Computers in Biology and Medicine 65 (2015) 1–9

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Audio-based detection and evaluation of eating behavior using the smartwatch platform

Haik Kalantarian, Majid Sarrafzadeh

Wireless Health Institute, Department of Computer Science, University of California, Los Angeles, United States

ARTICLE INFO

Article history:
Received 2 May 2015
Accepted 16 July 2015

Keywords:
Signal processing
Wireless Health
Nutrition
Smartwatch
Machine learning
Pervasive computing

ABSTRACT

In recent years, smartwatches have emerged as a viable platform for a variety of medical and health-related applications. In addition to the benefits of a stable hardware platform, these devices have a significant advantage over other wrist-worn devices, in that user acceptance of watches is higher than other custom hardware solutions. In this paper, we describe signal-processing techniques for identification of chews and swallows using a smartwatch device's built-in microphone. Moreover, we conduct a survey to evaluate the potential of the smartwatch as a platform for monitoring nutrition. The focus of this paper is to analyze the overall applicability of a smartwatch-based system for food-intake monitoring. Evaluation results confirm the efficacy of our technique; classification was performed between apple and potato chip bites, water swallows, talking, and ambient noise, with an *F*-measure of 94.5% based on 250 collected samples.

1. Introduction

There is little doubt that obesity is associated with various negative health outcomes such as an increased risk for stroke, diabetes, various cancers, heart disease, and other conditions. In 2008, medical costs associated with obesity were estimated to exceed $147 billion, with over one-third of adults in the United States estimated to be obese [1]. The two major contributors to weight gain are an inactive lifestyle and poor diet. Though the former has been addressed by many wearable devices in recent years both in research and the consumer electronics field, few works exist on automatic detection of dietary habits in an inconspicuous form-factor [2–4]. Instead, characterization of an individual's eating habits is possible through manual record keeping such as food diaries, 24-h recalls, and food frequency questionnaires. However, these approaches suffer from low accuracy, high user burden, and low rates of long-term compliance. Wireless health-monitoring technologies have the potential to promote healthy behavior and address the ultimate goal of enabling better lifestyle choices.

In recent years, several electronic devices have been proposed for monitoring dietary habits. However, most works attempt to characterize eating from patterns in chewing and swallow counts, and very few proposed attempt to identify the nutritive properties of the foods themselves. Therefore, a fundamental question in the field of electronic food monitoring is the validity of chew and swallow counts as a heuristic for estimation of Caloric intake. A recent work by Fontana et al. [5] addresses this issue by comparing several different techniques for estimation of Caloric intake: weighed food records (gold standard), diet diaries, and electronic sensor-based measurements of chews and swallows. Though the study was conducted under constrained conditions, the results suggest that chew and swallow counts may be a promising alternative to manual self-reporting techniques.

While many audio-based nutrition monitors are novel from a perspective of algorithmic techniques, they generally propose custom hardware solutions or bulky non-standard equipment which are of limited use outside of clinical environments. The primary challenge of monitoring a subject's eating habits is creating a system that provides passive monitoring of behavior, presenting a low level of user burden and providing no compromises on comfort and appearance: even the most accurate techniques have very limited scope if they do not encourage repeated use from users.

Recently, smartwatches have emerged as a new platform that provide several promising applications such as wrist-worn activity monitoring, heart rate tracking, and even stress measurement. Watch usage is well established and has a high level of social acceptance, as confirmed not only by our personal studies but by their ubiquity in day-to-day life. Furthermore, the smartwatch platform provides many useful services that can collectively improve user adherence rates, rather than specialized devices

E-mail addresses: kalantarian@cs.ucla.edu (H. Kalantarian), majid@cs.ucla.edu (M. Sarrafzadeh).

http://dx.doi.org/10.1016/j.compbiomed.2015.07.013
0010-4825/

with just one application that may fail to sustain a user's interest. These devices contain a multitude of sensors including but not limited to: a microphone, camera, accelerometer, and gyroscope. Due to the ubiquity of watches, this technology can be used for various wireless health monitoring applications discretely, with low user burden. Furthermore, from a user-acceptance standpoint, these systems have a clear advantage over other proposed solutions based on custom hardware, which may require that these bulky and non-standard devices be worn in unconventional ways. Clearly, the multitude of sensors available on the smartwatch platform, wireless connectivity, as well as the comfort and social acceptance of the form-factor warrant further study into their potential applications in the medical and health-monitoring domain.

This paper explores the idea of tracking eating habits using a custom Android application on the smartwatch platform. Though identifying eating-related gestures using wrist-worn devices is a viable application of the watch, the focus of our work is to explore the idea of using audio to detect eating behavior based on bites, rather than swallows as other works have done. A high-level system architecture is presented in Fig. 1. The first step is audio-based acquisition of eating-related sounds such as bites, acquired from the microphone integrated within the smartwatch. After data acquisition, the audio is processed using various classifiers to identify the sound and infer the associated activity.

In addition, we conducted two surveys in order to evaluate the potential of the smartwatch platform for nutrition monitoring. The surveys were conducted online, with 221 respondents in the first and 55 in the second. In the first survey, we asked subjects various questions about their general habits with respect to watches. For example, subjects were asked which hand they prefer to wear a watch, and whether they were willing to wear a watch on the opposite hand on which they were accustomed. In the second survey, respondents provided information about their opinion of various wearable form factors. Fifty-five subjects rated their receptiveness to smartwatches, necklace-based wearables, custom wrist-worn hardware, and smart glasses.

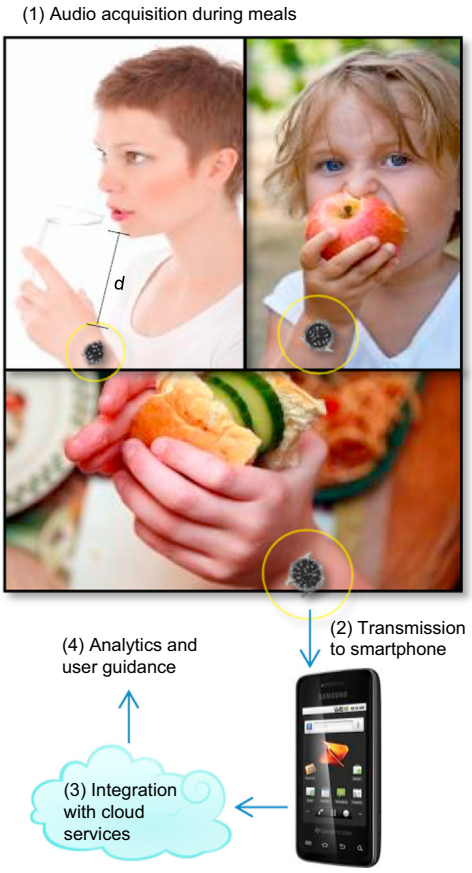

Fig. 1. A high level architecture of the proposed system is shown above. Many different forms of eating can be detected using a smartwatch, provided the appropriate hand is used and the watch is brought close enough to the mouth.

This paper is organized as follows. Section 2 provides an overview of related work, primarily in the scope of audio-based analysis of eating habits. Section 3 describes the hardware architecture of the system, based on around the Samsung Galaxy Gear smartwatch. Section 4 describes the algorithmic aspects of our work. Section 5 outlines the experimental procedure. Section 6 provides results and Section 7 provides concluding remarks.

2. Related work

The use of audio signals for analysis of swallows or eating behavior has been explored in several other works. For example, the work in [6] uses acoustic data acquired from a small microphone placed near the bottom of the throat. Their system is coupled with a strain gauge placed near the ear. In this work, acquired data is manually labelled to provide a benchmark for future classification. Analyzing wave shape in the time domain or feature extraction and machine learning [7] has resulted in an 86% swallow detection accuracy in an in-lab controlled environment. In [8], Pler et al. proposed a system geared towards patients living in ambient assisted living conditions and used miniature electret microphones which were integrated into a hearing aid case, and placed in the ear canal. In [9], the authors are able to achieve a food detection accuracy of 79% using hidden Markov models based on data acquired from microphones in the ear canal.

In the work by Amft et al. in [10], authors analyze bite weight and classify food acoustically from an earpad-mounted sensor. However, sound-based chewing recognition accuracy was low, with a precision of 60–70%. In [11], the authors present a similar earpad-based sensor design to monitor chewing sounds. Food grouping analysis revealed three significant clusters of food: wet and loud, dry and loud, soft and quiet. An overall recognition accuracy of over 86.6% was achieved.

Though the signal processing aspects of this application are relatively well developed, our work differs from prior approaches in two significant ways. First, we propose the use of an audio spectrogram for representing the changes in the frequency distribution of the signal over time, which is subsequently subdivided into bins and used for feature extraction and selection. Therefore, the classifier can distinguish between different foods based on the frequency distribution of the signal over time. Secondly, prior works rely on cumbersome multi-sensor hardware approaches that have little utility out of laboratory environments, while our algorithm runs on off-the-shelf Samsung hardware. Lastly, we show how the extensive openSMILE feature extraction tool for analyzing human vocalizations can be applied to other domains for classification of sounds unrelated to speech patterns.

3. System architecture

Our proposed system does not require any custom hardware: the Android application runs on Samsung Galaxy Gear smartwatch running Android 4.2.1. This device, shown in Fig. 2, features an 800 MHz ARM-based processor, 512 MB of RAM, and a 320×320 pixel 1.6 inch display. The device also supports transfer of data using the Bluetooth LE protocol, and can be configured to access the Internet using Bluetooth tethering with compatible

Fig. 2. The Samsung Galaxy Gear is the chosen hardware platform, which runs Android 4.2.1 and features an 800 MHz ARM-based processor.

smartphones. Once the on-board algorithm detects that a bite has been taken, a web-service call is made to store the data in a database for access by caregivers. In the case of algorithm inaccuracies and errors, subjects are permitted to manually make modifications and add annotations to the data.

Data was recorded using the Samsung Galaxy Gear microphone in MPEG-4 Part 14 (m4a) format at a rate of 96 kbps, as prior research has shown that the spectral energy for many common foods is between 0 and 10 kHz, with highest amplitude ranges between 1 and 2 kHz for water [12,13]. Of note is the availability of additional sensors on the Samsung Gear platform, including accelerometers and gyroscopes, which can be used for improved classification accuracy in future work, based on hand and wrist motion associated with eating behavior.

The Samsung Galaxy Gear has a 315 mAh capacity battery. This is significant because audio recording and transmission is a relatively energy-intensive task that may compromise battery life. This is partially mitigated by the decision to acquire data at a low sample rate. A more comprehensive evaluation is provided in Section 6.

4. Algorithm design

4.1. Frequency-domain evaluation: liquids

We begin our algorithm analysis with the objective of detecting liquid ingestion using a smartwatch. Because we have a priori knowledge about the kind of data we would like to identify, we could pre-process the recorded data before classification, as we describe in this section.

Fig. 3 shows a spectrogram corresponding with an audio clip consisting of five water swallows acquired from the smartwatch. A spectrogram is a visual representation of the frequency spectrum over time, and is an ideal representation for extracting distinguishing features in many classification problems. A spectrogram is typically generated using a short-time Fourier transform (STFT) with a fixed window size, the squared magnitude of which yields the spectrogram. Fundamentally, a spectrogram allows easy identification of changes in the frequency spectrum of a signal, over time. Fig. 4 shows a more detailed comparison between a brief interval of noise (1 s) and a water swallow. Generally speaking, the data of interest is between 600 Hz and 1 kHz, as shown by the deviation between the signals at this time, and confirmed by the spectrogram shown in Fig. 3. We conclude that analysis of this frequency range is critical for classification of liquid swallows. This observation is confirmed by Fig. 5, which shows the transformation of an audio signal corresponding with 10 swallows. The top waveform is the original, while the bottom is the post-processed filter output in which noise is substantially reduced. This is achieved by band-pass filtering the audio data with cutoffs of 600 Hz and 1 kHz and a rolloff of 48 dB – meaning the amplitude decreases by 48 dB for each octave outside the filter threshold.

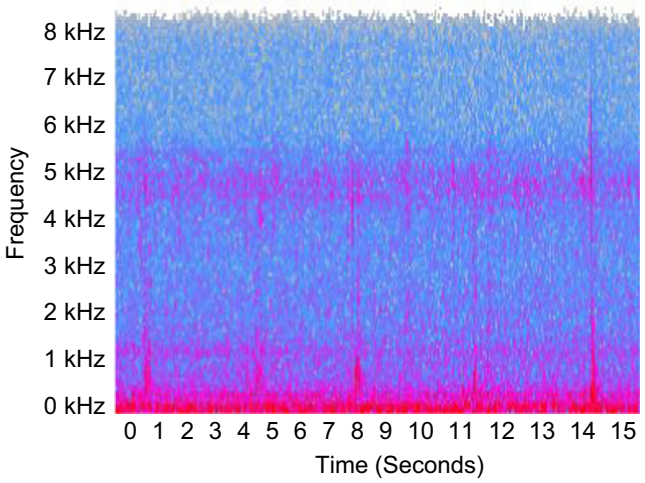

Fig. 3. A spectrogram of an audio clip consisting of five swallows, generated with a Hanning window of size 1024 samples. There is a visible change in the spectral density at points corresponding with swallows as shown above.

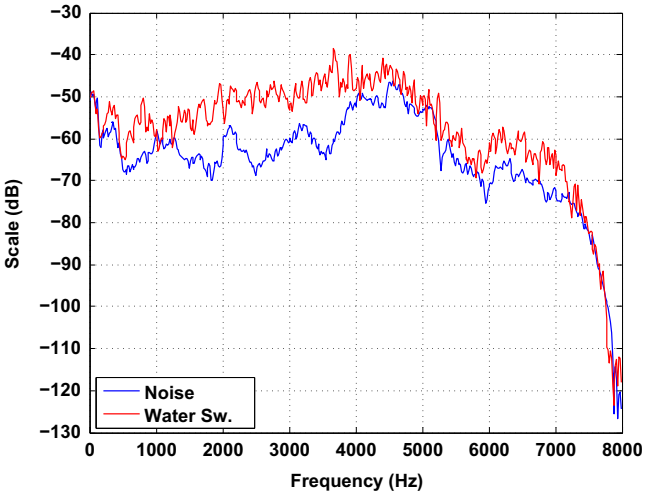

Fig. 4. Frequency distribution of a water swallow vs. silence (noise). This graph reveals that the frequency range between 500 Hz and 1000 Hz is the point of interest.

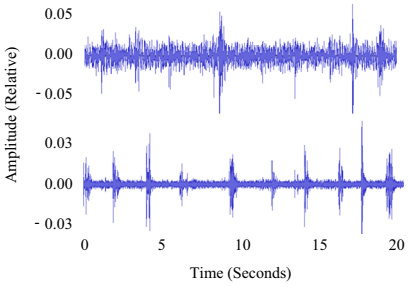

Fig. 5. Post-processing of the audio signal corresponding with water can dramatically improve signal-to-noise ratio. The top shows the original waveform. The bottom shows the waveform after a bandpass filter is applied.

While the resulting signal clearly shows the swallows, marked by pronounced peaks, this technique is not very generalizable to other foods besides water, because the data is pre-processed. In the case of the frequency distribution of a one second window around the initial bite of a potato chip, compared to an equal period of chewing, the amplitude of the bite signal is greater from 600 Hz to 4 kHz. However, the pattern is not as distinctive as for liquids, and may certainly vary between individuals with different eating styles. Therefore, a simplistic filter-approach may not be sufficient for foods with less uniformity. More significantly, removing a frequency band to simplify the recognition of one food may also remove crucial information necessary for identifying another, in systems which attempt to classify between very different food types. Therefore, a more generalizable approach is described in the next subsection.

4.2. Generalizable feature extraction

Detection of eating habits differs significantly from that of liquid consumption, as the smartwatch will not necessarily be near the throat during a swallow. When an individual is drinking water, the swallows happen almost immediately after each sip. However, chewing food takes a significant amount of time. Typically, the smartwatch would be brought toward the mouth during the first bite, after which it would be lowered once more during the chewing process. Once the individual swallows the food, it is difficult to predict the location of the microphone. Therefore, in these cases we attempt to identify when an individual bites into a food item rather than chewing. The smartwatch platform is particularly well suited for this application because the microphone will be nearest to the sound source during the times at which the signal is of interest. The proposed model must be flexible to identify biting and swallowing for many different foods and drinks, between individuals with varying eating styles.

The Munich open Speech and Music Interpretation by Large Space Extraction toolkit, known as openSMILE [14], is a feature extraction tool intended for producing large audio feature sets. This tool is capable of various audio signal processing operations such as applying window functions, FFT, FIR filterbanks, autocorrelation, and cepstrum. In addition to these techniques, openSMILE is capable of extracting various speech related features and statistical features. A partial list of extracted features is shown in Tables 1 and 2. More "low-level" audio-based features include frame energy, intensity, auditory spectra, zero crossing rate, and voice quality. Therefore, the capabilities of this tool are significantly more extensive than that of the spectrogram based approach described earlier, which relied only on statistical features from time–frequency decomposition. After data is collected from a variety of subjects eating several foods, feature selection tools can be used to identify strong features that are accurate predictors of swallows and bites for various foods, while reducing the dimensionality by eliminating redundant or weakly correlated features.

A microphone on a Smartwatch can either constantly record data, or be configured to record audio based on motion-based triggers indicative of eating-related gestures, in order to save battery life. The recorded audio is stored on a buffer in Smartwatch memory with storage for 4096 samples, corresponding with 0.25 s of data. Once the buffer is full, features are extracted using openSMILE (elaborated upon in subsequent sections), and the audio clip is classified divided into several distinct categories corresponding with the various foods the system has been trained to detect. A counter is incremented corresponding with the food type detected, which is necessary for long-term record keeping. In the event that eating behavior is detected, subsequent detection is disabled for a period of two seconds to prevent duplicate records caused by the same event. The algorithm is presented in Algorithm 1, with $\beta = 4096$ samples and $\tau = 2$ s.

Table 1
Partial list of openSMILE speech features from [20].

Speech-related features		
Signal energy	Loudness	Mel/Bark/Octave spectra
MFCC	PLP-CC	Pitch
Voice quality	Formants	LPC
Line spectral pairs	Spectral shape	CENS and CHROMA

Table 2
Partial List of openSMILE statistical features from [20].

Speech-related features		
Means	Extremes	Moments
Segments	Samples	Peaks
Zero crossings	Quadratic regression	Percentiles
Duration	Onset	DCT coefficient

Algorithm 1. Simplified classification scheme.

```
RecordAudio (Buffer);
if Buffer.Utilization = β then
    d = Buffer[1 : β];
    f = ExtractFeatures(d);
    s = {Water, Talk, Apple, Chips, Other};
    c = Classify(f, s);
Counter_c++;
if c ≠ Other then
    PauseRecording(τ)
```

To minimize the overlap between neighboring segments for performance reasons, the last 50 ms of buffer data are cleared after each classification activity, and classification resumes when the buffer is full once again (not shown).

5. Experimental procedure

5.1. Data collection for recognition

A total of 10 subjects were used for data collection, with ages ranging from 22 to 35 in order to develop a model for identifying swallows. The subjects included 8 males and two females. Each subject was asked to eat the following foods, in order: three apple slices with at least two bites per slice, one 8 oz. glass of room-temperature water, and one bag of potato chips. The moments at which the food was bitten into (or swallowed as in the case of the water) were manually annotated by the subject, though these events were clearly audible on the resulting waveform. The hand on which the smartwatch was worn was used to pick up the food items and water, which happened to be the left hand for all subjects.

Data collection took place in a laboratory environment which had a minimal level of background noise including talking and doors opening, most of which is barely audible in the recording. However, pre-recorded background noise from a public shopping square was combined with the original data, to produce clips that more accurately reflect a real-world use case. It was assumed that the background noise should be quieter than the original waveforms because in our experiments, the watch was inches away

from the mouth at the time of the extracted audio clips. Regardless of the food or activity type, each sample was exactly 0.25 s in length, and the peak of the wave amplitude was not necessary centered in the window. In some cases, such as during the biting of an apple, one quarter of a second was not sufficient to capture the entire bite. Therefore, the relevant information was partially truncated. Subjects were then asked to read a brief passage from a Wikipedia article, with no particular instruction about the rate at which they should read. The data was then automatically split into 0.25 s audio fragments using an audio processing program. Therefore, some samples were collected between phrases, and were relatively silent. Other fragments had periods of silence as well as vocalizations.

In order to evaluate if the classifier can distinguish between background noise and other classes of data, we added a separate "noise" class that we present in our classification results. However, the background noise used was relatively uniform: a 50 s clip of cafeteria noise consisting of movement, background chatter, and silverware noise. The clip was divided into 50 quarter-second samples. The environment was busy, and the noises were quite pronounced in comparison to the relatively quiet sounds associated with the other classes.

5.2. Smartwatch feedback: a survey

Before the system development phase, we had several important questions about how individuals feel about smartwatches. As described previously, a wearable device must have both high accuracy and high rates of user adherence for the subject to reach his or her intended goals. Furthermore, we proposed several questions about which hand a subject prefers to wear a watch. For example, our experimental evaluation requires that subjects wear a watch on the same hand with which they typically eat food such as chips or raise a glass of water. Though preliminary results suggest that data from individuals who pick up food with the hand on which they do not wear the watch can still be useful for classification, a thorough evaluation is left to a future work.

An online survey was conducted with a total of 221 responses in which various questions were posed with respect to how individuals feel about wearing a smartwatch. The participants in the study were anonymous, but represented a diverse set of ages, cultures, and genders. The study was originally conducted on January 28 for an internal data collection on smartwatch usage applied to the domain of medication adherence, but we found the majority of the questions were also applicable to food-intake monitoring as most questions pertained to smartwatch usage in general. This survey consisted of a total of 9 questions.

A separate online survey, with a total of 55 subjects, was later conducted to specifically investigate the attitude of individuals towards wearables in various form-factors. Specifically, subjects were asked to rate their receptiveness to smartwatch-based systems, custom wrist worn devices such as FitBit, necklace-based wearables, and smart glasses. Survey results and discussion can be found in Section 6.

6. Results and discussion

6.1. Audio classification

Results for classification between apples, chips, water, speaking, and ambient noise are shown in Table 3 based on 50 unprocessed samples collected from each of these foods, using the Random Forest classifier with 6555 extracted features from each sample. The Random Forest classifier consisted of 100 trees, each constructed using 13 random features, and was validated using leave-one-subject-out cross validation. Classifiers are generally evaluated on the basis of precision, recall, and *F*-measure. These terms are defined in Eq. (1) , where t_p is the number of true positives, and f_p is the number of false positives. The weighted average precision, recall, and *F*-Measure from our experimental results were 94.7%, 94.4%, and 94.4% respectively. In this case, the weighted average refers to the accuracy of the classifier across all different food groups, weighted according to the number of samples in each group. The majority of classification errors were between apples and potato chips. It should also be noted that while the ambient noise was disambiguated from the other classes in every sample, the ambient noise data was all recorded at the same location and therefore quite similar. Therefore, further work must be done to validate the ability of the proposed algorithm to recognize eating in real-world environments

Table 3
Audio: confusion matrix (random forest).

True class	Predicted class					
	Apple	Chips	Noise	Water	Talk	**Recall (%)**
Apple	40	9	0	1	0	80
Chips	3	47	0	0	0	94
Noise	0	0	50	0	0	100
Water	0	1	0	49	0	98
Talk	0	0	0	0	50	100
Precision	93.0%	82.5%	100%	98%	100%	

$$Precision = \frac{t_p}{t_p + f_p}$$
$$Recall = \frac{t_p}{t_p + f_n}$$
$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

Several other classifiers were also evaluated, many of which provided strong results using leave-one-subject-out cross-validation. The full comparison of classifiers is presented in Fig. 7. Though the RandomForest classifier produced the best results, the SimpleLogistic technique produced comparable results. The J48 decision tree classifier also performed well, with a precision, recall, and *F*-measure of 91.56%, 91.6%, and 91.5% respectively. The ROC curves for four classifier outcomes are shown in Fig. 6.

6.2. Feature extraction

From the 6555 extracted features, the Correlation Feature Selection (CFS) Subset Evaluator was used to evaluate 991,139 subsets of features. This is necessary to select the features best associated with the desired classifier outcomes. This subset evaluator considers both the individual predictive ability of features and the redundancy between them, and found the merit of the best subset to be 0.948. The search was stale after 5 node expansions. In other words, the subset evaluator aggregates the best features linearly beginning with those that show the highest correlation, and terminates after five consecutive subsets showing no improvement in classification accuracy.

The top 10 features are listed in Table 4. The first feature is the skewness of the logarithmic signal energy, in which skewness is defined as the asymmetry of the variable in comparison with a normal probability distribution [15]. More formally, skewness is defined in below, where μ_i is the ith central moment about the mean

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \quad (2)$$

For a probability density function $f(x)$, the first moment about the mean is always zero (with $s=1$), while the second moment is the variance. The third central moment is defined as skewness such that a distribution skewed to the right has a positive value, while one shifted towards the left has a negative skewness. The second most highly correlated feature is the mean peak distribution, which is defined as the mean distance between peaks for the logarithmic representation of the signal energy. The third feature is the number of non-zero values of the normalized log-energy signal.

Features 4–10, preceded by MFCC, are Mel-Frequency Cepstral Coefficients, which represent the spectral characteristics of the signal. A cepstrum is the result of the Inverse Fourier Transform of the logarithm of a signal spectrum. Mel-Frequency Cepstral Coefficients are based on the mel scale, which is a perpetual scale of pitches judged by listeners to be equidistant from one another [16]. The relationship between the frequency and mel scales is logarithmic, and can defined by the following formula (though other variations exist) [16]:

$$MEL(f) = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right) \quad (3)$$

However, the human ear can discern differences in frequency at low frequency ranges with a much higher resolution than at higher ranges, due to the physical properties of the cochlea. Therefore, a triangular Mel Filterbank is applied to the Discrete Fourier Transform of the original signal. Next, a dot product is computed between the filterbank and vector $P(k)$, which yields N intermediary coefficients – one for each triangle window function in the filterbank. Because humans do not perceive loudness on a linear scale, the logarithm is calculated for all N coefficients. Finally, the Discrete Cosine Transform (DCT) of the log powers is applied in order to decorrelate the energies of the overlapping filterbank energies. The resulting coefficients are used to extract statistical features as shown in Table 4.

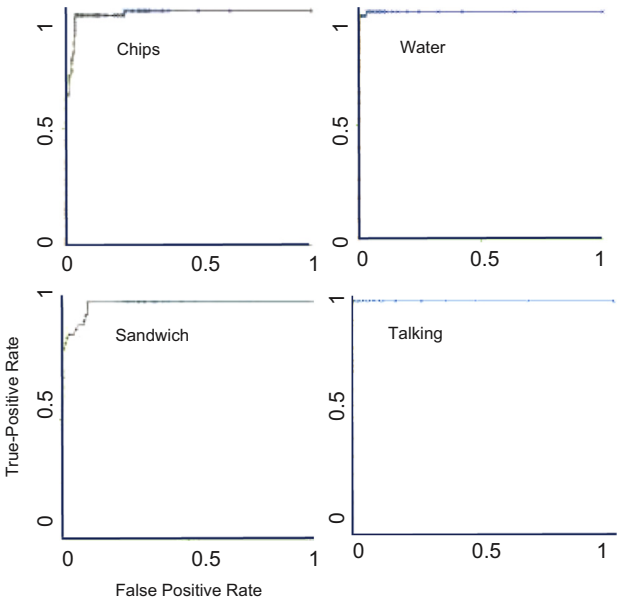

Fig. 6. The ROC curves for each classifier outcome, for the Random Forest classifier, show how the true-positive and false-positive rates vary based on the classifier threshold.

Table 4
Partial list of selected features.

Rank	Feature Name
1	Log Energy: Skewness
2	Log Energy: Mean Distance Between Peaks
3	Log Energy: Zero Crossings
4	Mel-Freq: Simple Moving Average[0] Quartile 3
5	Mel-Freq: Simple Moving Average[0] Mean Distance Between Peaks
6	Mel-Freq: Simple Moving Average[0] Zero Crossings
7	Mel-Freq: Simple Moving Average[1] Quartile 2
8	Mel-Freq: Simple Moving Average[1] Mean Distance Between Peaks
9	Mel-Freq: Simple Moving Average[1] Arithmetic Mean of Peaks
10	Mel-Freq: Simple Moving Average[1] Arithmetic Mean

6.3. Battery life implications

To realize the intended goal of minimizing burden, it is desirable for wearable devices to remain powered for weeks, or months, without interruption. Required nightly charging can be considered a burden to the user, which is undesirable because high user burden is typically associated with low compliance. Furthermore, energy-intensive applications can drain the battery completely, long before the user has an opportunity to recharge the device. Subsequently, the user will either uninstall the application or be unable to make proper use of it. Therefore, many activity monitoring devices have carefully factored power-efficiency into their design. Examples include that the Misfit Shine activity monitor claims a battery life of four months [17]. Other wearables

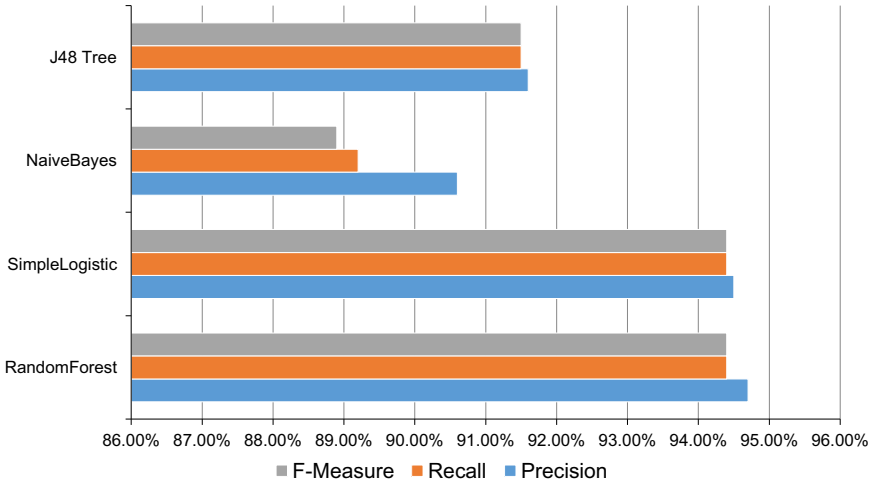

Fig. 7. Precision, recall, and *F*-measure are common measures of classification accuracy. This figure reports these values for different classifiers.

devices such as the Jawbone UP24 claim their devices can sustain seven days of continuous use [18].

Power-efficiency is a matter of particular concern in audio signal-processing applications such as the nutrition monitoring approach described in this paper. Audio signal processing typically requires that the signal be sampled at the Nyquist frequency, which is rather high compared to approaches that rely on inertial sensors such as accelerometers and gyroscopes. The Samsung Galaxy Gear has a 315 mAh capacity battery, which is significantly smaller than that of most mobile phones. In this section, we briefly describe our evaluations of the battery life implications of recording audio using the Samsung Galaxy Gear.

We evaluated the battery life of the smartwatch in three different use cases. In the first case, the screen of the phone was off and the watch was idle and unused. In the second case, the watch was idle but the screen was on. In the third case, the screen was on and the watch was recording audio at the same rate (96 kbps) as required for our audio analysis algorithms. Therefore, it can be inferred that the overhead of audio recording is the difference between the screen on and the screen on while recording. Our results are shown in Fig. 8. From these results, we can draw several conclusions. First, it is evident that the Smartwatch in a static mode with no computation and the screen off, consumes very little energy. Secondly, the overhead of recording audio is significant, and has a substantial effect on battery life. As the graph shows, an hour of recording audio with the screen on will consume 38% of the battery life, which amounts to approximately 119.7 mAh. Therefore, it can be assumed that the audio recording functionality consumes 10% of the watch's battery life per hour, as a rough approximation. Fig. 8 shows that the power dissipation of the screen is significantly larger than that of the audio recording and processing. Nevertheless, for long-term applications, the energy overhead of consistent audio recording may be prohibitive.

6.4. Smartwatch feedback: a survey

Fig. 9 provides several of the most pertinent questions from the survey. From the total sample of 221 respondents, 86% claimed to be right handed, 12% left-handed, and the remaining responded that they were ‘unsure’ or the question was ‘not applicable’.

In the following question, a total of 76% of respondents stated that they generally would wear a watch on their left hand, with an additional 19% who preferred to wear the watch on their right hand. The remaining 5% of those surveyed expressed no preference.

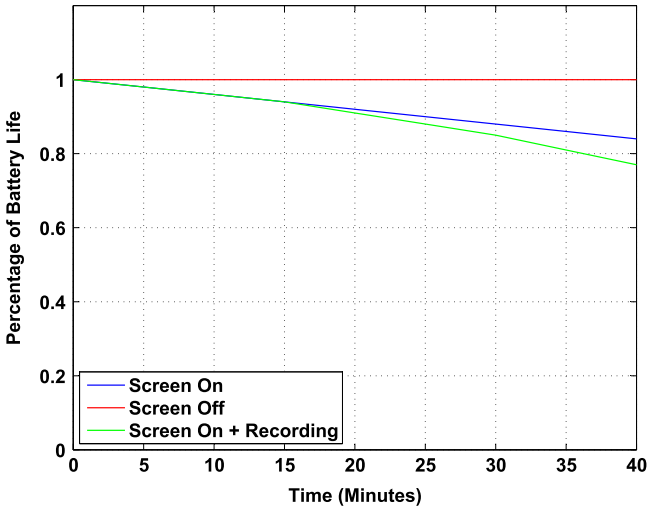

Fig. 8. A graph of battery life for three different use cases is shown above.

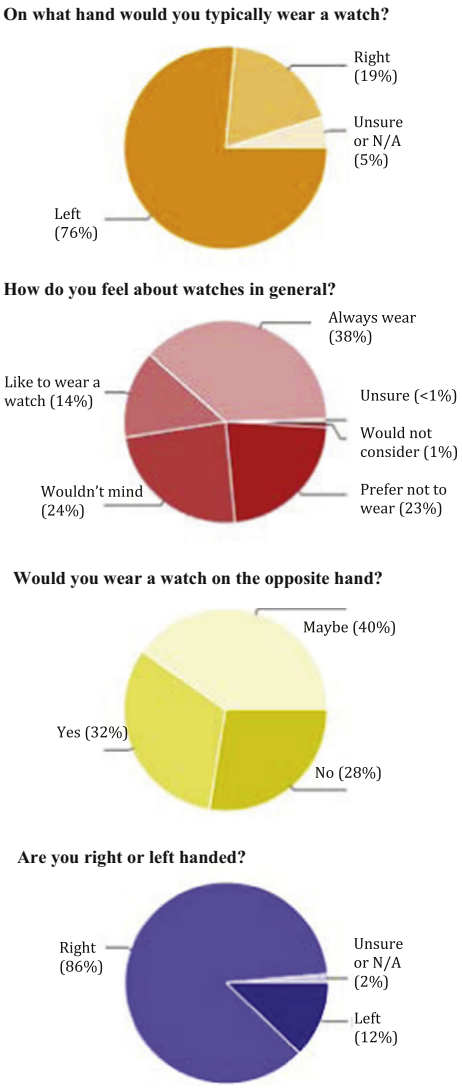

Fig. 9. Partial survey results are shown above.

The next question asked respondents how they felt about wearing watches in general. Most individuals stated that they always wear a watch (38%). However, 23% claimed that they preferred not to wear a watch, 24% stated that they would not mind, and 14% stated that they like to wear a watch. Only 1% of individuals claimed that they would not consider wearing a watch. The next survey question revealed that those who drank water out of a glass would use their primary hand to lift the cup for their mouth (69%), rather than the secondary hand on which the watch is worn (20%) with a remaining 10% claiming to be unsure. The final question asked respondents if they would be willing to wear a watch on the opposite hand to which they are accustomed. 40% of respondents answered ‘maybe’, 32% answering ‘yes’, and 28% answering ‘no’.

These results are generally promising: almost no individuals expressed an adamant refusal to wear a watch. Furthermore, results suggest that most subjects show some flexibility about which hand they wish to wear a watch. Consider foods that require both hands to be raised towards the mouth, such as large sandwiches or hamburgers. In such cases, the eating can be detected regardless of which hand the subject prefers to wear the watch upon. This is the case because during the initial bite, the

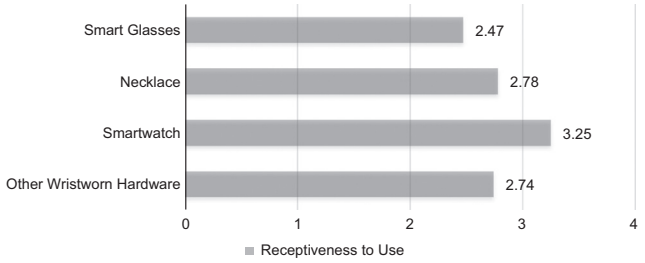

Fig. 10. Survey participants were asked for their receptiveness to wearable devices in several form factors.

watch will be close to the mouth and the microphone can detect the pertinent signals. However, failing to use the hand on which the watch is worn to raise a glass of water or eat potato chips may pose a challenge to detection, as the source is not as close to the microphone, and it is possible that the signal-to-noise ratio may be lower. The feasibility of detecting the ingestion of foods consumed with the secondary hand should be explored in future work.

It appears that enough individuals are willing to change which hand they wear their watch, to make detection of most eating habits possible if the algorithm settings are customized to their personal habits. However, generally speaking, the results suggest the importance of an adaptive algorithm that can be used to detect eating habits regardless of the hand on which the device is worn. This can potentially be achieved by detecting the distance between the watch and the audio source, and performing amplification and filtering accordingly. This will be explored in future works.

To evaluate the receptiveness of the public to the smartwatch platform, we conducted a separate survey to specifically investigate the attitude of individuals towards wearables in various form-factors. A total of 55 subjects participated in the online survey, of which 45.5% were male and 50.9% were female, and 3.6% who did not identify. 25.5% of subjects were 17 and younger, 27.3% ranged from 18 to 23, 27.3% were from 24 to 3, 12.7% were from 30 to 40, and 7.2% were over 40 years of age.

Subjects were asked to rate their willingness to wear health-monitoring wearable devices in four forms: glasses (such as Google Glass), smartwatches (such as the Galaxy Gear), custom wrist-worn hardware (such as FitBit), and necklaces (such as WearSens [19]). The scale ranged from 1, "not at all interested", to 5, "I would be completely comfortable wearing it". The results can be found in Fig. 10. As the data suggests, "Smart Glasses" was the least favorable option, with an average score of 2.47. The "Necklace" and "Other Wristworn Hardware" options scored similarly, at 2.78 and 2.74, respectively. The highest score was associated with the smartwatch, with a rating of 3.25 out of 5.

With respect to the number of individuals who assigned a rating of 5, the highest possible score, the smartwatch was also the favorite. 25.5% of individuals assign the smartwatch a rating of 5, compared to 9.1% with glasses, 10.9% for the necklace, and 10.9% for the "Other Wearables" option.

6.5. System limitations and future work

This paper provides an introduction to the applicability of the smartwatch platform for monitoring of eating habits, but there are several limitations in the proposed scheme that must be thoroughly evaluated in future works. These are described below:

1. The information from bites in close succession can be used to improve classification accuracy. For example, consider a subject who eats an entire bag of chips. The classification of each chip bite should not be conducted independently of the neighboring bites.

2. Smartwatch power optimization techniques should be developed, targeted towards selectivity in when audio recording is enabled based on hand gestures recognized using inertial sensors. Preliminary results show that the watch can only record audio for a few hours without optimizations, which is impractical. In addition, the sample rate of the audio signal could be substantially reduced based on the frequency ranges of the events of interest.

3. The range of foods tested should be expanded significantly, to evaluate the scalability of the proposed algorithm to real-world conditions.

7. Conclusion

This paper presents a novel approach to detecting ingestion of foods and liquids, using a smartwatch for identification of bites and swallows from acoustic signals. We also present a survey of users about smartwatch usage which confirms that a substantial portion of individuals would be willing to wear a watch on the hand with which they primarily eat. Future works will attempt to analyze eating behavior from the secondary hand, and explore the integration of audio-based detection of eating with inertial sensors for gesture recognition.

Conflict of interest statement

None declared.

Acknowledgements

This work is funded by the National Science Foundation AIR Option 1: Award no. 1312310.

References

[1] Centers for Disease Control and Prevention. Annual Medical Spending Attributable to Obesity: Payer-and Service-Specific Estimates, ⟨http://www.cdc.gov/obesity/data/adult.html⟩.
[2] P.S. Freedson, E. Melanson, J. Sirard, Calibration of the computer science and applications, inc. accelerometer, Med. Sci. Sports Exerc. 30 (5) (1998) 777–781.
[3] P.S. Freedson, K. Lyden, S. Kozey-Keadle, J. Staudenmayer, Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample, J. Appl. Physiol. 111 (6) (2011) 1804–1812.
[4] S. Patel, H. Park, P. Bonato, L. Chan, M. Rodgers, A review of wearable sensors and systems with application in rehabilitation, J. NeuroEng. Rehabil. 9 (1) (2012) 21.
[5] J. Fontana, J. Higgins, S. Schuckers, E. Sazonov, Energy intake estimation from counts of chews and swallows, Appetite 85 (2014) 14–21.
[6] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E.L. Melanson, M. Neuman, Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior, Physiol. Meas. 29 (5) (2008) 525.
[7] H. Tsujimura, H. Okazaki, M. Yamashita, H. Doi, M. Matsumura, Non-restrictive measurement of swallowing frequency using a throat microphone, IEEE Trans. Electron. Inf. Syst. 130 (2010) 376–382. http://dx.doi.org/10.1541/ieejeiss.130.376.
[8] S. Passler, W. Fischer, Acoustical method for objective food intake monitoring using a wearable sensor system, in: 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011, pp. 266–269.
[9] F.W. Pler S, M. Wolff, Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food, Physiol. Meas. (2012) 1073–1093.
[10] O. Amft, M. Kusserow, G. Troster, Bite weight prediction from acoustic recognition of chewing, IEEE Trans. Biomed. Eng. 56 (6) (2009) 1663–1672.
[11] O. Amft, A wearable earpad sensor for chewing monitoring, in: Sensors, 2010 IEEE, 2010, pp. 222–227, http://dx.doi.org/10.1109/ICSENS.2010.5690449.
[12] D. Brochetti, M. Penfield, S. Burchfield, Speech analysis techniques: a potential model for the study of mastication sounds, J. Text. Stud. 23 (2) (1992) 111–138. http://dx.doi.org/10.1111/j.1745-4603.1992.tb00515.x.

[13] W.E. HI, A.E. Deibel, C.T. Glembin, E. Munday, Analysis of food crushing sounds during mastication: frequency–time studies1, J. Text. Stud. 19 (1) (1988) 27–38. http://dx.doi.org/10.1111/j.1745-4603.1988.tb00922.x.
[14] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: The munich versatile and fast open-source audio feature extractor, in: Proceedings of the International Conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 1459–1462, http://dx.doi.org/10.1145/1873951.1874246, URL ⟨http://doi.acm.org/10.1145/1873951.1874246⟩.
[15] T. Pfister, P. Robinson, Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis, IEEE Trans. Affect. Comput. 2 (2) (2011) 66–78. http://dx.doi.org/10.1109/T-AFFC.2011.8.
[16] D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley, New York, NY, 1987.
[17] Misfit wearables faq, ⟨http://www.misfitwearables.com/supportl⟩.
[18] Jawbone Up Technical Specifications, ⟨http://jawbone.com/store/buy/up24⟩.
[19] H. Kalantarian, N. Alshurafa, T. Le, M. Sarrafzadeh, Monitoring eating habits using a piezoelectric sensor-based necklace, Else. Comput. Biol. Med. 58 (C) (2015) 46–55.
[20] opensmile faq, ⟨http://www.audeering.com/research/opensmile⟩.