Auracle: Detecting Eating Episodes with an Ear-mounted Sensor

SHENGJIE BI, Dartmouth College
TAO WANG, Dartmouth College
NICOLE TOBIAS, Clemson University
JOSEPHINE NORDRUM, Dartmouth College
SHANG WANG, University of Electronic Science and Technology of China
GEORGE HALVORSEN, SOUGATA SEN, Dartmouth College
RONALD PETERSON, KOFI ODAME, Dartmouth College
KELLY CAINE, Clemson University
RYAN HALTER, Dartmouth College
JACOB SORBER, Clemson University
DAVID KOTZ, Dartmouth College

In this paper, we propose ***Auracle***, a wearable earpiece that can automatically recognize eating behavior. More specifically, in free-living conditions, we can recognize when and for how long a person is eating. Using an off-the-shelf contact microphone placed behind the ear, *Auracle* captures the sound of a person chewing as it passes through the bone and tissue of the head. This audio data is then processed by a custom analog/digital circuit board. To ensure reliable (yet comfortable) contact between microphone and skin, all hardware components are incorporated into a 3D-printed behind-the-head framework. We collected field data with 14 participants for 32 hours in free-living conditions and additional eating data with 10 participants for 2 hours in a laboratory setting. We achieved accuracy exceeding 92.8% and F1 score exceeding 77.5% for eating detection. Moreover, *Auracle* successfully detected 20-24 eating episodes (depending on the metrics) out of 26 in free-living conditions. We demonstrate that our custom device could sense, process, and classify audio data in real time. Additionally, we estimate *Auracle* can last 28.1 hours with a 110 mAh battery while communicating its observations of eating behavior to a smartphone over Bluetooth.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile devices*; *Ubiquitous and mobile computing design and evaluation methods*; • **Applied computing** → *Law, social and behavioral sciences*;

Additional Key Words and Phrases: Wearable computing, Acoustic sensing, Activity recognition, Automated dietary monitoring, Eating detection, Eating episodes, Earables, Unconstrained environment, Field studies

ACM Reference Format:
Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, Ryan Halter, Jacob Sorber, and David Kotz. 2018. Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 92 (September 2018), 27 pages.
DOI: http://doi.org/10.1145/3264902

Author's addresses: 6211 Sudikoff Lab, Dartmouth College, Hanover, NH 03755. Contact email: shengjie.bi.gr@dartmouth.edu.

2474-9567/2018/9-ART92 $15.00
DOI: http://doi.org/10.1145/3264902

1 INTRODUCTION

Chronic disease is one of the most pressing health challenges faced in the United States, and around the world. According to one report, over 133 million Americans suffer from at least one chronic disease, and the number is likely to rise to 157 million by 2020 [8]. Chronic diseases are a tremendous burden to the individuals, their families, and to society. By 2023, diabetes alone is estimated to cost $430 billion to the US economy [8]. Many chronic diseases are an outcome of, or exacerbated by, an individual's lifestyle. Eating behaviour, in particular, is strongly related to chronic diseases like obesity, diabetes, and metabolic disorders. Scientists are still trying to fully understand the complex mixture of diet, exercise, genetics, sociocultural context, and physical environment that lead to these diseases.

Wearable devices present an opportunity to measure health-related behavior [23]. Many commercially available wearable devices can monitor a person's activity level, which can be related to caloric output. There is, however, no commercially available device that can automatically detect eating behavior in free-living conditions. The availability of such automatic dietary monitoring (ADM) systems would be a huge benefit to health-science research.

An ideal embodiment of an ADM system has several challenges: (a) identifying *when* and for *how long* an individual performed an eating activity, (b) identifying *what* and *how much* is consumed during the eating activity, and (c) ensuring that the system is *usable* in real-world settings, i.e., it is unobtrusive, energy-efficient, robust to environmental noise, and easy to use. In this paper, we focus on accurately identifying when an individual is eating and for how long the activity lasted. These two goals are the foundation for automatic dietary monitoring, and could help trigger other kinds of sensing or inquiries.

To automatically recognize eating in free-living conditions, we designed and built a wearable eating-recognition system *Auracle*.[1] We assume that chewing is a first-level indicator of eating activity, so *Auracle* uses a contact microphone mounted behind the ear to detect chewing sounds.

Although several researchers have proposed approaches to monitor eating activity, it is not yet possible to accurately and automatically recognize eating outside the lab; thus our interest is to develop a wearable system, which is effective and robust enough to automatically detect when people eat in out-of-lab, day-long, free-living conditions. Indeed, we designed our system for use primarily by health-science researchers. For instance, a health-science researcher may want to study how the eating habits of college students change during a semester; when and how often do they eat? for how long? how do these patterns change during exam periods? *Auracle* could be used for such research purposes.

In this paper, we make the following contributions:

- *Auracle* is the first system that demonstrates the possibility of using a self-contained, ear-mounted system with an in-built contact microphone for eating detection in free-living conditions.
- *Auracle* runs feature extraction and classification algorithms in an ultra-low-power microcontroller (MCU) (ARM Cortex M3). Previous researchers run their models for eating detection using platforms that are significantly more power hungry (such as a laptop, smartphone, or Arduino). Based on our power measurements, we estimate *Auracle* could last for 28.1 hours with a 110 mAh battery, all while transmitting eating notifications to a subject's smartphone.
- We demonstrated the success of *Auracle* in a field deployment involving 14 participants, despite challenges with environmental noise (ambient sound, motion artifacts), in a setting different from training conditions (e.g., subjects eating while walking), and with widely varying food types.

[1] http://auracle-project.org

Fig. 1. Tip of mastoid bone

2 BACKGROUND

In most (if not all) previous reports of eating-detection technologies, researchers do not provide a precise definition of eating, even though they set out to detect eating. We define *eating* in this paper as "an activity involving the chewing of food that is eventually swallowed." This definition may exclude drinking actions which usually does not involve chewing. On the other hand, consuming "liquid foods" that contain solid content (like vegetable soup) and requires chewing is considered eating. Our definition also excludes chewing gum, since gum is not usually swallowed.

For our work in this paper, we define an *eating episode* as: "a period of time beginning and ending with eating activity, with no internal long gaps, but separated from each adjacent eating episode by a long gap, where a *gap* is a period in which no eating activity occurs, and where *long* means a duration greater than a parameter δ." We chose δ = 15 minutes in our studies as suggested by Leech et al. [16]. We used this definition for the episode-based evaluations in Section 6.2.

Although several researchers have designed systems that use various cues to determine eating (e.g., audio information from the ear canal [1, 17, 22, 27, 31] or throat [20, 25, 27, 28, 37], first-person or third-person images [26, 32, 33], wrist-based gesture recognition [10, 29]), these systems have practical limitations. They are either obtrusive (microphone on throat), uncomfortable (bulky), privacy invasive (images capturing other people) or unnatural (wearing a watch on the dominant hand).

Distinct from all these prior approaches, we have designed a head-mounted device that is similar to a behind-the-head earphone and is comfortable to wear in everyday settings. Our device can detect eating infers episodes, in real-time, on the wearable device, and logs these events as they occur, or opportunistically alerts a smartphone or smartwatch about detected eating behaviors. We chose to place the sensor behind the ear, right on the tip of mastoid bone (Figure 1); this location has been shown to give a stronger chewing signal to a contact microphone than other locations on the jaw or neck [25]. Besides, a device placed behind the ear does not impede hearing, unlike earbuds or ear-canal sensors. Moreover, this location, once the device is miniaturized, may allow a user to wear the device privately, i.e., other people would not see it and would not know that it is there (as in modern hearing aids).

No other researchers have developed an eating-recognition system that can run on-board, real-time feature-extraction and classification algorithms, which can significantly decrease the latency, improve power efficiency, and protect user privacy. Auracle can locally capture, process, and classify sensor data collected in out-of-lab, day-long, free-living scenarios.

Similarly, prior eating-detection research has traditionally worked to detect eating without considering energy efficiency and battery life—both critical factors for improving the size, weight, comfort, cost, and usability of

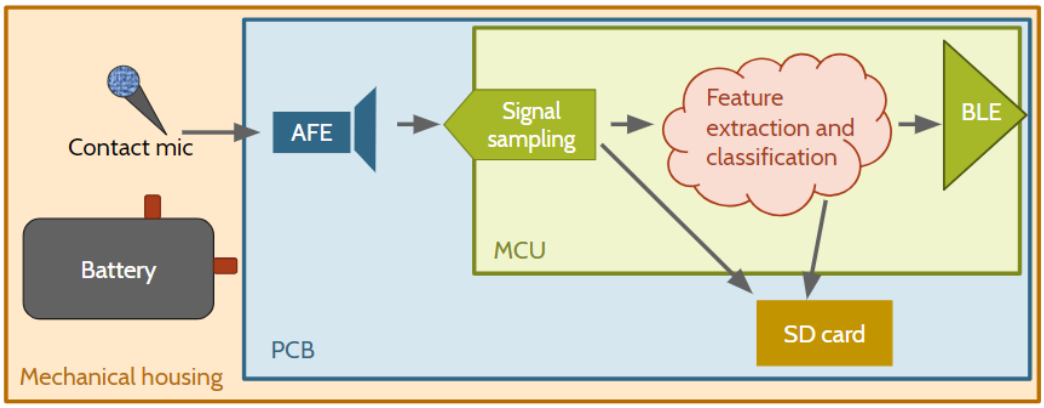

Fig. 2. Auracle prototype

any wearable device. In order to develop an eating detection system that works well beyond carefully-controlled laboratory settings, we developed Auracle using an ultra-low-power MCU (Section 3.3), evaluated Auracle's energy efficiency (Section 7) and estimated that our prototype can monitor eating continuously while lasting 28.1 hours when paired with a 110 mAh rechargeable battery.

3 SYSTEM DESIGN

The Auracle system (shown in Figure 2) includes a contact microphone (Figure 3), a battery, a custom-designed printed circuit board (PCB) for data acquisition, and a wearable mechanical housing. The PCB (Figure 4) incorporates an analog front end (AFE) for signal amplification, filtering, and buffering, an MCU for signal sampling and processing, feature extraction, eating activity classification, and system control, a Bluetooth radio for data transmission, and a micro-SD card socket for long-term data storage. The signal and data pipeline from the contact microphone includes AFE-based signal shaping, MCU-based analog-to-digital conversion, on-board feature extraction and classification, and data transmission and storage. We implemented data-logging functions to write raw data, feature values or prediction results to the SD card for our research. We also implemented Bluetooth Low Energy (BLE) functionality in the MCU so the Auracle prototype can also transmit these data through BLE, if needed. The total cost of the current prototype, including PCB fabrication and component costs, is $80 per unit, and would drop to $66 if ordered in quantities of 1,000 or more.

We developed Auracle in three stages. In Stage I, we built three prototypes and used them for acquiring field data (Section 4.1) and additional eating data (Section 4.2). We implemented only the functions required for data acquisition on the MCU. We analyzed the data (Section 5) and evaluated eating-detection performance (Section 6) offline on a laptop. In Stage II, we implemented on-board feature extraction and classification based on the most promising features (Table 1) and classification models determined in Stage I. We trained the classification model (Section 5.3) offline on a laptop using the in-lab and field data recorded (Section 4.1 and 4.2); the classification model was then implemented in embedded-C and ported to the MCU. The on-board classification model uses the feature values extracted from windows of audio samples as inputs to classify windows as periods of *eating* or *non-eating*. In Stage III, we added a Bluetooth radio on our PCB and implemented the BLE functionality in the MCU, which could be used to provide users with real-time interventions.

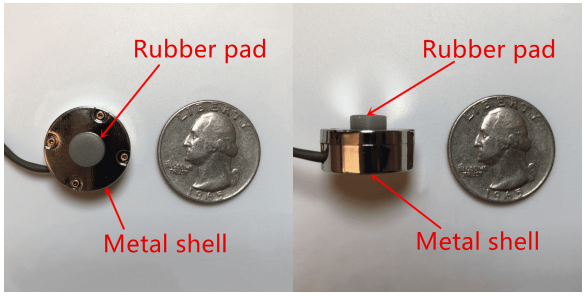

Fig. 3. Contact microphone

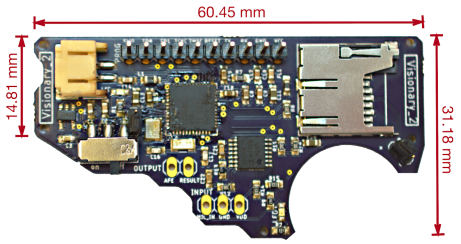

Fig. 4. Auracle's PCB Design

3.1 Contact Microphone

We used an off-the-shelf contact microphone (CM-01B from Measurement Specialties), shown in Figure 3, to capture chewing sound. This microphone uses a PVDF piezo film combined with a low-noise electronic preamplifier to pick up sound applied to the central rubber pad; a metal shell minimizes external acoustic environmental noise. The 3dB bandwidth of the microphone ranges from 8 Hz to 2200 Hz, and covers our frequency range of interest. According to the data sheet, when powered by 3.3V the power consumption of the microphone is 0.33 mW. This microphone has been used in electronic stethoscopes and, based on preliminary studies, we found it to be sufficiently sensitive to detect chewing sounds.

3.2 Analog Front End (AFE)

To make the most of the MCU's analog-to-digital converter's (ADC) input dynamic range, the contact microphone signal is conditioned by an analog front end (AFE). The AFE level-shifts the contact microphone signal, amplifies it by 15 dB, and bandlimits it to the 20–250 Hz frequency range. We chose the frequency range and amplification gain based on the experiment results from previous work [7].

3.3 Microcontroller Unit

An embedded microcontroller (MCU) samples the output signal from the AFE, processes data, and communicates results. A 500 Hz sampling rate with 10 bits of resolution is required to sample typical eating signals from a contact microphone [7]. To meet these requirements, the Auracle prototype employs a Texas Instruments (TI) CC2640R2F Simplelink Wireless MCU (ARM Cortex M3) with an integrated sensor controller and BLE module. The MCU samples and stores data over a 3-second window to construct a 1500-value array from which features are extracted and classified as *eating* or *non-eating* events. The MCU can record to the SD card raw data, summary data (i.e., feature values or prediction results), or both, depending on operating mode. The MCU can also transmit these data through BLE, if needed. The Auracle application leverages TI's operating system (TI-RTOS) for simplified task threading and automatic low-power optimization. We developed programs for the main CPU in TI's Code Composer Studio and designed and generated the firmware image for the Sensor Controller using TI's Sensor Controller Studio.

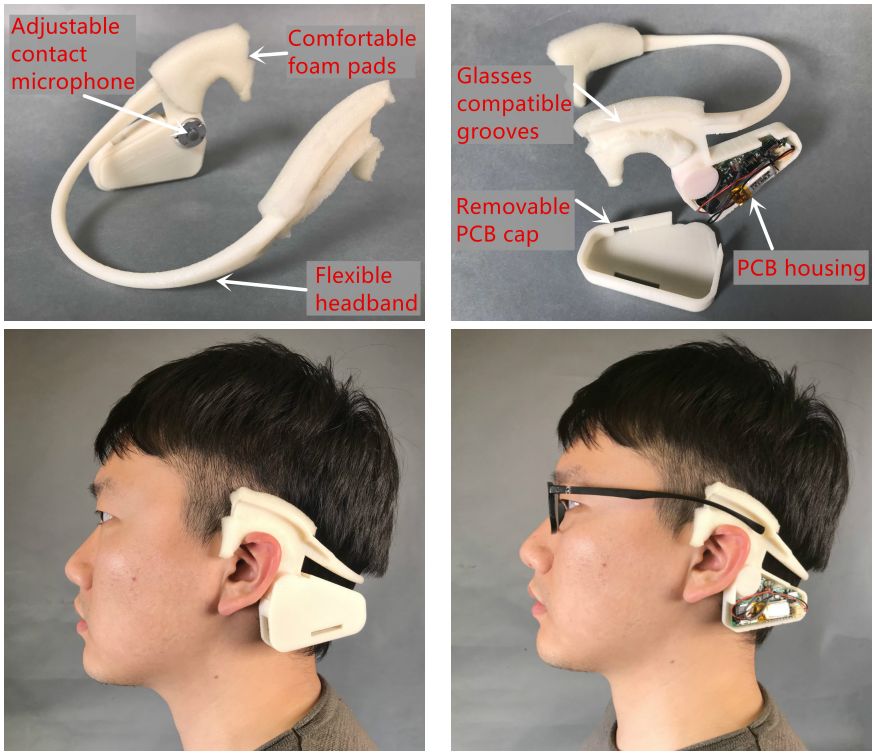

Fig. 5. Mechanical housing of Auracle

3.4 Printed Circuit Board

The Auracle prototype hardware integrates a custom printed circuit board (PCB) housed in a 3D-printed head-mounted plastic enclosure, detailed below. Figure 4 shows the PCB, which comprises the CC2640R2F MCU, a 110 mAh battery, the contact microphone (Section 3.1), a Bluetooth radio, a micro-SD card socket, and the custom AFE (Section 3.2). Our PCB implementation is small enough to be deployed in free-living conditions and its unique shape was designed to fit within the wearable form-factor of the head-mounted housing. The semicircular arc was added to the PCB design to provide a structured fit for the contact microphone.

3.5 Mechanical Housing

The Auracle enclosure consists of a 3D-printed ABS plastic frame that wraps around the back of a wearer's head and houses the PCB, battery, and contact microphone (Figure 5). Soft foam supports the enclosure as it sits above a wearer's ears. There are grooves in the enclosure making Auracle compatible with most types of eyeglasses. The contact microphone is adjustable, backed with foam that can be custom fit to provide adequate contact on different head shapes. This adjustment is necessary because Auracle is built on the premise that the contact microphone has proper contact with the mastoid bone. An adjustable microphone mount ensures that Auracle can cater to several head shapes and bone positions.

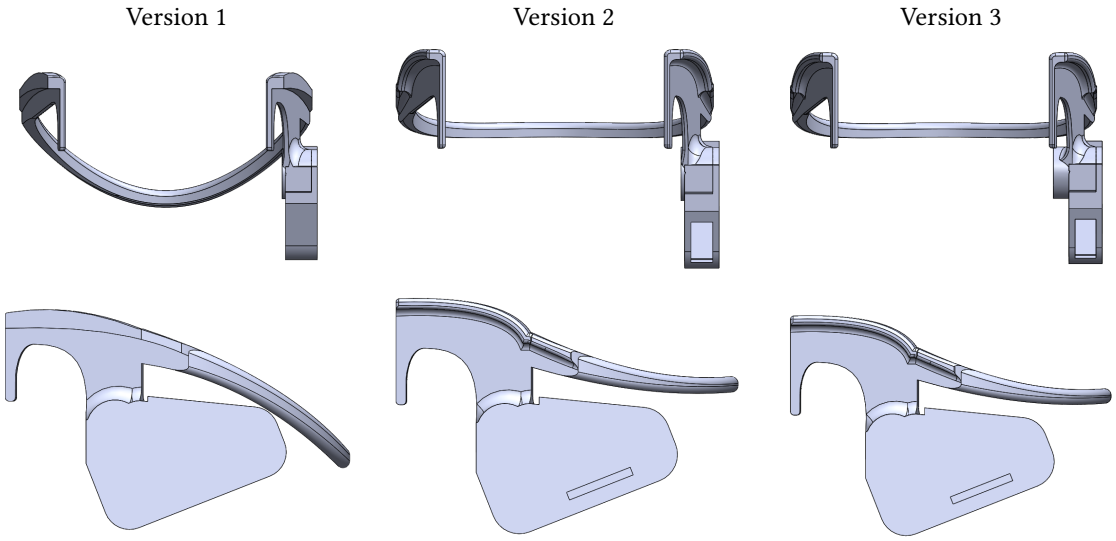

Fig. 6. Three versions of mechanical housing design

There are three versions of the enclosure to fit various head shapes (Figure 6). Version 1 wraps lower around the head than Versions 2 and 3. Version 3 has an extra extrusion to hold the contact microphone closer to the wearer if their mastoid bones are more recessed relative to their ears. All versions are 12.7cm × 12.7cm × 8.6cm.

3.6 Power Management

We plan to add a wake-up circuit in our AFE, which will keep the core MCU (i.e., Cortex M3) in the sleep state when the microphone signal is silent. Figure 7 shows the circuit we propose, which detects a surrogate measure of signal variance and compares with a preset threshold. As shown in Figure 8, when the wake-up circuit detects sound, it triggers the MCU to switch from sleep state to wake-up state and begin sampling, processing, and recording data. This process is similar to the first stage of our classification model (Section 5.4), in effect replacing the first software stage with hardware and allowing the Auracle to stay in low-power sleep state more than half of the time. There are three AD8609 in the circuit and the V_{dd} is 3.3 V. According to the data sheet, the total power consumption would be 0.5 mW, which we used as the estimated power consumption of the wake-up circuit.

4 DATA COLLECTION

Using sensor data recorded with Auracle in both field and laboratory settings, we determined an optimal set of features and an appropriate classification algorithm to implement in the digital back-end running on our PCB. We also used these data as training data for our classification model, and to derive the performance (Section 6) in terms of accuracy and power-consumption evaluation (Section 7). Under a protocol approved by our Institutional Review Board (IRB), we collected data in both free-living scenarios and a laboratory setting.

4.1 Field Data Collection

Auracle is aimed at use in free-living conditions, so we conducted a field study with 14 participants. The goal of this study was to collect raw audio data for the purpose of *developing* and *evaluating* the Auracle itself, as

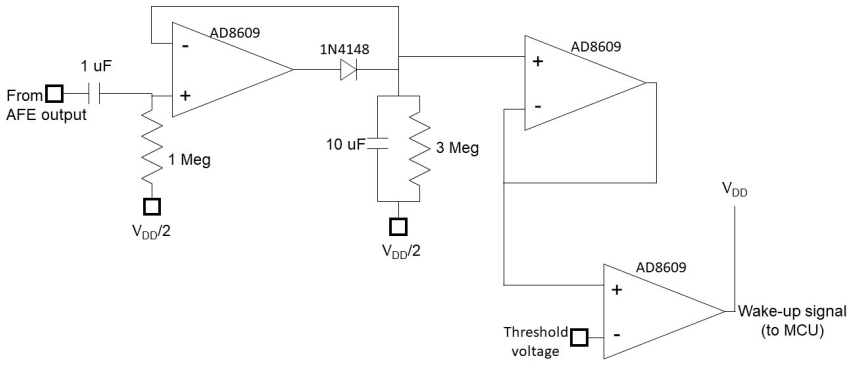

Fig. 7. Wake-up circuit

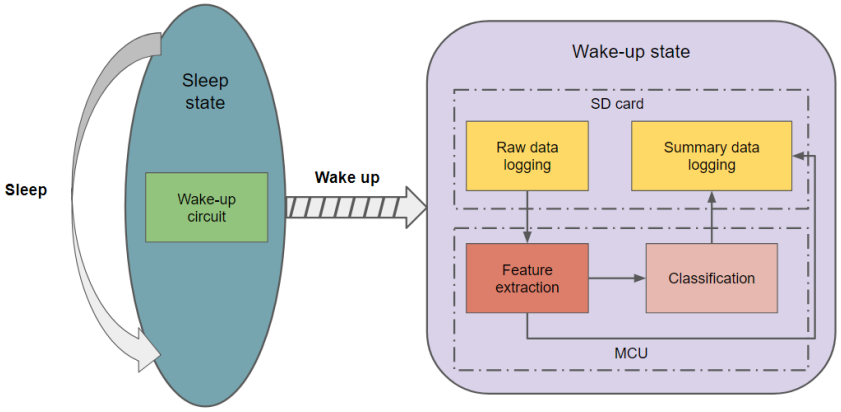

Fig. 8. State diagram

noted above. To do so, we had to address a critical challenge – we need a reliable way to obtain "ground truth" in free-living conditions. In short: *when* did the participants actually eat?

We thus developed an approach for ground-truth measurement. It is important to note that this mechanism is not part of the envisioned use of Auracle – just part of its development. We fused an off-the-shelf wearable miniature camera into a baseball cap and used the camera to record video during the field studies (Figure 9). The camera was fixed under the brim of the cap and directed at the mouth of the participants only; this orientation made it difficult to identify the participant by watching videos and also avoided recording anyone else, other than the study participant. The ambient microphone built into the camera was physically removed before the study so no audio would be captured. All the videos recorded during the study were stored in an SD card for later annotation. Compared with other similar apparatus [3], our ground-truth collector is relatively unobtrusive. Figure 10 shows two screen shots of the video recorded by the camera during eating and non-eating periods, respectively. Again, the ground-truth collector is not part of the operational Auracle and is used just for development and evaluation purpose.

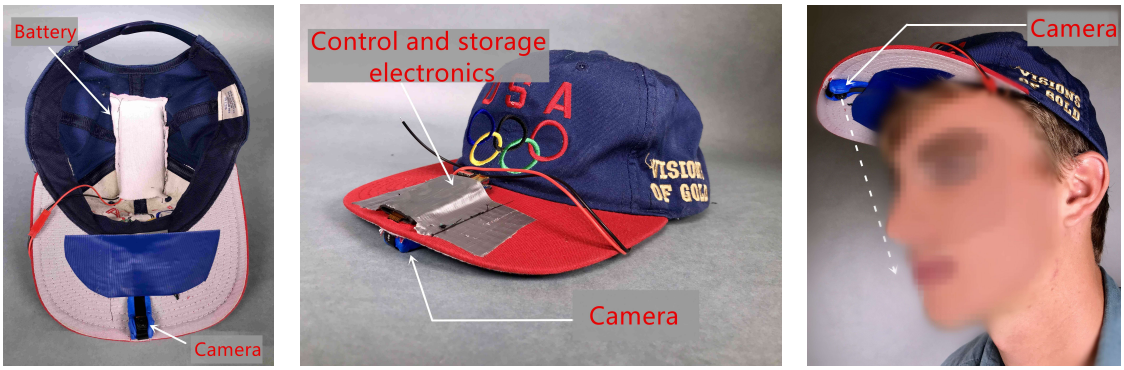

Fig. 9. Ground-truth collector

Fig. 10. Screen shots of the video recorded by ground-truth collector during eating and non-eating periods

4.1.1 Field Studies. We collected data from 14 participants (2 females, 12 males; aged 20-33; 10 wore glasses; 2 had long hair). These participants were mostly college students and staff. For each session, the participant was compensated with a $20 gift card. Among the 14 participants, 12 participants chose to participate in 1 session of the study while 2 participants chose to participate in 2 sessions. Each session lasted 2 hours. Overall, we collected a total of 32 hours of field data. After preliminary review, we found 2 sessions (4 hours) of the field data, collected from 2 different participants, could not be used for further analysis. In one session, the video recorded by cap-mounted camera was totally blocked by the participant's nose, making it hard to determine whether the participant was eating. In another session, the contact microphone signal was too weak due to poor contact and barely changed during session. We excluded the data collected during these two sessions. We used the remaining 28 hours of data recorded from 12 participants for analysis (Section 5) and evaluation (Section 6). During these 28-hour periods of field data acquisition, participants ate various types of food including rice, bread, noodles, meat, vegetables, fruit, eggs, nuts, cookies, crackers, soup and yogurt. Participants recorded data in diverse environments including houses, offices, cars, restaurants, dining halls, kitchens and streets.

Before the start of each session, the participant was asked to wear the Auracle prototype in Stage I (Section 3) and the ground-truth collector (Figure 9). To ensure the contact microphone in our prototype had good contact with mastoid bone (Figure 1), we first visually inspected whether the central rubber pad of the contact microphone remained in contact with the skin when the participant turned her or his head back and forth. We then asked the participant to stay silent for 10 seconds, followed by chewing a baby carrot for another 30 seconds. If the

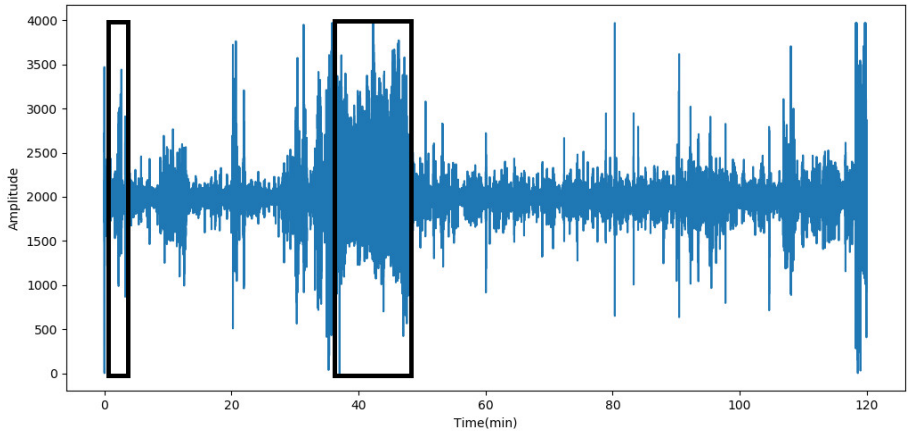

Fig. 11. Temporal signature of one session of field-data collection (black boxes indicate periods of eating)

amplitude of the data recorded during the chewing period was larger than that in silent period, we concluded there was good contact between the microphone and skin.

At the beginning of each session, we asked the participant to tap on their cheek and the mechanical housing of the prototype using their hands three times, which could be recorded by both head-mounted camera and Auracle. We then asked the participants to go about their normal daily activities outside the lab. Their behavior and location were uncontrolled, but the participants were asked to wear the Auracle and the cap continuously during their time in the field. Also, we requested that at least one eating episode take place at anytime during the session. At the end of the session, we asked the participant to perform the same three-tap event. We used these three-tap events at the beginning and end of the session to synchronize the video and audio data collected. A example of one session of field data collection is shown in Figure 11, where the parts in black boxes represent eating periods.

4.1.2 Video Annotation. To annotate the videos (i.e., labeling moments as *eating* or *not eating*), we used the video annotation service from Baidu.[2] We uploaded all field study videos to the Baidu Drive for review. Three Baidu annotators independently watched and annotated the periods of eating in each video, with 1-second resolution.

We calculated the proportion of the annotation-mismatch periods across each of the 3 reviews. Each 1-second window over which the three annotators disagreed were defined as annotation-mismatch periods. The proportion of the annotation-mismatch periods in 14 the field-study videos was small (mean: 2.79%; standard deviation: 1.85%). Thus we concluded all the videos were annotated carefully by three annotators.

We converted the three annotation results into a single label file used for experiments in Section 5 and 6. The label file was generated based on the majority annotation results from three annotators. For example, if two or more annotators annotated a 1-second period of video as *eating*, it was labeled *eating* in the final label file; otherwise it was labeled *non-eating*.

Finally, since our predictions were based on 3-second windows, we converted the resolution of the labeling result from 1 second to 3 seconds. We found that there were very few 3-second windows (less than 0.78%) that

[2]http://zhongbao.baidu.com/

Fig. 12. Six types of food used for additional eating-data collection

contained both *eating* and *non-eating* labels. We labeled a 3-second window *eating* if it contains any *eating* labels within the window; otherwise we labeled that window *non-eating*.

4.2 Additional Eating-data Collection

Since the data collected in free-living scenarios is unbalanced (i.e., much less time spent on *eating* than *non-eating*), we collected additional in-laboratory eating data to augment the training dataset. The additional data allowed us to explore whether the addition of in-laboratory eating data would improve the classification results (Section 6.2).

We collected data from 10 participants (2 females, 8 males; aged 21–33; 8 wore glasses; 2 had long hair) in the laboratory condition. At the start of each session, each participant was asked to wear the Auracle prototype described in Section 3. We used the same visual and data inspection methods used (Section 4.1.1) to verify Auracle placement in this cohort.

We asked the participants to eat six different types of food, one after the other. The food items (Figure 12) included three crunchy types (protein bars, baby carrots, crackers) and three soft types (canned fruits, instant foods, yogurts). We asked the participants to chew and swallow each type of food for two minutes. During this eating period, participants were asked to refrain from performing any other activity and to minimize the gaps between each mouthful. After every 2 minutes of eating an item, participants took a 1-minute break so that they could stop chewing gradually and prepare for eating another type of food. A signal plotting of one entire session of lab data collection is shown in Figure 13, where the parts in black boxes represent eating periods. We removed data collected during the 1-minute break periods and concatenated all 2-minute eating periods into the additional eating dataset we used in Section 6.1.

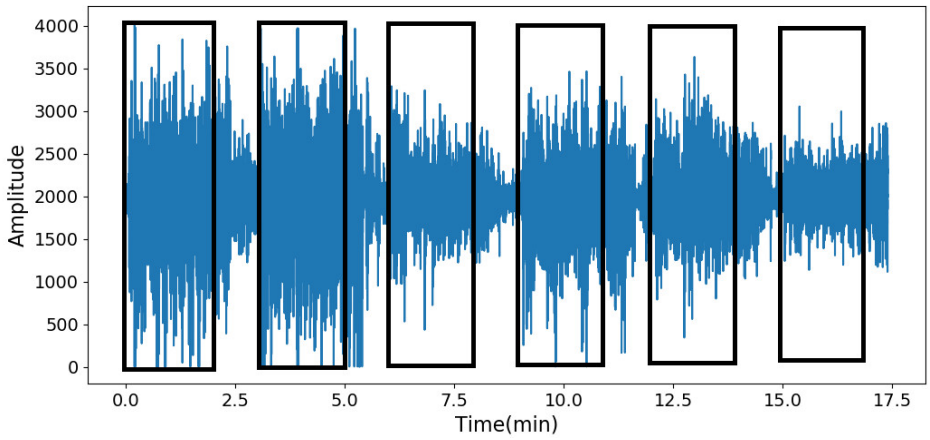

Fig. 13. Temporal signature of one session of additional eating-data collection (black boxes indicate periods of eating)

5 DATA ANALYSIS

In this section, we describe our evaluation metrics and multiple stages of our data processing pipeline (Figure 14) including data preprocessing, feature extraction, feature selection, classification, classification aggregation and ground-truth label aggregation.

5.1 Evaluation Metrics

We performed a Leave-One-Person-Out (LOPO) cross-validation to evaluate our classifier's performance in both *window-based evaluation* (described in Section 5.1.1) and *episode-based evaluation* (described in Section 5.1.2). A LOPO model is relatively unbiased because the classifier detects eating for a new person whose data it has not seen before. The model iterates over all possible combinations of the training and testing data set. For each iteration, the data set was divided into two subsets: the testing set (data from one participant) and the training sets (data from all other participants). The classifier is trained on the training sets and outputs metrics on the testing set for each iteration; we then compute average metrics across all iterations. For the LOPO experiments using additional eating data (Section 6.1), we added the additional eating dataset (Section 4.2) to the training sets in each iteration.

5.1.1 LOPO Window-based Evaluation. To evaluate the accuracy of our classifier, we compared its output for each 1-minute time window against the ground-truth label for that time window. In other words, each time window was an independent test case that resulted in one of four outcomes:

True positive: Both the classifier and ground truth indicated *Eating*.

False positive: The classifier indicated *Eating* and ground truth indicated *Non-eating*.

True negative: Both the classifier and ground truth indicated *Non-eating*.

False negative: The classifier indicated *Non-eating* and ground truth indicated *Eating*.

We defined TP, FP, TN and FN as the number of true positive, false positive, true negative and false negative cases in the testing set, respectively. We then evaluated our method using five metrics:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

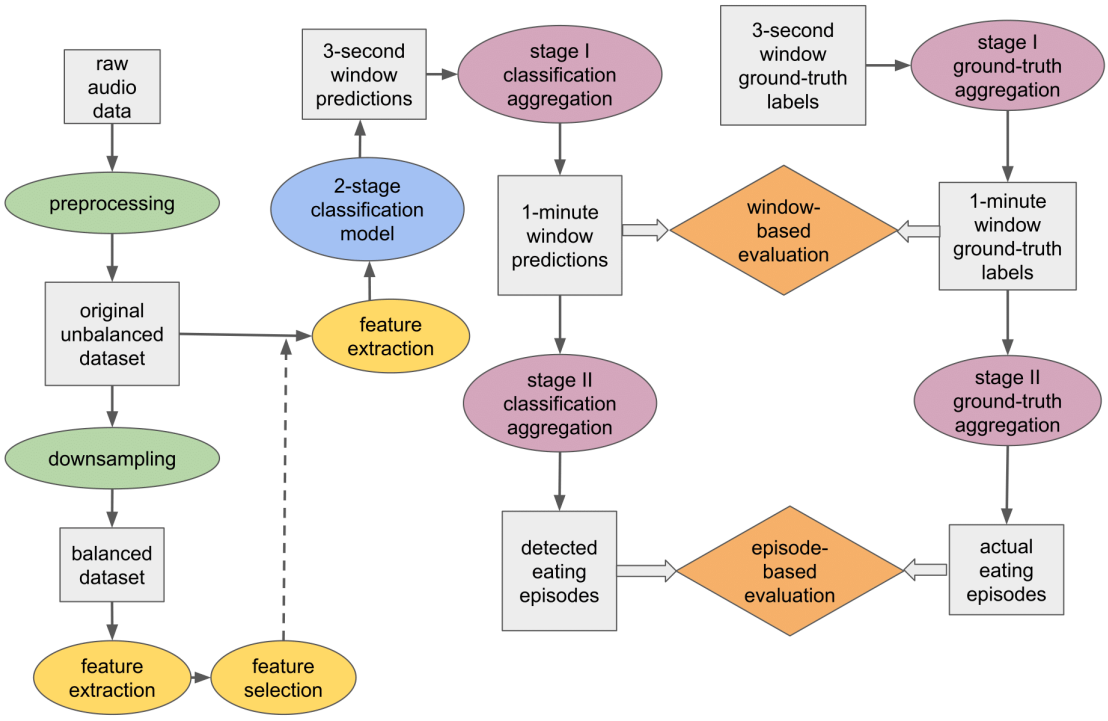

Fig. 14. Data-processing pipeline

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Weighted\ accuracy = \frac{w * TP + TN}{w * (TP + FN) + FP + TN}$$

where w is the ratio of non-eating period vs. eating period; setting $w = 1$ yields Accuracy (non-weighted) [10, 21]. We set w in weighted-accuracy metrics based on the ratio of non-eating and eating period in the testing set for each LOPO iteration. As summary metrics, we calculated the mean and standard deviation of these five scores across all iterations. Using this evaluation method, each participant affected the summary metrics equally, regardless of whether they had 2-hour or 4-hour data recordings.

5.1.2 LOPO Episode-based Evaluation. We evaluated our method's ability to detect *eating episodes* using two metrics, the Jaccard similarity coefficient and the activity-recognition metrics proposed by Ward et al. (Ward's metrics) [36].

Using an approach similar to previous work by Papapanagioto et al. [21], we matched each detected eating episode with either 0 or 1 ground-truth eating episode. We used the Jaccard similarity coefficient to determine whether this match led to a Correct Detection, False Detection or Missed Detection.

Let the detected episode be represented as $E_d = [t_s, t_e]$, where t_s is the start of the detected eating episode and t_e is the end of the detected eating episode. Similarly, the actual eating episode (obtained from ground truth) is represented by $E_a = [t'_s, t'_e]$, where t'_s is the start of the actual eating episode and t'_e is the end of the actual eating episode.

We then use Jaccard similarity coefficient:

$$J = \frac{E_a \cap E_d}{E_a \cup E_d}$$

Each detected eating episode is an independent test case that results in one of three outcomes:

$$Outcome = \begin{cases} J \geq 0.55, & \text{Correct Detection} \\ 0 < J < 0.55, & \text{False Detection} \\ J = 0, & \text{Missed Detection} \end{cases}$$

For each Correct Detection, we also calculated the mean and standard deviation of the *delay* and *duration difference*. The *delay* is defined as the absolute value of the difference between the starting time of a detected and corresponding actual eating episodes. The *duration difference* is defined as the sum of the absolute value of the difference between the starting time and ending time of a detected and corresponding actual eating episodes.

Additionally, we evaluated our method using Ward's metrics. Ward et al. define an *event* as a variable duration sequence of positive frames within a continuous time-series [36]. In our case, an eating episode represents an event and a 1-minute time window within the event represents a frame. An *event* can then be scored as either correctly detected (C); falsely inserted (I'), where there is no corresponding event in the ground truth; or deleted (D), where there is a failure to detect an event [36].

5.2 Data Preprocessing

As mentioned in Section 3.2, we first bandlimited signals to the 20–250 Hz frequency range using our AFE. The filtered signals were then segmented into non-overlapping windows of uniform duration. Based on some preliminary experiments testing a range of window sizes from 1 second to 5 seconds, we found that the 3-second window size gave us the best results so we chose 3 seconds as our default window size. Furthermore, because the signal amplitude was affected by the pressure applied to the contact microphone, which varied in each session due to different head shapes and microphone positioning, we used the RobustScaler function in Python's scikit learn package to normalize the data of each participant.

5.3 Feature Extraction and Selection

In our original field data set, the number of windows labeled as *non-eating* was significantly larger than that of the ones labeled as *eating* (the time-length ratio of data labeled as *non-eating* and *eating* is 6.92:1). When we selected features on this dataset, the top features returned provide us relatively good accuracy, but not always good recall and precision. However, recall and precision may be important metrics for some eating-behavior studies, so we first converted the original unbalanced dataset to a balanced dataset by randomly downsampling the number of *non-eating* windows so that we had equal number of *non-eating* windows and *eating* windows. We then performed feature extraction and selection on the balanced dataset (See Figure 14).

Table 1. Top 40 features selected by RFE algorithm

Feature category	Description	Number of features
FFT coefficients	Fourier coefficients of one-dimensional Discrete Fourier Transform	29
Range count	Count of values within a specific range	1
Value count	Count of occurrences of a specific value	1
Number of crossings	Count of crossings of a specific value	3
Sum of reoccuring values	Sum of all values that present more than once	1
Sum of reoccuring data points	Sum of all data points that present more than once	1
Count above mean	Number of values that are higher than mean	1
Longest strike above mean	Length of the longest consecutive subsequence that is bigger than mean	1
Number of peaks	Number of peaks at different width scales	2

For each time window, we used the open-source Python package tsfresh[3] to extract a common set of 62 categories of feature from both time and frequency domains. Each feature category in this set can consist of up to hundreds of features when the parameters of the feature category vary. In our case, we extracted more than 700 features in total. We then selected relevant features based on feature significance scores and the Benjamini-Yekutieli procedure [6]. We evaluated each feature individually and independently with respect to its significance in detecting eating, and generated a p-value to quantify its significance. Then, the Benjamini-Yekutieli procedure evaluated the p-value of all features to determine which ones to keep. After removing irrelevant features, considering the limited computational resources of wearable platforms, we further selected a smaller number of k features using the Recursive Feature Elimination (RFE) algorithm with a Lasso kernel ($5 \leq k \leq 60$). Table 1 summarizes the top 40 features.

Finally, we then extracted the same k features from the original unbalanced dataset to run the classification experiments ($5 \leq k \leq 60$).

5.4 Classification

We designed a two-stage classification model to perform a binary classification on the original unbalanced dataset, using the set of features selected above. In Stage I, we used simple thresholding to filter out the time windows that seemed to include silence. We calculated the threshold by averaging the variance of audio data across multiple silent time windows. We collected this silent data during a preliminary controlled data-collection session. We identified time windows in the field data that had lower variance than the pre-calculated threshold and marked them as *evident silence periods*. After separating training and testing data, we trained our classifier on the training set excluding the *evident silence periods*. During testing, we labeled the time windows in the testing set that were *evident silence periods* as *non-eating*. In the future, this Stage I software will be replaced by the wake-up circuit discussed in Section 3.6.

In Stage II, after experimenting with different commonly used classifiers (shown in Table 2), we chose a Logistic Regression (LR) classifier to perform a 2-class classification to classify *eating* and *non-eating* using the features

[3]http://tsfresh.readthedocs.io/en/latest/

Table 2. Results when using different classifiers with 40 features

Classifier	Accuracy	Precision	Recall	Weighted accuracy	F1 score
Logistics regression (LR)	0.928	0.757	0.808	0.879	0.775
K-nearest neighbors (K = 5)	0.888	0.621	0.810	0.858	0.689
Random forest	0.891	0.629	0.866	0.881	0.718
Decision tree	0.753	0.394	0.914	0.819	0.539
Gradient boosting	0.924	0.769	0.757	0.856	0.751

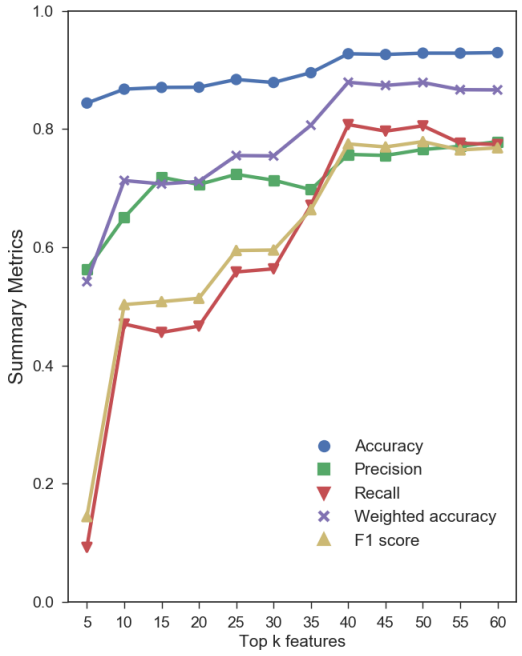

Fig. 15. Results when using only field data for training classification model

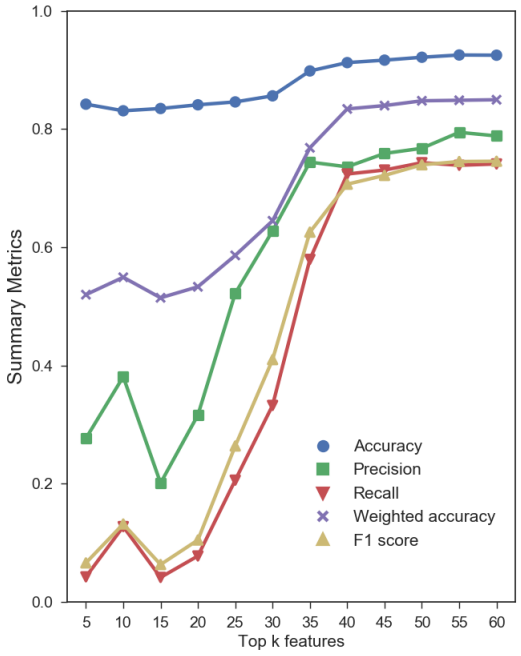

Fig. 16. Results when using both field data and additional eating data for training classification model

we described in Section 5.2. We chose the LR classifier because it yielded the best F1 score in our experiment (shown in Table 2) and it is lightweight enough to be implemented in a resource-limited wearable such as our CC2640R2F MCU (Section 3.3). Figure 15 and Figure 16 show performance of the classification model in detecting *eating* or *non-eating*, when the top k features were used ($5 \leq k \leq 60$).

5.5 Classification Aggregation

Given the classification results produced by the classifier on each 3-second window, we then decided to aggregate these results into coarser windows. We conducted a two-stage aggregation process. In Stage A, since completing

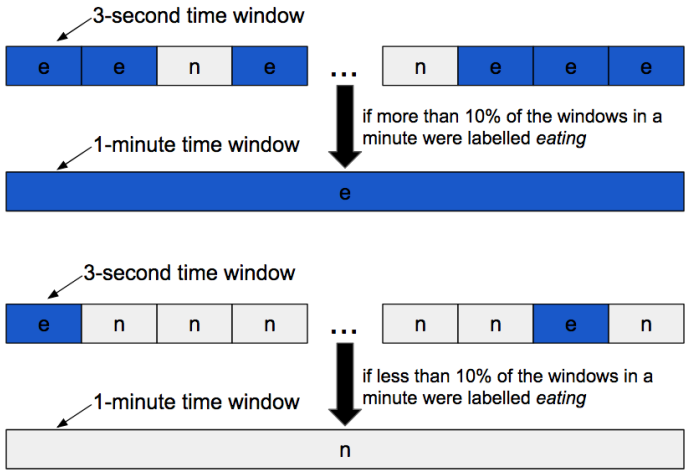

Fig. 17. Stage A aggregation (e indicates time window labeled as eating; n indicates time window labeled as non-eating)

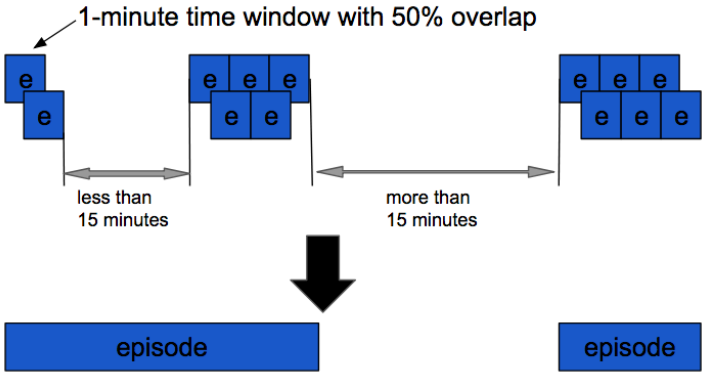

Fig. 18. Stage B aggregation (e indicates time window labeled as eating; episode indicates eating episode)

a mouthful usually lasts longer than 3 seconds, we chose to aggregate prediction results of twenty 3-second time windows to a result every 1 minute according to a threshold: if more than 10% of the windows in a minute were labeled *eating*, we labeled that minute as *eating* (shown in Figure 17). The evaluation results in Section 6.1 are based on the results after Stage A aggregation. Additionally, in Stage B, we aggregated 1-minute prediction results from Stage A to *eating episodes*, which can last for several minutes. We used 50% overlap between consecutive 1-minute time windows. We used one parameter γ to achieve eating episodes: if the gap between two 1-minute time windows prediction result is less than γ, we merged them into one eating episode (shown in Figure 18). We chose γ = 15 minutes, which is same as δ used in our definition of eating episode (Section 2). The evaluation results in Section 6.2 are based on the results after stage B aggregation.

Table 3. Results when using field data only and combining additional eating data for training (mean value ± standard deviation)

Training data	Accuracy	Precision	Recall	Weighted accuracy	F1 score
Field data	0.928 ±0.042	0.757 ± 0.158	0.808 ±0.133	0.879 ±0.074	0.775 ±0.128
Field data with additional eating data	0.913 ±0.047	0.736 ± 0.155	0.724 ±0.224	0.834 ±0.108	0.707 ±0.174

5.6 Ground-truth Label Aggregation

We used a similar two-stage aggregation approach on the ground-truth data to obtain ground-truth labels of 1-minute windows and eating episodes, and used them for window-based evaluation (Section 6.1) and episode-based evaluation (Section 6.2), respectively. In Stage A, we aggregated the ground-truth labels using the same method and threshold as in Section 5.5. In Stage B, we merged the 1-minute ground-truth labels into eating episodes using our definition in Section 2.

6 PERFORMANCE EVALUATION

To evaluate the performance of our approach, we evaluated Auracle's accuracy at two levels of detail: how well Auracle detected short periods of eating (using 1-minute windows of data) and how well those windows were aggregated into longer eating episodes.

6.1 Window-based Evaluation

Using the LOPO cross validation from Section 5.1, Figure 15 shows how well our classifier detects eating and non-eating data windows, when we vary the number of top features, k, from 5 to 60. In the experiment, adding features improved the F1 score up to $k = 40$, after which adding more features yielded little-to-no improvement. To achieve a reasonably high F1 score and avoid high power consumption when we later run feature-extraction algorithms in a wearable platform, we chose to use the top 40 features (Table 1) for evaluation in Section 6.2 and implemented these features in the MCU of our prototype (Section 3.3).

We also tried adding the laboratory-based eating data we collected in Section 4.2 into the training data set for each iteration of LOPO cross validation, and explored whether it helped to improve results. Figure 16 shows the performance of the classification model for different feature set sizes. Table 3 shows summary metrics in the two above cases when using top 40 features. From the figure and table, we see that the addition of this data did not improve the classification performance. We speculate that the reason is eating behaviour of participants in the laboratory and free-living conditions are different. Participants sat and ate without many body movements in the laboratory, but they sometimes ate while moving (and even walking) in free-living conditions.

To better understand the difference between the eating data collected in the laboratory and free-living conditions, we conducted another experiment. We trained another LR classifier with all the field data and used all the laboratory eating data for testing. The data prepossessing and feature extraction and selection approach are same as those mentioned in Sections 5.2 and 5.3. We found this classifier could only recognize 61.9% of laboratory eating data as *eating* and misclassified other laboratory eating data as *non-eating*. As a result, adding eating data collected in the laboratory setting did not help the classifier to better recognize eating in free-living conditions.

Table 4. Results for episode-based evaluation using Jaccard similarity coefficient

	Ground truth	CD	MD	FD
Number	26	20	6	12
Maximum duration (minutes)	42	49	42	21.5
Minimum duration (minutes)	1.5	1.5	0	1
Mean duration (minutes) ± standard deviation (minutes)	17.8 ±11.9	19.7 ± 10.9	12.8 ±15.8	5.2 ±6.7

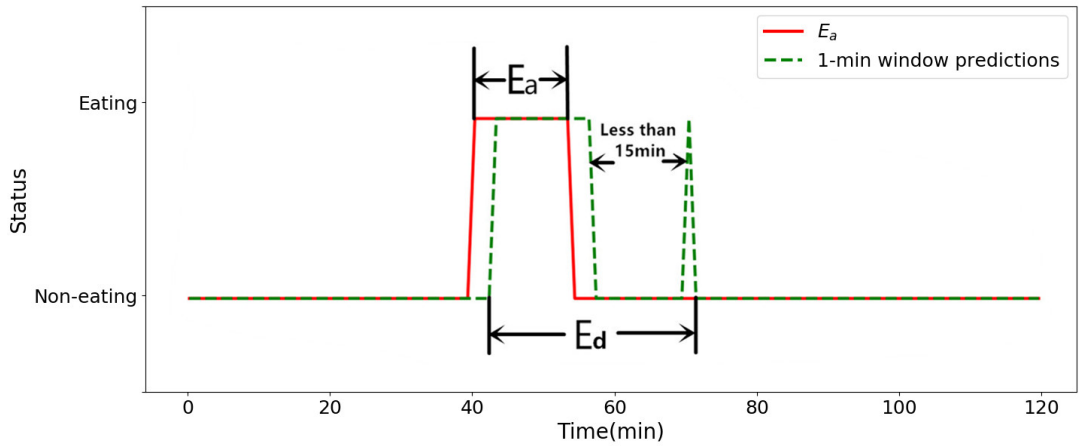

Fig. 19. An example of Missed Detection (E_a indicates the actual eating episode; E_d indicates the detected eating episode)

6.2 Episode-based Evaluations

According to our definition of *eating episode* in Section 2, there were 26 actual eating episodes in our field data, ranging in duration from 1 minute to 41 minutes. As shown in Table 4, when using Jaccard similarity coefficient as metrics, we correctly detected 20 eating episodes out of 26 and missed 6 eating episodes. We also falsely detected 12 eating episodes. For the Correct Detection (CD) cases, the mean and standard deviation of delay and duration difference were 3.0 ± 3.8 minutes and 5.3 ± 5.9 minutes. As we aggregated 1-minute time windows with 50% overlap to eating episodes, the resolution of episode-based evaluation is 30 seconds. In other words, our method will take at least 30 seconds to detect an eating episode.

To understand the source of Missed Detection (MD), we visually analyzed the data. In certain cases we identified the eating episode correctly, but within the subsequent (or previous) 15 minutes, the participant performed an activity (e.g., face touching) that our 1-minute window inferred as eating. This widened the span of the detected eating episode, with low overlap between the detected eating episode and actual eating episode. Thus, the Jaccard similarity coefficient in these scenarios was less than 55% and the eating episode was considered as MD. Figure 19 shows an example.

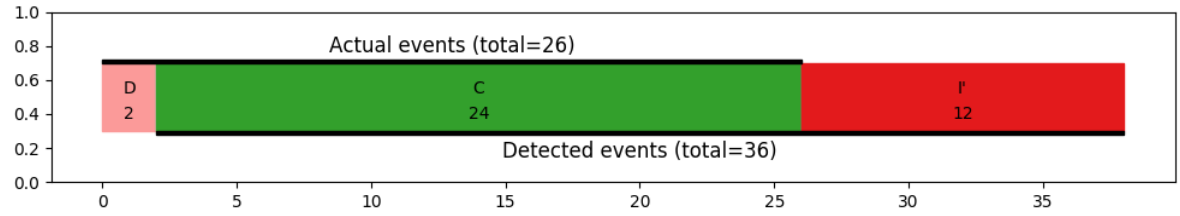

Fig. 20. Results for episode-based evaluation using Ward's metrics

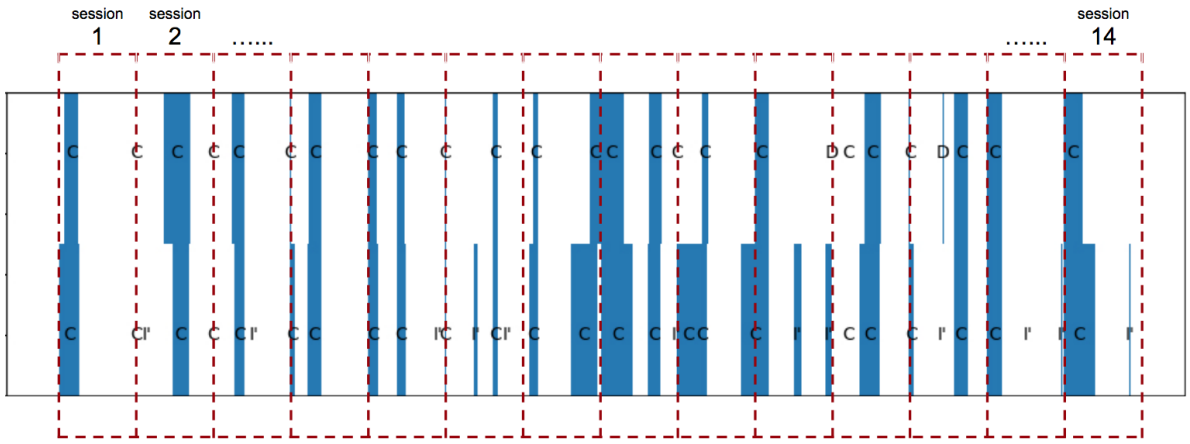

Fig. 21. Eating-episode assignment for 14 field-study sessions (each red box indicates a different session)

Moreover, we found 5 of the 12 False Detections (FD) lasted only for 1 minute, which is shorter than all the actual eating episodes. It may be necessary to better tune the thresholds used, or the metrics used, for future analysis.

In addition, we evaluated our method to detect *eating episodes* using Ward's metrics [36]. As shown in Figure 20, we achieved 24 correct detection (C) among 26 actual eating episodes with 12 false insertions (I') and 2 deletions (D). Figure 21 shows the *eating episode* assignment for 14 two-hour sessions in the field study.

7 POWER AND MEMORY EVALUATION

In this section, we estimate the power consumption of the Auracle during operation. Although the current prototype runs continuously at full power, we anticipate adding a wake-up circuit (Section 3.6) that would allow the MCU to remain in a lower-power, sleep mode when no sound is detected. We model the power consumption of the Auracle, with that addition, but must first measure the consumption of the current prototype. We used a Monsoon Power Monitor (Monsoon Solutions Inc., FTA22J) to conduct all power measurements. For each measurement, we use the Monsoon to recorded power data for half a hour at each activity level, from which we calculated the average power consumption.

We first define three different modes: verbose mode P_v, development mode P_d and realistic mode P_r. In *verbose mode*, the MCU logs both raw data and summary data to the SD card whenever it is not sleeping. In *development mode*, the MCU logs summary data to the SD card whenever it is not sleeping. In *realistic mode*, the MCU

Table 5. Average power consumption of each component

	Average Power Draw (mW)
Sleep state (P_s)	0.89
Data processing (P_d)	+18.29
Summary data logging (P_{c1})	+2.29
Raw data and summary data logging (P_{c2})	+7.28
BLE (P_b)	+3.37

continuously transmits prediction results through Bluetooth Low Energy (BLE) to a smart phone (and logs no data to SD card).

We estimated the power consumption in verbose mode P_v, development mode P_d and realistic mode P_r as follows:

$P_v = S * P_s + (1 - S) * (P_d + P_{c1})$

$P_d = S * P_s + (1 - S) * (P_d + P_{c2})$

$P_r = S * P_s + (1 - S) * P_d + P_b$

where S is the fraction of time spent sleeping, and P_s indicates the power consumption when the system is sleeping. We estimated P_s by summing the power consumption of the wake-up circuit (0.5 mW – as shown in Section 3.6), the power consumption of the contact microphone (0.33 mW – as shown in Section 3.1) and the power consumption of the MCU in standby mode (0.06 mW – as measured by Monsoon). By summing the power consumption of these three parts, We achieved P_s to be 0.89 mW. Based on the fraction of 3-second windows when the audio signal was below threshold in the field data we collected (Section 4.1.1), we estimate that S = 0.503.

P_d indicates the power consumption when the MCU samples sensor data, and runs feature extraction and classification algorithms on chip. We achieved P_d by directly measuring the power consumption of our PCB when data is processing on board.

P_{c1} indicates the power required to write the raw data (500 Hz sampling rate, 1000 bytes written / second) and summary data (feature values and prediction results; less than 200 bytes written / 3 seconds) to the SD card. P_{c2} indicates the power required to log only the summary data to SD card. We determined both P_{c1} and P_{c2} by calculating the difference in the power consumption of our PCB with and without SD card writing enabled.

P_b indicates BLE power consumption of our PCB when transmitting only classification results (2 bytes / 3 seconds) to an IPhone via BLE. We used the TI BLE-Stack software development kit to interface with the on-chip BLE radio, and the LightBlue[4] iOS and Android app to receive the data on smartphone. We determined P_b by calculating the difference in the power consumption of our PCB with and without BLE transmission enabled.

Based on above assumptions, we estimated the power consumption of each component in Table 5 and power consumption in each system mode in Table 6.

Auracle is powered by 3.3 V. Assuming use of a 110 mAh battery, we estimated Auracle can last 27.6 hours, 34.0 hours, and 28.1 hours in *verbose mode*, *development mode*, and *realistic mode*, respectively.

We also implemented our feature extraction and classification algorithms in the 20 KB SRAM of the MCU. Based on our measurement of the memory usage, we used 8.2KB SRAM when the MCU is in sleep date, and 19.2KB SRAM during other periods.

[4]https://punchthrough.com/

Table 6. Average power consumption in each system mode

System Mode	Average Power Draw (mW)
Verbose mode (P_v)	13.16
Development mode (P_d)	10.68
Realistic mode (P_r)	12.91

8 DISCUSSION AND FUTURE WORK

Handling misclassification: To identify the reasons that lead to misclassification of eating as non-eating and vice versa, we watched the videos during all the periods that were misclassified by our system. Some scenarios where false positives occurred include instances when the monitored individual was talking while walking, continuously touching face, excessively moving body, or making constant contact between neck or hoods and the mechanical housing. We also observed that several false negatives occurred when the individual was eating while walking or eating a soft food item like yogurt. We found that among all these reasons, the motion artifacts caused by walking and body movement played an important role in the misclassification. We believe that adding an accelerometer or an IMU to Auracle in the future may reduce the effect of the motion artifacts. One possible technique to reduce classification errors is to design non-standard features based on the data. In the future, we intend to explore such features.

Personalized modeling: As shown in Section 5, we used the same feature set and a general classification model for all participants, and evaluated the performance using LOPO cross validation. We assumed a general feature set and classification model would be preferred over a personalized model for most health-science projects, as researchers do not have to collect training data every time there are new participants. In certain cases, however, researchers may require higher eating-detection performance than we show in Section 6. In the future, we plan to explore whether personalization of feature set or classification model can help to further improve the eating-detection results in free-living conditions.

Additional sensing modality: Auracle relies heavily on chewing detection. Based on the chewing action, Auracle determines whether a person is eating. However, if a participant performed an activity with a significant amount of chewing but no swallowing (e.g., chewing gum), which is not 'eating' based on our definition, our system may output false positives. In the future, fusing data from additional sensors (e.g., a throat microphone for swallowing detection or wrist-worn devices for eating gesture recognition) might help handle situations that involve chewing but are not eating.

Day-long monitoring: The goal of a system like Auracle is to monitor an individual throughout the day and identify periods where the person was eating. In this project, we performed field studies where each session lasted for 2 hours. In the future, we plan to conduct day-long field studies with Auracle. As noted in Section 3.6, we plan to add a wake-up circuit to our AFE. This optimization will ensure Auracle can monitor an individual's eating behavior for more than a waking day. In the future day-long deployment, we will be able to measure the power consumption more accurately for different scenarios. Moreover, we also plan to make the mechanical housing comfortable enough for day-long experiments.

Real-time intervention: During the field data collection (Section 4.1), we used a micro-SD card to store all the raw data. However, the MCU used in Auracle has an on-chip BLE module, which we plan to use to communicate with a smartphone. Using smartphone apps, we can provide real-time interventions to users or collect additional contextual or self-report data.

Mechanical design: The current design of Auracle works for individuals with standard head-shapes and is compatible with eyeglasses. However, we noticed that the standard deviation of F1 score for eating detection results among all the participants is relatively large (0.128 as shown in Table 3). One reason could be that the pressure between contact mic and the skin of some participants was significantly different from that of others. More specifically, the mechanical housing was either too tight or to loose for them. In the future, we will further refine our mechanical housing design (e.g., enable the adjustment of both the distance and angle between the microphone and skin) to ensure it can fit better for different head shapes. More personalization of mechanical design can be explored to achieve this goal.

9 RELATED WORK

Health-science researchers are interested in various measurable parameters including eating-specific data such as the time, duration and rate of eating, and meal-specific data such as food quantity, food group classification and calorie estimation [24]. For all of these parameters, accurate recognition of *when* people eat is the foundation of effective automatic dietary monitoring (ADM) systems. Several review papers [2, 14, 24, 35] covered aspects of eating detection and summarized ADM systems developed. Here we focus only on technologies developed to recognize when people eat using wearable sensors; the variety of sensors explored include acoustic, physiological, piezoelectric, proximity, visual, inertial and fusion approaches in both laboratory and free-living scenarios.

9.1 Laboratory Studies

Below is a brief overview of existing methods evaluated in laboratory conditions, which we categorize into two types: *acoustic* and *other*. Acoustic approaches can be further classified as *air* (using microphones designed for recording sound from the air) and *contact* (using microphones designed for recording sounds conducted through the body). For the second type, these microphones typically require direct contact with the skin.

Air-conducted sound: Amft et al. evaluated the air-conducted sound intensity of chewing and speech when a microphone is placed at different locations on the body [1]. They identified the optimal location to be the inner ear, directed towards the eardrum, rather than 2 cm in front of mouth, at the cheek, collar bone, behind the outer ear or 5 cm in front of the ear canal opening. Since then, much effort has been put in developing ADM systems using air microphones positioned in the ear [17, 22, 27]. Sazonov et al. explored the option of using the neck as the sensing locations and achieved 84.7% average weighted accuracy in detection of swallowing events [27, 28].

Body-conducted sound: To capture and recognize a diverse range of body-conducted sounds, including eating sounds, Rahman et al. designed a mobile sensing system consisting of a customized contact microphone placed on the neck, an ARM microcontroller and Android smartphone [25]. They achieved an average recall of 71.2% for a nine-class classification of different body sounds (eating, drinking, deep breathing, clearing throat, coughing, sniffling, laugh, silence, speech) in laboratory conditions. Several other acoustic-based ADM systems also used body-conducted sound recorded from the neck [20, 27, 37] or in the ear canal [31] to detect swallowing or chewing events.

Compared with normal air microphones, contact microphones capture internal vibrations directly from the body surface and are naturally immune to ambient noise, making these sensors promising for eating detection in out-of-lab, free-living scenarios, where ambient noise is variable and can be large in magnitude. Because we are most interested in detecting eating in free-living scenarios, Auracle was designed with a contact microphone as the eating-detection sensor.

Other eating-detection approaches evaluated in laboratory conditions include physiological, piezoelectric, proximity and fusion approaches. The two primary physiological signals explored for eating detection include electroglottography (EGG) and electromyography (EMG). EGG sensors capture the motion-induced variations of electrical impedance recorded between two electrodes positioned on the larynx [15]. Farooq et al. placed an EGG setup around participants' necks to capture swallowing events and achieved an average per-epoch classification

accuracy of 90.1% [11]. Zhang et al. fused three EMG electrodes into an eyeglasses frame to capture muscle signals during eating [38, 39, 41]. Using dry fabric electrodes, they could detect chewing with a precision and recall of 80%. Piezoelectric sensors can produce a voltage at their terminals in response to mechanical stress [13]. To automatically monitor eating behavior, piezoelectric film sensors were placed on the jaws [12] or throat [13] for motion capture. Kalantarian et al. developed a necklace to capture swallowing events [13] and were able to detect more than 81.4% of swallows. Finally, many systems fused two or more of these approaches with the aim of improving automatic intake monitoring systems [9, 18, 21]. Merck et al. presented a multi-sensor study of eating recognition, which combines head motion, wrist motion and audio [18]. In their study, using audio sensing alone achieved 92% precision and 89% recall in finding meals, while motion sensing was needed to find individual intakes.

9.2 Field Studies

Ultimately, we aim to develop an ADM system that works in real life; thus field studies in free-living scenarios will be crucial for evaluating an ADM system. Bedri et al. evaluated optical, inertial and acoustic sensors, and ended up using a behind-the-ear inertial sensor and achieved an F1 score of 80.1% for detecting eating episodes [3]. Using a proximity sensor, Chun et al. developed a necklace that captures head and jawbone movement [9]. They achieved 78.2% precision and 72.5% recall for detecting eating episodes in the free-living study. In another ADM system, Outer Ear Interface (OEI), three proximity sensors are encapsulated in an earpiece to monitor jaw movement by measuring ear-canal deformation during chewing [4, 5]. In a field experiment, OEI classified five-minute segments of time as eating or non-eating with 93% (user dependent) and 82% accuracy (user independent) [4]. Thomaz et al. collected wrist-mounted audio data and tried to use ambient sound to infer eating activities [34]. Their system was able to identify meal eating with an F1 score of 79.8% in a person-dependent evaluation. Sen et al. built and tested an approach based on wrist motion and achieved false-positive and false-negative rates of 6.5% and 3.3% respectively [29, 30]. Zhang et al. evaluated smart eyeglasses they proposed in free-living scenarios and achieved precision and recall more than 77% for chewing detection [40]. Mirtchouk et al. experimented with different combinations of motion (head, wrist) and audio (air microphone) data collected in laboratory and free-living conditions [19]. They found a combination of sensing modalities (audio, motion) was needed; yet sensor placement (head vs. wrist) was not critical.

In these previous field studies, researchers logged field data in free-living scenarios and ran offline experiments. Even though we currently run experiments offline, the Auracle can do real-time eating detection. We have developed an ADM system that can locally capture, process, and classify sensor data collected in out-of-lab, day-long, free-living scenarios.

10 CONCLUSION

In this paper, we propose Auracle, a wearable system for eating detection in free-living scenarios. We first implemented the Auracle hardware, which includes a contact microphone, battery, wearable mechanical housing and PCB with data acquisition function. Using this device, we collected field data with 14 participants for 32 hours in free-living scenarios and additional eating data with 10 participants for 2 hours in laboratory scenarios, respectively. Based on these data, we designed a data-processing pipeline and evaluated its performance using LOPO cross validation. We achieved accuracy exceeding 92.8% and F1 score exceeding 77.5% of eating detection, and successfully detected 20-24 eating episodes (depending on the metrics) out of 26 in free-living conditions. Finally, we implemented the data-processing method on our prototype and estimated the power consumption of Auracle. We anticipate Auracle can last 28.1 hours with a 110 mAh battery in realistic mode. Please follow us at auracle-project.org.

ACKNOWLEDGMENTS

We thank the other members of Auracle project (Xing-Dong Yang, Jun Gong, Byron Lowens, Nathan Grice) for their feedback and help. We also appreciate the advice about eating-behavior research from Tauhidur Rahman, Tanzeem Choudhury, John Batsis, William Kelley, Lisa Marsch, Tam Vu, Adam Hoover, Eric Muth, Tobias Kowatsch and Joseph Paydarfar.

This research results from a research program at Dartmouth College and Clemson University, supported by the National Science Foundation under award numbers CNS-1565269 and CNS-1565268. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

[1] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. 2005. Analysis of Chewing Sounds for Dietary Monitoring. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. https://doi.org/10.1007/11551201_4

[2] O. Amft and G. Troster. 2009. On-Body Sensing Solutions for Automatic Dietary Monitoring. *IEEE Pervasive Computing* 8, 2 (April 2009), 62–70. https://doi.org/10.1109/mprv.2009.32

[3] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj P. Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Y. Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interactive, Mobile and Wearable Ubiquitous Technology* 1, 3 (Sept. 2017). https://doi.org/10.1145/3130902

[4] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. Detecting Mastication: A Wearable Approach. In *Proceedings of the ACM on International Conference on Multimodal Interaction*. https://doi.org/10.1145/2818346.2820767

[5] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. A wearable system for detecting eating activities with proximity sensors in the outer ear. In *Proceedings of the ACM International Symposium on Wearable Computers*. ACM, 91–92. https://doi.org/doi:10.1145/2802083.2808411

[6] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency, In Annals of Statistics. *Annals of Statistics*. http://dx.doi.org/DOI:2010.1214/aos/1013699998

[7] Shengjie Bi, Tao Wang, Ellen Davenport, Ronald Peterson, Ryan Halter, Jacob Sorber, and David Kotz. 2017. Toward a Wearable Sensor for Eating Detection. In *Proceedings of the 2017 Workshop on Wearable Systems and Applications (WearSys)*. ACM Press, 17–22. https://doi.org/10.1145/3089351.3089355

[8] Thomas Bodenheimer, Ellen Chen, and Heather D. Bennett. 2009. Confronting The Growing Burden Of Chronic Disease: Can The U.S. Health Care Workforce Do The Job? *Health Affairs* 28, 1 (1 Jan. 2009), 64–74. https://doi.org/10.1377/hlthaff.28.1.64

[9] Keum S. Chun, Sarnab Bhattacharya, and Edison Thomaz. 2018. Detecting Eating Episodes by Tracking Jawbone Movements with a Non-Contact Wearable Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (26 March 2018), 1–21. https://doi.org/10.1145/3191736

[10] Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover. 2014. Detecting periods of eating during free-living by tracking wrist motion. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (July 2014), 1253–1260. http://view.ncbi.nlm.nih.gov/pubmed/24058042

[11] Muhammad Farooq, Juan M. Fontana, and Edward Sazonov. 2014. A novel approach for food intake detection using electroglottography. *Physiological measurement* 35, 5 (May 2014), 739–751. https://doi.org/10.1088/0967-3334/35/5/739

[12] Muhammad Farooq and Edward Sazonov. 2016. A Novel Wearable Device for Food Intake and Physical Activity Recognition. *Sensors* 16, 7 (11 July 2016). https://doi.org/10.3390/s16071067

[13] Haik Kalantarian, Nabil Alshurafa, and Majid Sarrafzadeh. 2014. A Wearable Nutrition Monitoring System. In *Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. https://doi.org/10.1109/BSN.2014.26

[14] Haik Kalantarian, Nabil Alshurafa, and Majid Sarrafzadeh. 2017. A Survey of Diet Monitoring Technology. *IEEE Pervasive Computing* 16, 1 (Jan. 2017), 57–65. https://doi.org/10.1109/mprv.2017.1

[15] F. L. E. Lecluse, M. P. Brocaar, and J. Verschuure. 1975. The Electroglottography and its Relation to Glottal Activity. *Folia Phoniatrica et Logopaedica* 27, 3 (1975), 215–224. https://doi.org/10.1159/000263988

[16] Rebecca M. Leech, Anthony Worsley, Anna Timperio, and Sarah A. McNaughton. 2015. Characterizing eating patterns: a comparison of eating occasion definitions. *The American Journal of Clinical Nutrition* (7 Oct. 2015). https://doi.org/10.3945/ajcn.115.114660

[17] Jindong Liu, Edward Johns, Louis Atallah, Claire Pettitt, Benny Lo, Gary Frost, and Guang-Zhong Yang. 2012. An Intelligent Food-Intake Monitoring System Using Wearable Sensors. In *Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, 154–160. https://doi.org/10.1109/bsn.2012.11

[18] Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. 2016. Multimodality Sensing for Eating Recognition. In *Proceedings of the EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM Press. https://doi.org/10.4108/eai.16-5-2016.2263281

[19] Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. 2017. Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (Sept. 2017), 85+. https://doi.org/10.1145/3131894

[20] Temiloluwa Olubanjo and Maysam Ghovanloo. 2014. Real-time swallowing detection based on tracheal acoustics. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4384–4388. https://doi.org/10.1109/icassp.2014.6854430

[21] Vasileios Papapanagiotou, Christos Diou, Lingchuan Zhou, Janet van den Boer, Monica Mars, and Anastasios Delopoulos. 2016. A novel chewing detection system based on PPG, audio and accelerometry. *IEEE Journal of Biomedical and Health Informatics* (2016). https://doi.org/10.1109/jbhi.2016.2625271

[22] Sebastian Päßler, Matthias Wolff, and Wolf-Joachim Fischer. 2012. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological Measurement* 33, 6 (01 June 2012), 1073–1093. https://doi.org/10.1088/0967-3334/33/6/1073

[23] Mitesh S. Patel, David A. Asch, and Kevin G. Volpp. 2015. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA* 313, 5 (03 Feb. 2015), 459–460. http://view.ncbi.nlm.nih.gov/pubmed/25569175

[24] Temiloluwa Prioleau, Elliot Moore, and Maysam Ghovanloo. 2017. Unobtrusive and Wearable Systems for Automatic Dietary Monitoring. *IEEE Transactions on Biomedical Engineering* 64, 9 (Sept. 2017), 2075–2089. https://doi.org/10.1109/tbme.2016.2631246

[25] Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A Mobile System for Sensing Non-speech Body Sounds. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. https://doi.org/10.1145/2594368.2594386

[26] Sasank Reddy, Andrew Parker, Josh Hyman, Jeff Burke, Deborah Estrin, and Mark Hansen. 2007. Image browsing, processing, and clustering for participatory sensing. In *Proceedings of the Workshop on Embedded Networked Sensors (EmNets)*. ACM Press, 13–17. https://doi.org/10.1145/1278972.1278975

[27] Edward Sazonov, Stephanie Schuckers, Paulo Lopez-Meyer, Oleksandr Makeyev, Nadezhda Sazonova, Edward L. Melanson, and Michael Neuman. 2008. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. *Physiological measurement* 29, 5 (May 2008), 525–541. https://doi.org/10.1088/0967-3334/29/5/001

[28] Edward S. Sazonov, Oleksandr Makeyev, Stephanie Schuckers, Paulo Lopez-Meyer, Edward L. Melanson, and Michael R. Neuman. 2010. Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior. *IEEE Transactions on Bio-medical Engineering* 57, 3 (March 2010), 626–633. http://view.ncbi.nlm.nih.gov/pubmed/19789095

[29] Sougata Sen, Vigneshwaran Subbaraju, Archan Misra, Rajesh K. Balan, and Youngki Lee. 2015. The case for smartwatch-based diet monitoring. In *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 585–590. https://doi.org/10.1109/percomw.2015.7134103

[30] Sougata Sen, Vigneshwaran Subbaraju, Archan Misra, Rajesh K. Balan, and Youngki Lee. 2017. Experiences in Building a Real-World Eating Recogniser. In *Proceedings of the International on Workshop on Physical Analytics (WPA)*. ACM, 7–12. https://doi.org/10.1145/3092305.3092306

[31] Masaki Shuzo, Shintaro Komori, Tomoko Takashima, Guillaume Lopez, Seiji Tatsuta, Shintaro Yanagimoto, Shin'ichi Warisawa, Jean-Jacques Delaunay, and Ichiro Yamada. 2010. Wearable Eating Habit Sensing System Using Internal Body Sound. *Journal of Advanced Mechanical Design, Systems, and Manufacturing* 4, 1 (2010), 158–166. https://doi.org/10.1299/jamdsm.4.158

[32] Mingui Sun, Lora E. Burke, Zhi H. Mao, Yiran Chen, Hsin C. Chen, Yicheng Bai, Yuecheng Li, Chengliu Li, and Wenyan Jia. 2014. eButton: A Wearable Computer for Health Monitoring and Personal Assistance. In *Proceedings of the Annual Design Automation Conference*. https://doi.org/10.1145/2593069.2596678

[33] Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D. Abowd. 2013. Feasibility of identifying eating moments from first-person images leveraging human computation. In *Proceedings of the International SenseCam & Pervasive Imaging Conference (SenseCam)*. ACM Press, 26–33. https://doi.org/10.1145/2526667.2526672

[34] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D. Abowd. 2015. Inferring Meal Eating Activities in Real World Settings from Ambient Sounds. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. ACM Press, 427–431. https://doi.org/10.1145/2678025.2701405

[35] Tri Vu, Feng Lin, Nabil Alshurafa, and Wenyao Xu. 2017. Wearable Food Intake Monitoring Technologies: A Comprehensive Review. *Computers* 6, 1 (2017). https://doi.org/10.3390/computers6010004

[36] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. 2011. Performance Metrics for Activity Recognition. *ACM Trans. Intell. Syst. Technol.* 2, 1, Article 6 (Jan. 2011), 23 pages. https://doi.org/10.1145/1889681.1889687

[37] Koji Yatani and Khai N. Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In *Proceedings of the ACM Conference on Ubiquitous Computing (UbiComp)*. 341–350. https://doi.org/10.1145/2370216.2370269

[38] Rui Zhang and Oliver Amft. 2016. Bite Glasses: Measuring Chewing Using EMG and Bone Vibration in Smart Eyeglasses. In *Proceedings of the ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/2971763.2971799

[39] Rui Zhang and Oliver Amft. 2016. Regular-look Eyeglasses Can Monitor Chewing. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 389–392. https://doi.org/10.1145/2968219.2971374

[40] Rui Zhang and Oliver Amft. 2018. Monitoring Chewing and Eating in Free-Living Using Smart Eyeglasses. *IEEE Journal of Biomedical and Health Informatics* 22, 1 (Jan. 2018), 23–32. https://doi.org/10.1109/jbhi.2017.2698523

[41] Rui Zhang, Severin Bernhart, and Oliver Amft. 2016. Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses. In *Proceedings of IEEE International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. https://doi.org/10.1109/bsn.2016.7516224

Received May 2018; revised July 2018; accepted September 2018