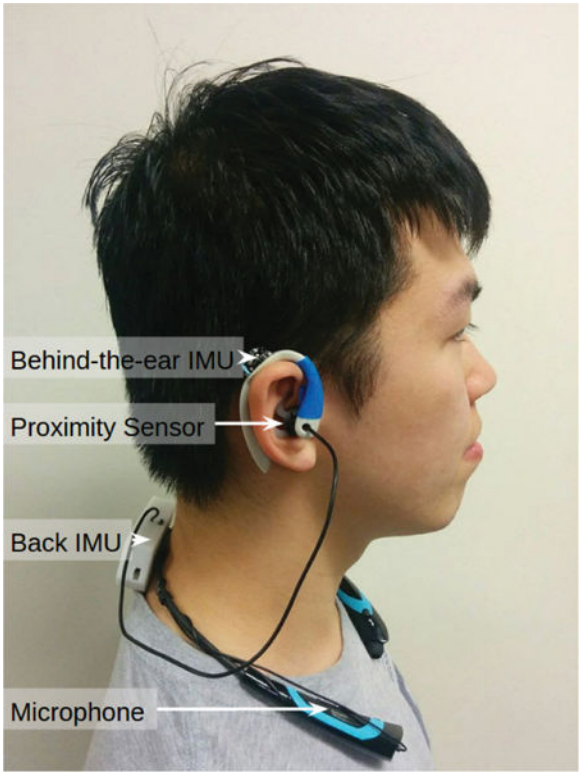

Fig. 1.

EarBit's data collection prototype with multiple sensors. Our semi-controlled and Outside-the-Lab evaluations show that the Behind-the-Ear IMU is enough to achieve usable performance. We envision such a sensor to be part of future eyeglasses or augmented reality head-mounted displays.

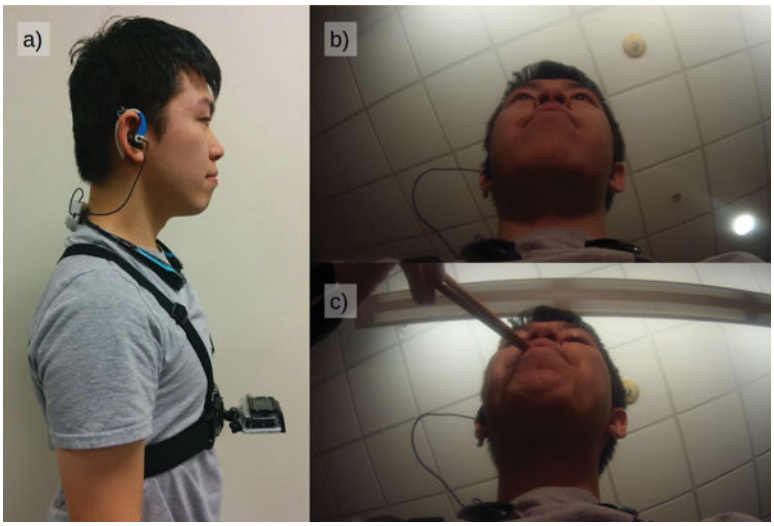

Fig. 2.

Outside-the-lab study configuration: a) A user wearing the EarBit system and GoPro camera. b) A picture from the GoPro camera of the user working at a desk. c) A picture from the GoPro camera of the user eating with a pair of chopsticks.

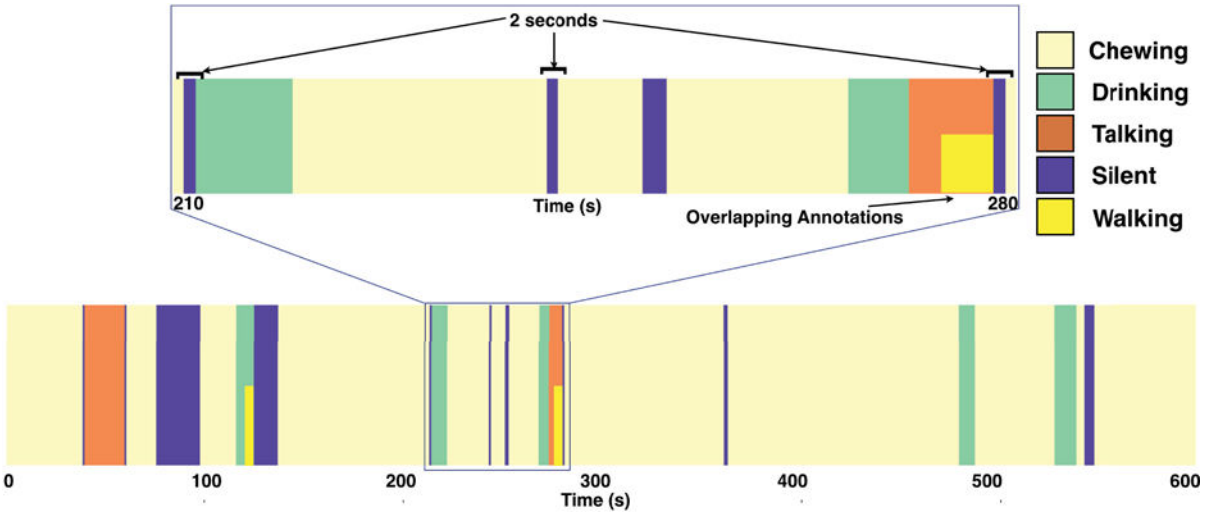

Fig. 3.

An example of annotations for eating activity. We annotated our video data at a 1-second resolution. In this 600-second example of a user having a meal, we capture all minute transitions and capture various 2-second intervals where the user stopped chewing. Mixed activities would have overlapping annotations as indicated in the example of walking and talking. For all the instances when the user is not moving a stationary label is also added.

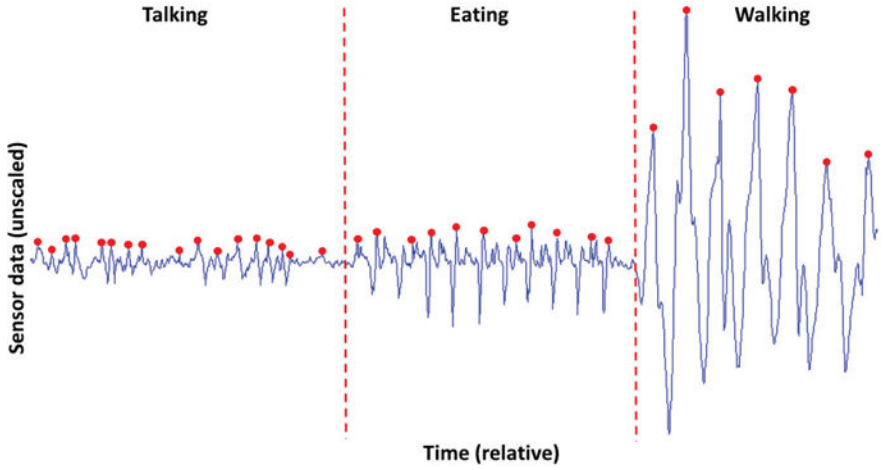

Fig. 4.

Example data from the *y*-axis of the behind-the-ear gyroscope. The dots indicate local maxima with high energy in the signal. As compared to talking, the peaks for eating are more periodic and ”spiky”.

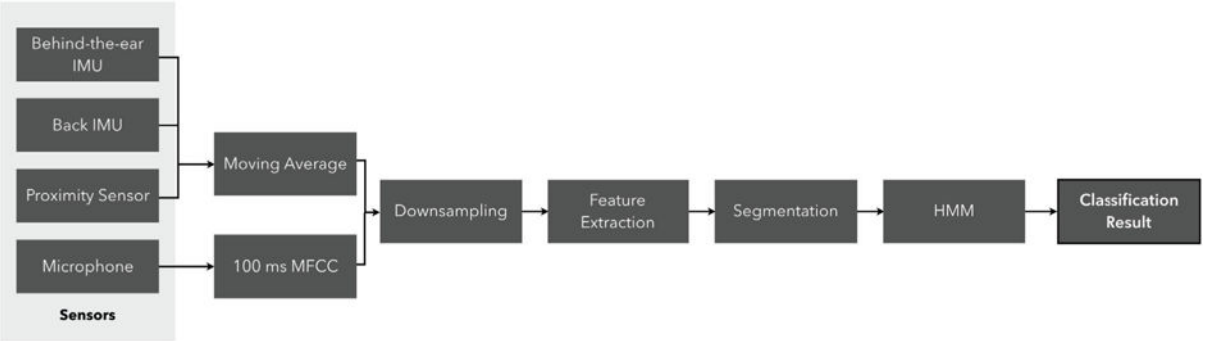

Fig. 5.
Flowchart for initial evaluation of the multi-sensor setup

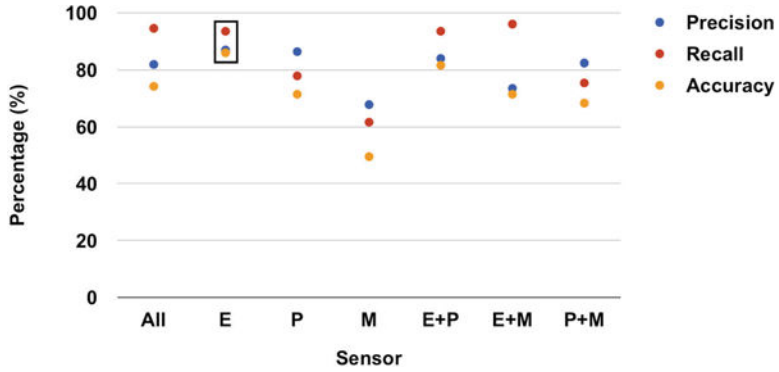

Fig. 6.

Comparison between sensing modalities. E = behind-the-ear IMU, P = outer-ear proximity sensor, M = neck microphone. The back IMU is used in all condition to detect if the user was walking. The performance of behind-the-ear IMU (E) was most consistent for all three metrics. It was also considered most comfortable to wear by the users.

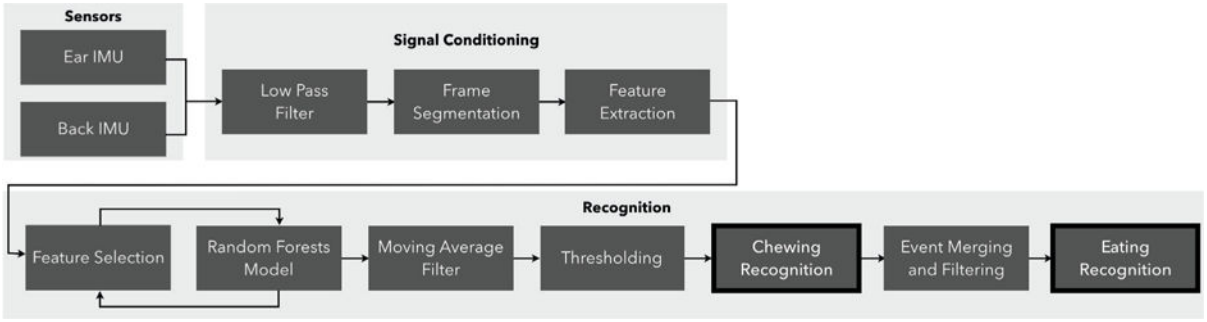

Fig. 7.
Flowchart for EarBit algorithm

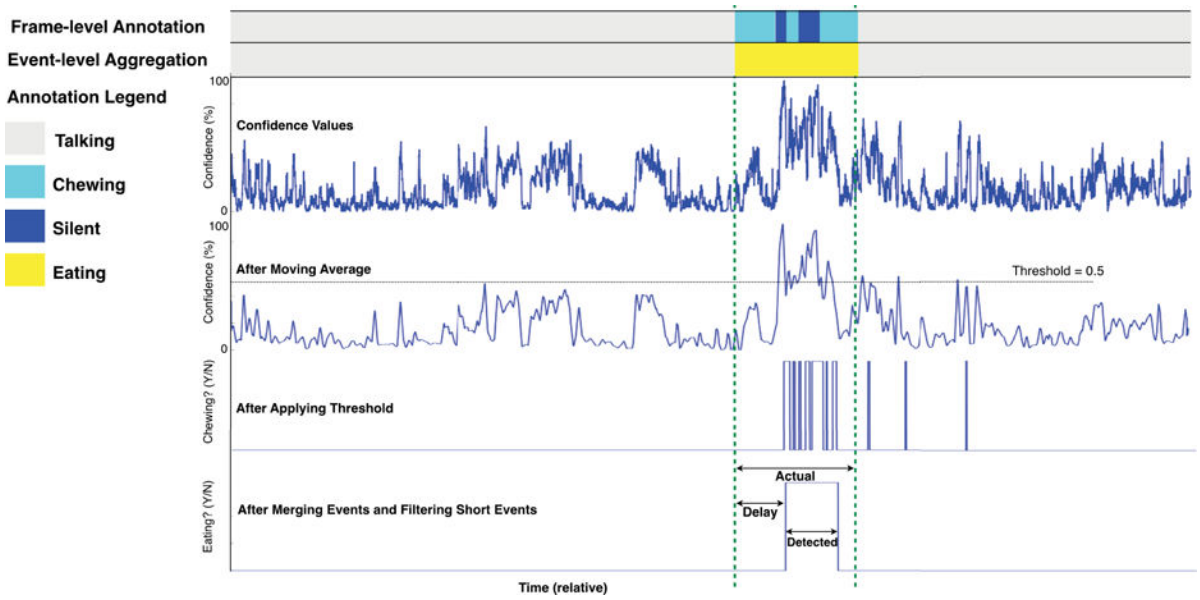

Fig. 8.

An example of conversion of confidence values from Random Forests to frame-level results (chewing) and then to event-level predictions (eating episodes).

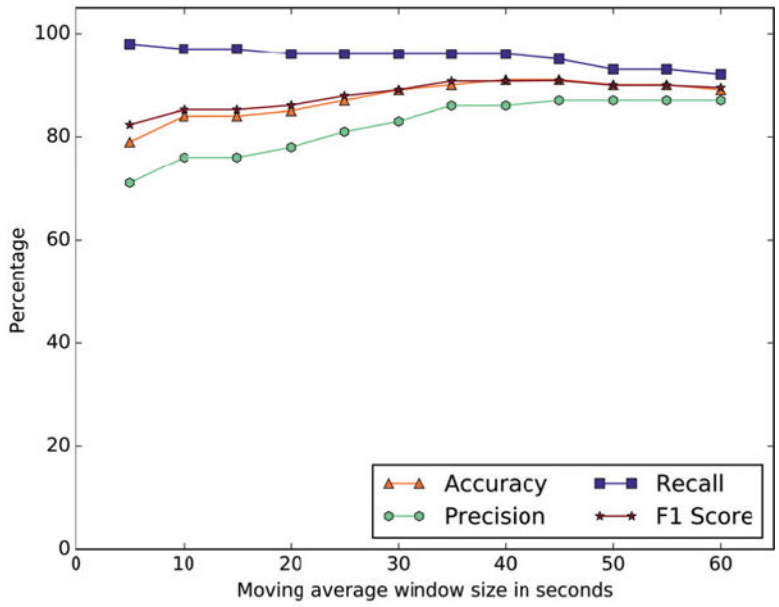

Fig. 9.
Chewing recognition results for semi-controlled lab