A novel chewing detection system based on PPG, audio and accelerometry

Vasileios Papapanagiotou, Christos Diou, Lingchuan Zhou, Janet van den Boer, Monica Mars, Anastasios Delopoulos

***Abstract*—In the context of dietary management, accurate monitoring of eating habits is receiving increased attention. Wearable sensors, combined with the connectivity and processing of modern smart phones, can be used to robustly extract objective, and real-time measurements of human behaviour. In particular, for the task of chewing detection, several approaches based on an in-ear microphone can be found in the literature, while other types of sensors have also been reported, such as strain sensors. In this work, performed in the context of the SPLENDID project, we propose to combine an in-ear microphone with a photoplethysmography (PPG) sensor placed in the ear concha, in a new high accuracy and low sampling rate prototype chewing detection system. We propose a pipeline that initially processes each sensor signal separately, and then fuses both to perform the final detection. Features are extracted from each modality, and support vector machine (SVM) classifiers are used separately to perform snacking detection.**

Finally, we combine the SVM scores from both signals in a late-fusion scheme, which leads to increased eating detection accuracy. We evaluate the proposed eating monitoring system on a challenging, semi-free living dataset of 14 subjects, that includes more than 60 hours of audio and PPG signal recordings. Results show that fusing the audio and PPG signals significantly improves the effectiveness of eating event detection, achieving accuracy up to 0.938 and class-weighted accuracy up to 0.892.

I. Introduction

THE emergence of obesity and eating disorders as major health concerns has triggered intensive research efforts both for prevention and treatment of the disease. Monitoring of individual eating behaviour through self-reports, such as questionnaires, has proven to be highly unreliable since people tend to significantly underestimate their food intake (the reported energy intake is sometimes less than the minimum required to avoid starvation [1]). As a result, it is not possible to rely on such data for analysis, prevention or treatment purposes.

More recently, rapid advancements of technology in mobile computing, wearable sensors and computer networks have provided the tools to create reliable, objective and non-intrusive systems of monitoring dietary and nutrition habits. Crude approaches that use questionnaires in electronic/digital form (e.g. mobile phone-based logging systems) have given way to more sophisticated methods and systems that do not rely on manual user input, but on measuring specific physiological and behavioural parameters using wearables, and on real-time analysis of the collected data to infer useful and actionable information.

Previous approaches for monitoring eating occurrences have been based on audio recordings aiming to detect the distinct sound of food being crushed during each chew [2]–[4]. Various types of microphones have been used (such as open-air, bone-conduction, etc), usually placed inside the outer ear canal, where such chewing sounds are naturally amplified due to the ear physiology. Other approaches have opted for detecting swallowing sounds, as in [5], based on evidence that the frequency of swallowing occurrences can be used as a detector of snacking events or meals [6]. Alternatively, microphones have also been placed near the throat [7], aiming to detect swallowing sounds. In [8], audio is used to detect patterns in chewing and swallowing in order to detect the number of food items during a meal. Furthermore, other prototype sensors have been reported such as strain sensors [7], [9] that capture muscle activity (usually masseter and temporalis muscles), inertial sensors placed on the hand [10] or, more recently, proximity sensors placed on the head and hands of the subject to detect the hand movement that transfers the food from the plate to the mouth [11].

In this work, we focus on detecting chewing activity as a means to detect eating events (either meals or snacks). The chewing detection system is comprised of an open-air microphone, a photoplethysmography (PPG) sensor, and a data logger. The data logger is used to store the recorded signals, and is also equipped with a triaxial accelerometer in order to detect physical activity, since higher physical activity levels have been found to lead to false chewing detections. The data logger is attached to the subject's belt, while the chewing sensors are placed on an ear-hook and are positioned at the subject's ear (see Figure 1). We evaluate each of the microphone and PPG sensors separately, and also propose a late-fusion pipeline that combines signals from both sensors to increase the system's effectiveness. The work is performed in the context of the European funded SPLENDID project [12].

The rest of this paper is organised as follows. Section II presents related work. Section III presents the hardware of the proposed chewing detection system and Section IV the signal processing algorithms for each sensor component, as well as for their combination. In Section V we present the semi-free living experimental dataset and component-level evaluation of our system. Section VI discusses the evaluation results. Finally, Section VII concludes the paper.

V. Papapanagiotou, C. Diou and A. Delopoulos are with the Multimedia Understanding Group, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece. E-mail: vassilis@mug.ee.auth.gr, diou@mug.ee.auth.gr, adelo@eng.auth.gr

L. Zhou is with CSEM SA, Switzerland. E-mail: lingchuan.zhou@csem.ch

J. van den Boer and M. Mars are with Wageningen University, Netherlands. E-mail: janet.vandenboer@wur.nl, monica.mars@wur.nl

Manuscript received ; revised .

II. Related Work

In [13], a total of seven audio signal processing algorithms are evaluated on a dataset of 51 subjects. Audio is captured at 11 kHz by an in-ear microphone, while a second microphone placed outside the ear is used by one of the algorithms as reference [14]. Each subject consumes a total of 6 food types, in "laboratory" conditions (subjects eat specific foods, and are instructed to avoid talking or other sounds). Evaluation includes new algorithms, as well as algorithms known from the bibliography (such as [15] and [16]). The best results indicate chewing event detection precision and recall in the range of 70% to 80% in the paper's dataset.

In [17] the authors report the use an in-ear microphone sampled at 44.1 kHz both for automatic dietary monitoring and food type detection. Spectral analysis is performed on chewing sounds, and a subject-specific algorithm (that requires a training phase) is proposed. Experiments with a single subject achieve 52% precision and 93% recall for chewing event detection and two food types.

A multi-sensor system for chewing detection is proposed in [11]; it includes a strain sensor that detects jaw motion, a hand gesture sensor, and an accelerometer. At the first stage, an algorithm performs detection of food intake intervals by combining the sensor measurements. For these intervals, various features are extracted from each sensor individually, and Artificial Neural Networks are used to perform leave-one-subject-out (LOSO) validation. Authors report an average accuracy of almost 90%.

Swallowing sounds detection is performed in [18]. Authors use electromyography sensors and a microphone housed in a soft fabric worn around the subject's throat. Two detection algorithms are examined, one based on signal energy peaks (activations), and one on detecting a predetermined pattern. A dataset of 5 subjects is used to evaluate the method under strict laboratory conditions. The set of food types includes only water, yogurt and bread, however the data collection was repeated on two different days to account for physiologic variations. A total of 4.85 hours were recorded, out of which 27.2 minutes are swallowing sounds. Authors report accuracy of 73% to 75%.

In [19], a wearable jaw motion sensor is used to detect chewing activity. Authors use a feature selection method and perform classification using SVM with both linear and radial basis function (RBF) kernel. Evaluation is performed on a semi-free living dataset of 7 subjects; each subject participated for 3 days. Each day session lasted approximately 50 minutes, and the subject followed a routine of talking, walking, eating, and finally resting, each for 10 minutes. Authors report average accuracy of 90.5%.

Contribution: Our work contributes in this area of automated food intake monitoring by introducing a novel, low-rate chewing detection system based on PPG and audio sensors, which, combined with an accelerometer, can lead to accurate snacking event detection and analysis in real-life operating conditions. We significantly extend our previous work reported in [4] and [20] by introducing a complete, non-invasive detection system, proposing a complex signal processing pipeline with late fusion of SVM scores, and exploring the effect of accelerometer signals on the system's effectiveness. We support these claims through extensive experiments on a challenging dataset consisting of over 60 hours of semi-free living recordings with 14 subjects, 86 meals and snacks of various food types and durations.

(a) The chewing sensors (b) The data logger (c) PPG placement

Fig. 1: The prototype chewing sensors (Figure 1a); the data logger (Figure 1b); the PPG LED transmitter and receiver ear placement (Figure 1c).

III. Chewing Detection Hardware

The proposed chewing detection system hardware consists of an in-ear housing of a microphone and a PPG sensor (shown in Figure 1). The microphone is placed inside the outer ear canal, while the PPG transmitter and receiver are placed on both sides of the ear concha, as there is strong evidence that chewing activity significantly affects blood flow in that area [21]. The chewing sensors are connected via a wire to a data logger that samples and stores the audio and PPG signals (a version of the data logger which transmits the signals via Bluetooth to a mobile phone is currently under development). Similarly to [11], the data logger is also equipped with a triaxial accelerometer, that is used for detecting intervals of high physical activity (such as walking, running, etc). During experimentation we have observed that such intervals are frequently misclassified by our system as eating when in fact they are not. Interference of such activities in detection has also been reported in [9].

A. PPG sensor

Chewing mainly involves the use of the masseter, the temporalis, the medial pterygoid and the lateral pterygoid muscles. These are used to progressively process each bite, transforming it to a wet bolus that can be swallowed. Activation of these muscles affects blood flow in various points around them; one such point is the ear concha. These variations have long been detected and reported, in [21]. However, to the extent of our knowledge, no approach exists that relies on blood flow variations in order to detect chewing activity and food intake.

PPG is a method for optically obtaining volumetric measurements, and is widely used to measure perfusion via pulse oximetry [22]. It has been lately applied in applications such as heart-rate monitoring in wearables [23]. In our chewing sensors, a light-emitting-diode (LED) is used to illuminate the skin, from the outer side of the ear. A photo-diode placed

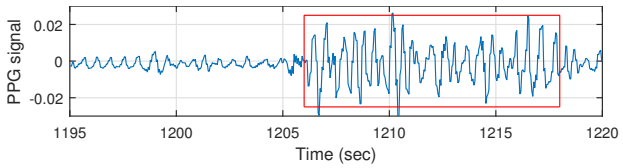


Fig. 2: Signal from the PPG chewing sensor after high-pass filtering. The red box indicates chewing activity. The signal pulses from 1195 to 1206 seconds correspond to heart-rate. The difference in amplification between the captured heart-rate and chewing components is essential to the PPG-based chewing detection.

inside the ear concha is used to measure the amount of light transmitted through the skin. Thus, the PPG can be used to capture the heart rate, by detecting periodicities in the range of 1 to 1.5 Hz (60 to 90 heart-bpms). However, during chewing, variations are created by the masseter and mainly the temporalis. These variations are produced as pressure is applied by the jaw to crush the food, and thus occur in synchronisation with each individual chew.

During the processing of each bite, a sequence of chews occurs, called a chewing bout. A bout starts with the first chew after biting or inserting the food in the mouth, and ends with the last chew before swallowing. During a bout, the individual chews appear with an approximately constant rate. This rate is in the band of 1 to 2 Hz (where in this case Hz is chews-per-second, Figure 2). Thus, detection of such intervals, where these frequencies are dominant, is the basis upon which we have designed the algorithm for processing the PPG signal.

Using a PPG sensor offers many advantages over a microphone and other sensors used in the literature. The in-ear PPG sensor is small and highly non-intrusive, compared to sensors housed in collars placed around the subject's throat. As it does not capture sound, it is not affected by ambient noises, talking, and other types of non-useful signal. It can also be combined with a microphone sensor, as we show in this work, in order to further increase the detection effectiveness and robustness. Finally, it has low power consumption and processing requirements, since it is sampled at a low frequency of 21.3 Hz.

However, the PPG signal is not entirely noise-free. Abrupt changes of environmental lighting can create significant artefacts, and can also lead to signal saturation (Figure 3). Such changes are usually caused by the subject moving to different places, for example walking out of a building into a sunny day or vice versa. To reduce the effect of such events, (a) adaptive amplification of the captured PPG signal is performed at hardware level, and (b) a pre-processing stage in the proposed signal processing algorithm (please refer to Section IV-A for details on these pre-processing steps). Furthermore, correct placement of the sensor is very important to achieve higher amplification for chewing signals compared to heart-beat (Figure 2) since both these signals significantly overlap in the frequency domain. This limitation is imposed by the current one-size-fits-all design of the system's hardware, and future work includes development of a better PPG sensor housing to overcome this problem.

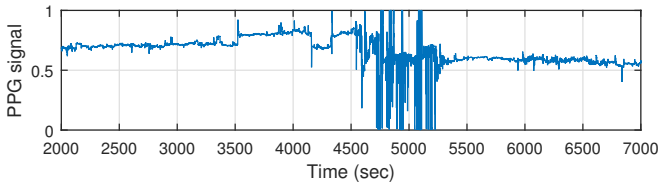


Fig. 3: Raw signal from the PPG chewing sensor. Abrupt changes are caused by the adaptive amplifier.

The PPG sensor used in our work consists of photo-diode BPW34FS and LED SFH4247, both manufactured by Osram. The photo-diode is especially designed to operate in the wavelength range of 870 to $1100\,nm$, while the LED emits at $940\,nm$; light is sampled at $64/3$ Hz (approximately 21.3 Hz). The PPG signal is adaptively amplified at the hardware level of the data logger. This requires that ambient light is also measured, by temporarily switching off the LED and measuring light intensity at the photo-diode. Ambient light is then subtracted from values measured while the LED is switched on (and emitting light). Based on these ambient-light-free values, the amplification level changes to ensure that the PPG signal is neither insufficiently amplified nor saturated. The data logger provides control signals that indicate the amplification level, which are used by the pre-processing stage of the PPG algorithm (Section IV-A).

B. Microphone sensor

The chewing detection system is also equipped with a microphone. We use an omni-directional model (FG-23329-D65) manufactured by Knowles that exhibits a sensitivity of $-53\,dB$ around the 1 kHz band. As shown in Figure 1, it is housed in an off-the-shelf ear bud, commonly used by mobile phone earphones; as a result, the sensor is placed inside the outer ear canal. This setup allows capturing of body-generated sounds, such as the crushing sounds of chews as well as talking, at higher level compared to external sounds, such as ambient noise, other people talking, etc.

In the dataset used in our experiments, audio was originally recorded at 48 kHz. Then, each recording was processed by a low-pass anti-aliasing filter and was down-sampled at 2 kHz. The down-sampling was required since the data logger employed in was built for experimental evaluation of various design parameters, including audio sampling frequency. Feature models will directly sample audio at 2 kHz. The entire processing pipeline has been implemented in MATLAB, using the libSVM library for SVM training and prediction.

C. Data logger

The chewing detection system is connected via a wire to a data logger, worn at the subject's waist. The data logger is used to sample and store the signals from the chewing detection system in a memory card. It is also equipped with a triaxial accelerometer (LIS3DH by STMicroelectronics), sampled in the same frequency of 21.3 Hz, as is the PPG sensor. The

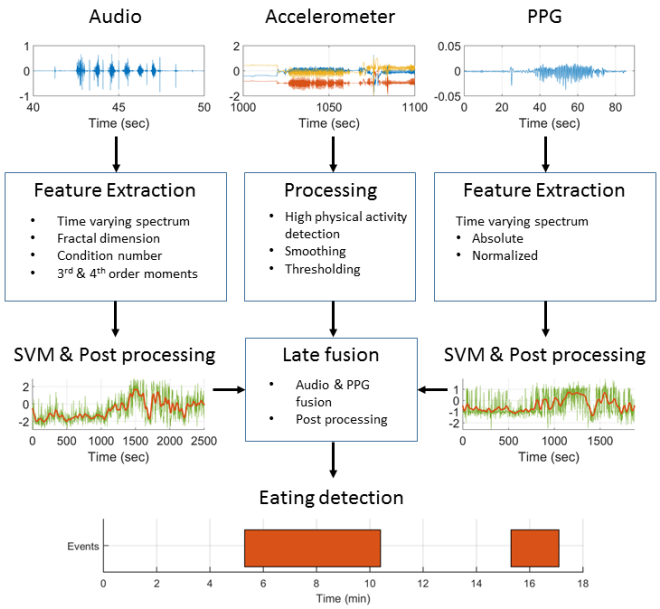


Fig. 4: The proposed eating detection pipeline. Features are extracted from windows of the audio and PPG signals and a score is computed for each window of each modality using SVMs. The SVM outputs are combined in a late-fusion approach in order to detect eating events. The effectiveness can further be improved by introducing physical activity information in the last stage of the detection.

accelerometer operates with a very low supply current of 6 to $11\,\mu A$ and very high sensitivity (1 to $12\,mg/digit$).

The signal from the accelerometer is used to provide information regarding the subject's physical activity level in order to assist our algorithms into discriminating between chewing and physical activity which occurs at similar frequencies as chewing, such as walking, running, etc.

IV. Chewing Detection Algorithms

In this Section we present the proposed algorithm for detecting eating events. We first present an algorithm for detecting eating events based solely on PPG, then another algorithm for audio, and finally the fusion of the two cues. The signal processing pipeline is shown in Figure 4.

All chewing detection algorithms are based on the following assumptions. Eating produces sequences of chews. Each chew lasts approximately 0.1 to 0.8 seconds, and subsequent chews are usually close to each other. Chews can be grouped into chewing bouts; a bout starts the moment food is placed into the subject's mouth and ends at swallowing. Each bout can last several seconds. Finally, bouts can be grouped into eating events; we use the term eating event to denote any complete session of eating activity. For example, eating a banana as a snack is a eating event; a full dinner (that can include first and main dishes and desert) is also an eating event. In the following, chews, bouts as well as eating events are represented by time intervals; they require a start and an end timestamp to be defined.

A. PPG-based chewing detection

The algorithm that processes the PPG signal aims at detecting intervals in the signal where the energy in the chewing frequency band is high. Initially, a pre-processing step is performed where the signal is smoothed using an high pass FIR filter with a cut-off frequency of 0.5 Hz. Subsequently, the control signals are used to smooth the filtered signal. More specifically, the time moments that amplification changes can be obtained using the derivative of the control signals. Amplification changes cause sudden offsets in the signal; this offsets survive the FIR filter as spikes. For this reason, a 5-second interval centred around each amplification change is replaced with 0-values on the filtered signal, essentially removing these spikes.

Next, we compute the time varying spectrum (TVS) of the signal based on Welch's method. This is achieved by first computing the discrete Fourier transform (DFT) over a sliding window of $N = 128$ samples (6 seconds) length and 1 sample step; the mean of each window is subtracted from its samples' values and a hamming window is applied before computing the DFT. The sliding window of 6 seconds is long enough to capture the effect of high activation due to chewing. We then compute the ensemble average of the DFT coefficients over several overlapping windows for each frame. In particular, let $X_n[k]$ denote the DFT coefficients for $k = 1, 2, \ldots, N$ for the n-th window; note that since the signal is real-valued, we need only compute $X_n[k]$ for $k = 1, 2, \ldots, \lfloor \frac{N}{2} \rfloor + 1$. We select a number of $2q + 1$ windows to perform spectrum estimation using

$$S_n[k] = \frac{1}{2q+1} \sum_{i=-q}^{q} \|X_{n+i}[k]\|^2 \quad (1)$$

In our experiments, we have picked $q = 12$. Thus, TVS is estimated on a longer sliding window of 152 samples with 1 sample step.

We then compute the energy $u_n[i]$ in 5 log-scale frequency bands ($i = 1, \ldots, 5$); the second and third bands are centred around the chewing band (as per [20]). More specifically, the 65 coefficients of TVS are grouped in 5 bands, corresponding to the following analog frequency ranges (in Hz): $0.0 - 1.0$, $1.0 - 1.8$, $1.8 - 3.3$, $3.3 - 5.9$ and $5.9 - 10.7$. A 10-dimensional feature vector is constructed (see Table I); the first 5 features are the energy values $u_n[i]$ of each band, whereas the latter 5 are the energy histogram $u'_n[i]$ obtained by normalising as $u'_n[i] = u_n[i] / \sum_{j=1}^{5} u_n[j]$. The histogram features may seem linearly dependant on the energy values; however, the scaling parameter $(\sum_{j=1}^{5} u_n[j])^{-1}$ is different for each feature vector, and thus the histogram provides additional information.

An SVM classifier with RBF kernel is used to perform detection; parameters C of SVM and γ of the RBF kernel are computed using a grid search (details are provided in Section V-C). The SVM score is computed as

$$s'_{\text{ppg}}[n] = \mathbf{w} \cdot \mathbf{f}[n] + b \quad (2)$$

where $\mathbf{w}$ is the separating hyperplane normal vector, $\mathbf{f}[n]$ is the feature vector corresponding to the n-th window, and b is the offset. To account for the slow transitions from eating to non-eating behaviour (and vice versa), we apply a relatively long

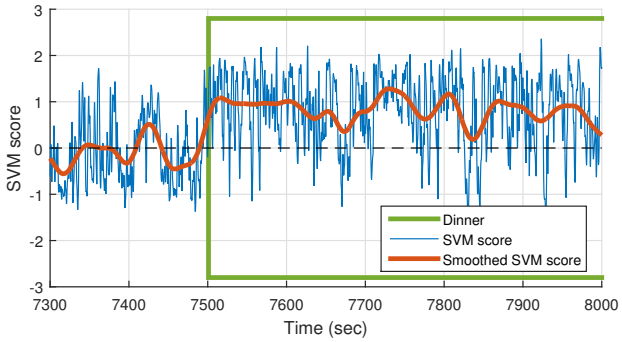


Fig. 5: Example of smoothing SVM scores. Dinner starts at 7501 seconds according to ground truth. Temporal average of SVM scores clearly increases when eating activity starts; this is captured very clearly by the smoothed SVM scores.

TABLE I: PPG and audio features.

	Feature	Dimension	Window
	PPG features		
1	Energy of log-band	5	6 sec
2	Energy histogram	5	6 sec
	Audio features		
1	Energy of log-band	9	0.2 sec
2	Fractal Dimension	1	0.1 sec
3	Condition number	1	0.1 sec
4	Skewness $m_3(0,0)$	1	0.1 sec
5	Kurtosis $m_4(0,0,0)$	1	0.1 sec
6	Moment $m_4(0,1,1)$	1	0.1 sec
7	Moment $m_4(0,2,2)$	1	0.1 sec

(1-minute) smoothing hamming filter on the decision scores s'_{ppg}, obtaining the PPG score signal $s_{\text{ppg}}[n]$; this is essential to eliminate the high variation of the decision scores. These variations can occur in-between chews, where the classifier can temporarily yield lower scores. Figure 5 shows an example of original SVM scores and their smoothed version.

Thresholding s_{ppg} yields intervals with eating activity that correspond to chewing bouts. Specifically, given a threshold A_{ppg}, we can directly identify detected chewing bouts as time intervals $b_i = [t_s[i], t_e[i]], i = 1, \ldots, N_{\text{b}}$, where

$$s_{\text{ppg}}[n] > A_{\text{ppg}} \tag{3}$$

only for those n such that $t_s[i] < n < t_e[i]$. The effect of various threshold values allows a trade-off between precision and recall; this is presented in Section V-C.

We then compute eating events $e_i, i = 1, 2, \ldots, N_e$ based on the detected bouts using a two stage process. At first, successive bouts are merged if they are no more than T_{gap} seconds apart (we have picked $T_{\text{gap}} = 60$ seconds); in particular, given two consecutive bout intervals $b_i = [t_s[i], t_e[i]]$ and $b_{i+1} = [t_s[i+1], t_e[i+1]]$, if

$$t_e[i+1] - t_s[i] < T_{\text{gap}} \tag{4}$$

holds, b_i and b_{i+1} are discarded and replaced with the new interval $[t_s[i], t_e[i+1]]$; note that the duration of this new interval is greater than the sum of the durations of the two discarded ones. Completing all merges yields a set of intervals which are possible eating events e_i.

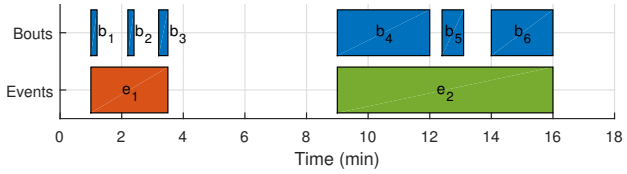


Fig. 6: Two examples of computing events from bouts. Bouts b_1, b_2 and b_3 are merged into interval e_1 and bouts b_4, b_5 and b_6 into e_2. However, interval e_1 is discarded as more than 75% of its duration is not covered by bouts. In contrast, interval e_2 is retained, and is the one and only detected eating event of the example.

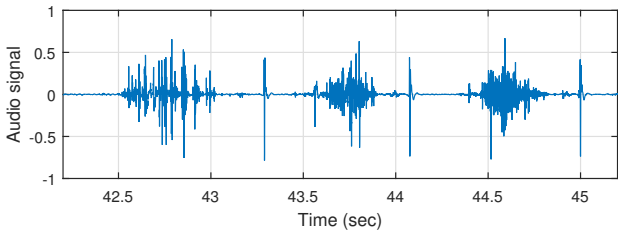


Fig. 7: Audio signal after pre-processing; three distinct chews are shown.

The merging process can however create intervals e_i that are long enough, yet they have been created by merging very short (and probably erroneously-detected) bouts. The second stage aims at removing such intervals. In particular, given an eating event e_i that has been produced by a series of consecutive bouts, we require that at least 25% of the duration of e_i is occupied by the bouts. Figure 6 demontrates this two-stage process.

B. Audio-Based chewing detection

This Section describes the audio signal processing pipeline. As a pre-processing step, we apply a high-pass FIR filter with a cut-off frequency of 20 Hz in order to remove the low-frequency components of the signal, since it provides no useful information. We have observed that this also removes some noises that register low frequency content on the microphone, such as the blowing wind, vehicle sounds, etc. An example of such a pre-processed signal with 3 chews is shown in Figure 7. A feature vector is then computed that consists of (a) signal energy in 9 log-scale frequency bands based on TVS estimation, (b) fractal dimension (FD), (c) condition number (CN) of the auto-correlation matrix, and (d) higher order statistics (3-rd and 4-th order moments); Table I lists all the features extracted from the audio signal. In the following, we provide details and rationale for each of these features.

The duration of the sliding window is 0.2 seconds for the spectral features and 0.1 seconds for non-spectral ones; the selection of short windows allows the detection of very brief chews. Each window is pre-processed by subtracting its mean, applying a hamming window, and dividing it by its standard deviation (STD). This last step essentially eliminates amplification-level changes that occur between subjects (due to different fitting of the chewing detection system) as well

as during a single session recording (due to the subject transitioning to noisier environments or interfering with the sensor).

Audio TVS estimation is also performed; however, it is important to note that audio TVS estimation captures different frequency information from PPG. Indeed, while PPG TVS aims at capturing the 1 to 2 Hz frequency of chewing occurrences, audio TVS captures frequencies up to 1000 Hz in much shorter windows, which correspond to texture information of each individual chew (or sound). The 129 coefficients of TVS are grouped in 9 bands, corresponding to the following analog frequency ranges (in Hz): $0.0-4.0$, $4.0-7.4$, $7.4-15.8$, $15.8-31.6$, $31.6-63.0$, $63.0-125.9$, $125.9-251.2$, $251.2-501.2$, and $501.2-1000$.

As shown in our earlier work [4], chewing sounds are highly fractal, and can be easily discriminated from talking, and potentially ambient noise based on their FD. We have also shown that this property is preserved, even for severely down-sampled versions of the signal. The FD is computed using the algorithm of [24] (more details on the FD-based chewing detection can be found in [4]). This feature can discriminate talking from chewing sounds with very high accuracy; talking is also detected by CN. CN is defined as the ratio of the greatest to the smallest eigenvalue of the auto-correlation matrix R; we estimate a 6×6 auto-correlation matrix, where $R(i,j) = m_2(i-j)$. The auto-correlation matrix R is computed for each pre-processed window of audio y.

Finally, higher order statistics help differentiate chewing from other noise-like sounds (city-buzz coming in from a window or while walking on the street, sounds produced by a vacuum cleaner, etc), as they are known to be insensitive to white noise. We estimate 4 moments [25]: (a) skewness

$$m_3(0,0) = \sum_{i=0}^{l-1} (y[i] - \mu_y)^3 \tag{5}$$

and (b) 4-th order moments with lags $(l_1, l_2, l_3) = (0,0,0)$ (or kurtosis), $(l_1, l_2, l_3) = (0,1,1)$, and $(l_1, l_2, l_3) = (0,2,2)$

$$m_4(l_1, l_2, l_3) = \sum_{i=0}^{N'_y} \left((y[i] - \mu_y) \prod_{j=1}^{3} (y[i + l_j] - \mu_y) \right) \tag{6}$$

where y is a pre-processed window of audio of N_y samples, μ_y is the mean of y (and is equal to 0 due to the window pre-processing), and $N'_y = N_y - 1 - \max\{l_1, l_2, l_3\}$. We must emphasise that these moments are in fact normalised by the power of the signal window, thanks to the normalisation in the window pre-processing step.

To account for the high variance of the spectrum and moment estimators we apply a smoothing filter on each feature (against time) using a 3.75-second hamming window. An SVM classifier with RBF kernel is then used with the smoothed feature vector; parameters C and γ are computed with a grid search (more in Section V-C). The SVM score $s'_{\text{audio}}[n]$ is computed as per Equation 2. The same long smoothing filter that is used to obtain s_{ppg} from s'_{ppg} is also applied on the audio SVM scores, yielding $s_{\text{audio}}[n]$.

A threshold A_{audio} (see Section V-C) is then used to identify individual chews. Chews are first merged into bouts (when they are closer than 2 seconds, and a minimum duration of 5 seconds is also required for each bout), and bouts are then merged into eating events, in exactly the same way as of PPG detection (Section IV-A).

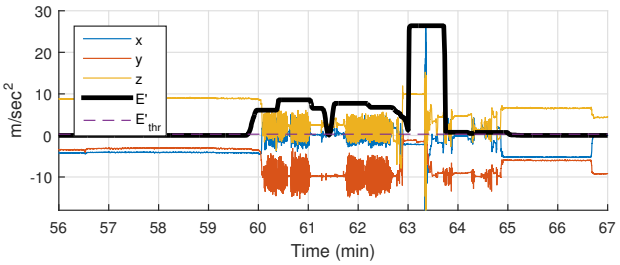


Fig. 8: An example of thresholding the acceleration-based signal E'. From approximately minute 60 to 64, high physical activity is detected which probably indicates no eating activity.

C. Physical activity thresholding

One of the biggest challenges of dietary monitoring in real-life conditions is interference from physical activity, such as walking or running. Walking naturally occurs at a frequency of 1 to 2 Hz, which is the same as the chewing band that the PPG chewing sensor captures. To improve the effectiveness of our system, we use a triaxial accelerometer that is embedded in the data logger. Processing of the accelerometer signal involves first computing the total acceleration $a[n]$ from the axis measurements $a_x[n]$, $a_y[n]$, and $a_z[n]$ as $a[n] = \sqrt{a_x^2[n] + a_y^2[n] + a_z^2[n]}$. A high-pass FIR filter with cut-off frequency at approximately 1 Hz is then applied to remove the DC offset due to gravity [26]. Signal energy E_a is then estimated at a rate of 0.28 seconds (6 samples) using a 5 second window, centred around n, as $E_a[n] = \frac{1}{N_a} \sum_i (a[i] - \mu_a)^2$ where the summation is across all N_a samples of the current window, and μ_a is the mean of the current window samples. Finally, the energy signal is dilated [27] using a structure element of ones, of 6.5 seconds length, thus expanding the effect of physical activity in order to avoid confusion during transient moments.

The resulting dilated signal $E'_a[n]$ is then thresholded using an empirically set value E'_{thr}, so that only high activity levels exceed the threshold (the same threshold value was used throughout all experiments of this work). An example is shown in Figure 8. Intervals exceeding the threshold are interpreted as intervals with high physical activity, walking, running, etc, which are less likely to contain chewing activity. We thus discard any detected eating during these intervals, in order to avoid false detections. There exists however the case, where a subject can be walking while simultaneously eating; the algorithm is then bound to miss this eating event. In general however, the gain in effectiveness is higher when making use of E'_a (Section V-C).

D. Fusion - proposed pipeline

In order to increase effectiveness, we combine the microphone and the PPG sensor in a late-fusion scheme. We use the

smoothed decision scores $s_{\text{audio}}[n]$ and $s_{\text{ppg}}[n]$ of the the audio and PPG SVM classifiers respectively. Since these signals have different sampling frequencies, we down-sample s_{audio} (which has the highest rate) to the frequency of s_{ppg} using linear interpolation.

The final decision is carried out using the following equation

$$s_{\text{ppg}}[n] + \alpha \cdot s'_{\text{audio}}[n] > A_{\text{fusion}} \quad (7)$$

where s'_{audio} denotes the down-sampled version of s_{audio}. Parameter α defines the factor by which the microphone contributes to the final decision, and parameter A_{fusion} defines the strictness of the classification (see Section V-C).

V. Evaluation

A. Dataset

The proposed system is evaluated on a semi-controlled dataset collected using the prototype sensors. The data collection took place in the Wageningen University, during the summer of 2015. A total of 22 subjects used the system; 19 females and 3 males with a mean age of 22.9 years and a mean body mass index of 28.0 km/m^2. Out of the 22 subjects, 19 participated in the data collection for 2 days, which were two weeks apart; the remaining 3 participated only for one day. Each of the 41 day sessions contains recordings of approximately 5 hours, split over two or three data-files, depending on the overall length. Due to hardware failures (missing recordings, recordings included only digital noise or extremely saturated signals) we have collected a total of 26 such data-files from 14 subjects. The total duration of the 26 data-files is approximately 60 hours (per sensor), out of which 7.6 hours correspond to eating activity.

The recordings started and concluded at the university premises where the subjects were monitored; eating and physical activity occurrences for each subject were recorded in a diary. Subjects were instructed to wear the data logger and chewing detection system throughout the recording, and were informed that they could abort the study at any point without consequences. Prior to the recording the subjects were familiarised with the sensor. The recording protocol began with some free-time, and then each subject ate lunch; a selection of food types were available for consumption (see Table II). After lunch, the subjects were able to spend the rest of the afternoon freely, and were allowed to leave the university premises. They were instructed to eat 3 snacks of their choice (e.g. an apple or candy bar; the full list of consumed snacks is shown in Table III) and perform at least 4 high physical activity tasks of their choice during that time. Subjects were also free to drink anything, however we have not marked drinking events for detection.

Activities included lying inside, sitting inside or outside, walking inside including stairs and elevators, walking outside, washing dishes, vacuuming, playing ball, and cycling inside and outside. At the end of the afternoon, subjects assembled to the dining room for dinner; available food types for dinner are shown in Table IV. In the end, subjects ate two main meals during each day, and 2 to 6 snacks.

TABLE II: Food types consumed during lunches

Type	Day 1	Day 2
Bread	Sliced bread, crackers	Soft buns, baguette, rusk
Topping	Butter, jam, chocolate sprinkles, chocolate spread, peanut butter, cheese, sliced meat	Butter, jam, chocolate sprinkles, chocolate spread, peanut butter, cheese, sliced meat
Fruit	Grapes, banana, apple	Grapes, banana, apple
Drinks	Water, milk, orange juice	Water, milk, orange juice

TABLE III: Food types consumed during snacks

Type	Day 1	Day 2
Fruit	Grapes, banana, apple	Orange, strawberry, kiwi,
Cookie	Bastogne cookie, gingerbread, fruit biscuit	Hazelnut waffle, spongecake caramel waffle
Chips	-	Potato chips
Candy	Hard boiled candy, liquorice, twix bar, chewing gum	Lollipop, wine gums mars bar
Drinks	Coffee, tea, hot chocolate, water, lemonade, orange juice, coke	Coffee, tea, hot chocolate, water, lemonade, orange juice, coke, milk

Annotation was performed in a two-stage process. During the first stage, organisers of the data collection study monitored the subjects and created detailed diaries regarding their eating and physical activities. This was possible since only 3 subjects were using the system at any given day. In the second stage, we marked eating events (start and end timestamps) by listening to the entire audio files and simultaneously observing the waveform visually. The second stage annotations were then cross-checked with the diaries, to minimise the chance of no lost eating events.

B. Evaluation metrics

Three types of evaluation are used in the experiments: (a) leave-one-subject-out (LOSO) duration-based, (b) cumulative duration-based and (c) event-based.

1) LOSO duration-based evaluation: To evaluate an eating event detector based on duration, each data-file is manually partitioned into consecutive, non-overlapping intervals using the 'Audacity'[1] software; each interval is marked either as positive ('eating') or negative ('non-eating') based on ground truth. Drinks and chewing gum are not marked as positive. A second partitioning is derived in the same way, however it is based on the output of the evaluated detector instead of ground

[1]http://www.audacityteam.org/

TABLE IV: Food types consumed during dinners

Type	Day 1	Day 2
Potatoes	Boiled	Puree
Vegetables	French beans	Salad (lettuce, tomato, cucumber, boiled egg)
Meat	Meatball, wrapped in a slice of meat	Chicken schnitzel
Condiment	Gravy	Salad dressing
Dessert	Custard, vanilla & chocolate	Vanilla ice cream

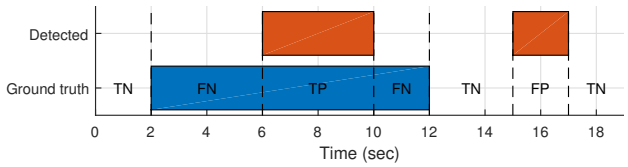


Fig. 9: Example of confusion matrix calculation for duration-based evaluation; for this 19-second duration recording we derive: $TP = 4$, $FP = 2$, $TN = 7$, $FN = 6$.

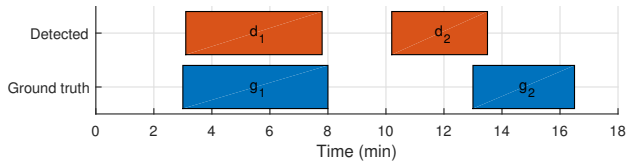


Fig. 10: Example of event-based evaluation. Ground truth event g_1 is matched with detected event d_1; g_2 is not matched to d_2 as they do not overlap more than the required threshold. Thus, $CD = 1$ (corresponding to pair g_1 and d_1), $FD = 1$ (corresponding to d_2), and $MD = 1$ (corresponding to g_2).

truth (the process for obtaining detector eating event intervals is described in Section IV-A). We calculate:

- True positive (TP) time as the total duration of the recording during which both the detector and ground truth indicate 'eating'.
- False positive (FP) time as the total duration of the recording during which the detector indicates 'eating' and ground truth indicates 'non-eating'.
- True negative (TN) time is calculated as the total duration of the recording during which both the detector and ground truth indicate 'non-eating'.
- False negative (FN) time is calculated as the total duration of the recording during which the detector indicates 'non-eating' and ground truth indicates 'eating'.

An example is illustrated in Figure 9. It is important to note that time is not quantised; the duration of each interval is used to perform the computations.

Since there exist multiple data-files for each subject, we calculate one confusion matrix per subject by summing all the confusion matrices from the data-files of that subject. Various metrics (e.g. precision, recall) are then calculated for each subject, as described in Section V-B4. Finally, we present the mean (across subjects) for each metric.

2) Cumulative duration-based evaluation: In this evaluation scheme, we compute the confusion matrices per data-file in exactly the same manner as in Section V-B1. However, all confusion matrices are then summed into one. This confusion matrix partitions into TP, FP, TN and FN the entire duration of the dataset (approximately 60 hours). The same metrics as in Section V-B1 are calculated.

This evaluation differs from 'LOSO duration-based evaluation' as it takes into account the duration of the recording for each subject (i.e. subjects with longer recordings affect the evaluation more).

3) Event-based evaluation: For event-based evaluation, we initially process each data-file individually, and derive the ground truth and evaluated detector partitionings as in Section V-B1. Each positive ground truth interval is regarded as a ground truth eating event, and each positive detected interval is regarded as a detected eating event. Note that we do not allow two consecutive intervals to be labelled with the same label (both positive, or both negative). We then perform a matching between detected and ground truth eating events that satisfies the following criteria:

- Each detected eating event is matched with either 0 or 1 ground truth eating event.
- Each ground truth eating event is matched with either 0 or 1 detected eating event.
- The overlap duration for each matched pair must be at least 75% of the duration of the union of the matched events.

Each matched pair contributes as one correct detection (CD). Each non-matched detector event contributes as one false detection (FD), and each non-matched ground truth event contributes as one missed detection (MD). An example is demonstrated in Figure 10.

As a result, total CD for the entire dataset are computed as the sum of all CD across each of the 26 recordings. Total FD and MD are also computed by summing the results across the recordings.

4) Evaluation metrics: Given a confusion matrix, we compute precision $= \frac{TP}{TP+FP}$, recall $= \frac{TP}{TP+FN}$, as well as

$$\text{F1 measure} = 2\frac{TP}{2 \cdot TP + FP + FN} \tag{8}$$

Weighted accuracy is also used as in [28], and computed as

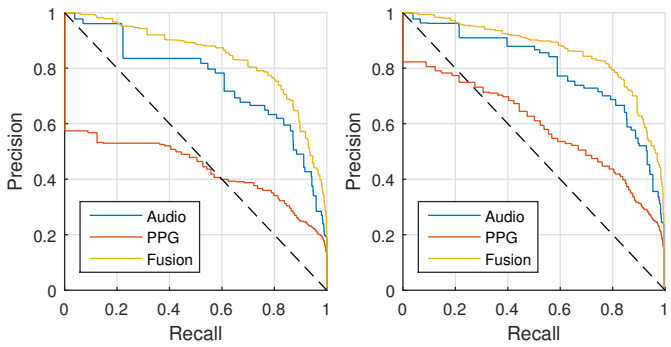
$$\text{weighted accuracy} = \frac{w \cdot TP + TN}{w(TP + FN) + FP + TN} \tag{9}$$

where w is the positive class weight; setting $w = 1$ yields accuracy (non-weighted). Authors of [28] use $w = 20$, based on the hypothesis that in real life, eating vs non-eating activity occurs at a $1:20$ ratio in a 24-hour cycle. In our results, we computed both non-weighted accuracy ($w = 1$, simply denoted accuracy), as well as weighted accuracy (with $w = 6.9$, based on our dataset's prior probability of eating $7.6 : (60 - 7.6)$).

C. Experiments

Since each of the three algorithms requires the training of an SVM model, we perform LOSO experiments to test their effectiveness on the dataset. In particular, for each subject we assemble all data-files of that subject as the test set. From the remaining recordings (that belong to other subjects) we randomly select an equal amount of positive and negative feature vectors to assemble the development set. In particular, we select 2000 positive and 2000 negative feature vectors for audio, and 1000 positive and 1000 negative feature vectors for PPG. The development set is then randomly split into a training set (which contains 70% of the development set) and a validation set (which contains the remaining 30%).

A grid search is then used to determine optimal values for parameters C and γ of the SVM and RBF kernel using the



(a) Without accelerometer (b) With accelerometer

Fig. 11: ROC curves for the threshold of the smoothed decision score of SVM and Equation 7. Evaluation is based on duration.

development set. In particular, we search in $C \in \{10^i, i = -2, -1, 0, 1, 2\}$ and $\gamma \in \{\gamma_0^i, i = -2, -1, 0, 1\}$, where $\gamma_0 = D^{-1}$ is the default libSVM value and D is the number of features ($D = 15$ for audio and $D = 10$ for PPG). For each point on the grid, a model is trained on the training set and is evaluated on the validation set. The pair that leads to the highest validation set accuracy is selected and a final model is trained using the entire development set. Evaluation is performed on the held-out test set (i.e. the "left-out" subject data). This process is repeated for all subjects.

Receiver operator characteristic (ROC) curves are calculated by varying the threshold values A_{ppg}, A_{audio} and A_{fusion} of the smoothed SVM scores s_{ppg} and s_{audio}, as well as the parameter α of Equation 7. For the thresholds we select 500 equally spaced values in the range of the decision scores, and for parameter α we select values from 0 to 2 with a step of 0.25. LOSO duration-based evaluation curves are shown in Figure 11. It is important to note that these thresholds and parameter values are always the same across all subjects (no inter-subject tuning).

In Table V we present evaluation results for each of the three methods of Section V-B; in particular, we present those points that yield the highest precision while maintaining recall over 0.8. Table Va shows results for duration-based evaluation; five metrics are shown for each combination of sensors, as well as the values of thresholds $A_{\text{ppg|}}$, A_{audio}, A_{fusion} (presented in the same column, A_{sensor}) and parameter α (for fusion only) for which these results are obtained. Similar results are shown in Table Vb for cumulative duration-based evaluation. Finally, we present in Table Vc event-based results.

As a comparison, we present the effectiveness of 3 audio-based algorithms presented in [13]: maximum sound energy algorithm (MSEA), maximum energy slope algorithm (MESA), and low-pass filtered signal algorithm (LPFSA) in Table VI. Since these algorithms detect chews, we apply a simple aggregation method first that computes chewing bouts by merging the detected chews that are closer than 2 seconds and then discards all bouts that are less than 5 seconds long. We then use the same aggregation method to compute eating events as in Section IV-A. All 3 algorithms require setting a parameter value; we have selected such as value that yields recall slightly higher than 0.8 for comparison with our algorithms.

TABLE V: Evaluation results for all three detection algorithms, with and without the use of the accelerometer signal ("+" denotes with acc.). Parameter selection is based on maximising precision while maintaining a minimum recall of 0.8. Bold indicates highest value for each metric and lowest for FD and MD.

(a) LOSO duration-based evaluation

sensor	**prec.**	**rec.**	**acc.**	**w. acc.**	**F1**	A_{sensor}	α
PPG	0.341	**0.814**	0.753	0.767	0.448	0.206	-
PPG+	0.436	0.805	0.814	0.800	0.522	0.206	-
Audio	0.633	0.809	0.880	0.861	0.650	0.220	-
Audio+	0.687	0.811	0.912	0.879	0.693	0.175	-
Fusion	0.760	0.802	0.928	0.886	0.729	0.654	1.25
Fusion+	**0.794**	0.807	**0.938**	**0.892**	**0.761**	0.509	1

(b) Cumulative duration-based evaluation

sensor	**prec.**	**rec.**	**acc.**	**w. acc.**	**F1**	A_{sensor}	α
PPG	0.278	0.801	0.710	0.749	0.413	0.106	-
PPG+	0.336	0.809	0.773	0.788	0.475	0.073	-
Audio	0.476	0.811	0.861	0.840	0.600	0.311	-
Audio+	0.561	**0.818**	0.895	0.862	0.666	0.220	-
Fusion	0.641	0.805	0.918	0.870	0.714	0.618	1
Fusion+	**0.702**	0.800	**0.931**	**0.875**	**0.748**	0.581	1

(c) Event-based evaluation; the high number of FDs is due to the requirement that at least 80% of all snacks ($CD + MD$) is correctly detected (CD).

sensor	**No. of CDs**	**No. of MDs**	**No. of FDs**	A_{sensor}	α
PPG	70	16	202	0.306	-
PPG+	69	17	153	0.239	-
Audio	**72**	14	89	0.356	-
Audio+	71	**15**	63	0.356	-
Fusion	69	17	51	0.618	0.75
Fusion+	69	17	**33**	0.618	0.75

TABLE VI: Cumulative duration-based evaluation for audio-based algorithms of Päßler et al. [13].

	prec.	**rec.**	**acc.**	**w. acc.**	**F1**
MSEA	0.288	0.804	0.720	0.756	0.424
MESA	0.304	0.813	0.738	0.770	0.443
LPFSA	0.289	0.811	0.720	0.759	0.426

VI. Discussion

Based on the results of the previous Section, PPG significantly under-performs compared to audio. Precision is 0.341 for LOSO and 0.278 for cumulative duration-based evaluation. These values are much lower compared to the ones reported in [20]. However this is expected as the dateset used in this work is much more challenging compared to the one of [20], where subjects were always seated on a table, alternating between talking, coughing, eating, drinking, and silence. Furthermore, PPG-based detection relies on the chewing-related signal being recorded with significantly higher energy compared to other signals of similar frequency content, such as heart rate; this in turn requires that the photo-diode is directly facing the LED. In our dataset this was not always the case, either because the subject did not properly position the sensor at his/her ear, or

due to inappropriate sensor size for the subject's ear anatomy (only one sensor size was used in the experiments). A solution to this problem would be to explore alternative housings for the PPG sensor; however, in this work we wish to combine the PPG sensor with an in-ear microphone and we therefore chose to embed the sensor in an ear hook.

For audio-based detection, on the other hand, precision is almost double for LOSO duration-based evaluation, 0.633 compared to 0.341 of PPG. Precision of our algorithm for cumulative evaluation is 0.476 which is also higher than the precision of the MSEA, MESA, and LPFSA algorithms of [13]. As presented in Table VI, these 3 algorithms achieve precision of approximately 0.29; this is a clear indicator of the challenge of our dataset.

Even though audio-based detection clearly outperforms PPG-based detection, fusion of both detection signals further increases the detection accuracy. In particular, LOSO duration-based evaluation for fusion yields precision of 0.76; this is more than 12 percentage points higher than audio, or 20% higher. Student's t-test for the null-hypothesis (that fusion does not improve over audio) yields $p < 0.02$ for all metrics against both PPG and audio based detections. In Table VII we present p-values for all metrics (given the 0.8 recall requirement) for 9 cases. Furthermore, the values of parameter α for which this effectiveness is achieved is either 1 or close to 1 (1.25 or 0.75) indicating a balanced contribution of both modalities in the detection.

Introducing the accelerometer signal to each detection method improves the results; precision increases 9, 5 and 3 percentage points for PPG, audio, and fusion-based detection respectively. Student's t-test about the improvement introduced by the accelerometer is by chance yields $p = 5 \cdot 10^{-7}$ for precision and $p = 10^{-5}$ for F1 score regarding PPG, and for the null-hypothesis that Audio+ does not improve over Audio $p = 0.02$ and $p = 0.1$ for the same metrics. Introducing the accelerometer for the fusion-based detection does not improve results significantly, since for most subjects effectiveness remains approximately the same. However, it increases for some few subjects, and which explains the higher values for Fusion+ in Table V. Furthermore, given the requirement for at least 80% CD, all 6 detection methods achieve approximately 70 CD and 16 MD (see Table Vc); however, fusion-based detection (and especially fusion-based combined with accelerometer) greatly decreases the number of FD, yielding only 33 FD for the best case.

In addition, as described in Section IV-B, a sampling frequency of as low as 2 kHz is sufficient to detect chewing activity (also shown in our earlier work in [4]). This allows us to overcome obstacles in real-world deployment of the chewing detection system as a wearable system; an audio signal of 2 kHz can be easily transmitted via Bluetooth to a mobile phone, and the required calculations can be easily carried out on modern mobile phones without significantly affecting battery consumption or CPU load. Tests with a preliminary prototype implementation of the entire streaming pipeline on Android devices (manufactured by Samsung, LG, Motorolla and HTC) do not indicate a significant impact on battery consumption or CPU load.

TABLE VII: Student's t-tests (p-values) for null-hypothesis that algorithm A produces the same results as algorithm B when comparing "A vs. B". Bold values indicate the cases where the null-hypothesis was not rejected based on the 0.05 probability threshold.

	p-value			
t-test	**prec.**	**acc.**	**w. acc.**	**F1**
PPG+ vs. PPG	0.0000	0.0004	0.0001	0.0000
Audio+ vs. Audio	0.0214	**0.1083**	**0.2729**	0.0132
Fusion+ vs. Fusion	**0.4058**	**0.8600**	**0.4276**	**0.4494**
Audio vs. PPG	0.0130	0.0044	0.0015	0.0018
Fusion vs. PPG	0.0000	0.0015	0.0000	0.0000
Fusion vs. Audio	0.0047	0.0147	0.0008	0.0013
Audio+ vs. PPG+	0.0391	0.0217	0.0394	0.0092
Fusion+ vs. PPG+	0.0007	0.0093	0.0016	0.0002
Fusion+ vs. Audio+	0.0107	**0.1778**	0.0493	0.0261

VII. Conclusions

We have presented a novel chewing detection system relying on audio, PPG and accelerometry to identify eating events. The proposed system integrates a microphone and a PPG sensor in an ear hook, connected via wire to a datalogger equipped with an accelerometer. Validation on an experimental, yet challenging real-life-like dataset, shows that the combination of signals from all sensors yields better results compared to results from either audio or PPG alone. For duration-based evaluation, accuracy reached 0.938 (weighted accuracy is 0.892), while for evaluation based on eating events, it reached precision 0.794 and recall 0.807 (with F1 score of 0.761). These results are particularly encouraging, given the challenging nature of the evaluation dataset.

The system's effectiveness, low sampling rate and low computational requirements indicate potential for use in dietary monitoring applications. However there is still a lot of room for improvement, both in research and in technical work. Real-time integration with mobile devices is a necessary first step, and we are currently working on integrating a Bluetooth transmitter to the data logger. Furthermore, user comfort is an important factor affecting the usability of the system and future study includes the exploration of alternative sensor housing designs. Finally, there are also limitations that should be taken into account; the integrated audio and PPG system does not detect individual chews, but rather chewing bouts. In addition, it does not detect drinking and future work includes studying the potential of in-ear microphone sounds for drinking detection.

Acknowledgement The work leading to these results has received funding from the European Community's ICT Programme under Grant Agreement No. 610746, 01/10/2013 - 30/09/2016, and is performed in the context of the SPLENDID project "Personalised Guide for Eating and Activity Behaviour for the Prevention of Obesity and Eating Disorders" (http://splendid-program.eu/).

References

[1] E. Archer, G. A. Hand, and S. N. Blair, "Validity of us nutritional surveillance: National health and nutrition examination survey caloric energy intake data, 1971–2010," *PloS one*, vol. 8, no. 10, p. e76632, 2013.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *UbiComp 2005: Ubiquitous Computing*. Springer, 2005, pp. 56–72.

[3] S. Päßler and W.-J. Fischer, "Acoustical method for objective food intake monitoring using a wearable sensor system," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. IEEE, 2011, pp. 266–269.

[4] V. Papapanagiotou, C. Diou, Z. Lingchuan, J. van den Boer, M. Mars, and A. Delopoulos, "Fractal nature of chewing sounds," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9281, pp. 401–408. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23222-5_49

[5] O. Makeyev, P. Lopez-Meyer, S. Schuckers, W. Besio, and E. Sazonov, "Automatic food intake detection based on swallowing sounds," *Biomedical signal processing and control*, vol. 7, no. 6, pp. 649–656, 2012.

[6] E. S. Sazonov, S. A. Schuckers, P. Lopez-Meyer, O. Makeyev, E. L. Melanson, M. R. Neuman, and J. O. Hill, "Toward objective monitoring of ingestive behavior in free-living population," *Obesity*, vol. 17, no. 10, pp. 1971–1975, 2009.

[7] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman, "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior," *Physiological Measurement*, vol. 29, no. 5, p. 525, 2008.

[8] P. Lopez-Meyer, S. Schuckers, O. Makeyev, J. M. Fontana, and E. Sazonov, "Automatic identification of the number of food items in a meal using clustering techniques based on the monitoring of swallowing and chewing," *Biomedical signal processing and control*, vol. 7, no. 5, pp. 474–480, 2012.

[9] E. S. Sazonov and J. M. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing," *Sensors Journal, IEEE*, vol. 12, no. 5, pp. 1340–1348, 2012.

[10] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008.

[11] J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior," *Biomedical Engineering, IEEE Transactions on*, vol. 61, no. 6, pp. 1772–1779, 2014.

[12] C. Maramis, C. Diou, I. Ioakeimidis, I. Lekka, G. Dudnik, M. Mars, N. Maglaveras, C. Bergh, and A. Delopoulos, "Preventing obesity and eating disorders through behavioural modifications: the splendid vision," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*. IEEE, 2014, pp. 7–10.

[13] S. Päßler and W.-J. Fischer, "Evaluation of algorithms for chew event detection," in *Proceedings of the 7th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012, pp. 20–26.

[14] ——, "Food intake activity detection using a wearable microphone system," in *Intelligent Environments (IE), 2011 7th International Conference on*. IEEE, 2011, pp. 298–301.

[15] J. Nishimura and T. Kuroda, "Eating habits monitoring using wireless wearable in-ear microphone," in *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*. IEEE, 2008, pp. 130–132.

[16] M. Shuzo, S. Komori, T. Takashima, G. Lopez, S. Tatsuta, S. Yanagimoto, S. Warisawa, J.-J. Delaunay, and I. Yamada, "Wearable eating habit sensing system using internal body sound," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 4, no. 1, pp. 158–166, 2010.

[17] O. Amft and G. Tröster, "Recognition of dietary activity events using on-body sensors," *Artificial intelligence in medicine*, vol. 42, no. 2, pp. 121–136, 2008.

[18] ——, "Methods for detection and classification of normal swallowing from muscle activation and sound," in *Pervasive Health Conference and Workshops, 2006*. IEEE, 2006, pp. 1–10.

[19] J. M. Fontana and E. S. Sazonov, "A robust classification scheme for detection of food intake through non-invasive monitoring of chewing," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 4891–4894.

[20] V. Papapanagiotou, C. Diou, L. Zhou, J. v. d. Boer, M. Mars, and A. Delopoulos, "A novel approach for chewing detection based on a wearable ppg sensor," in *In Proceedings of the 38th annual conference of the IEEE Engineering in Medicine and Biology society (EMBC-2016)*, 2016 (accepted for publication).

[21] O. Rasmussen, F. Bonde-Petersen, L. Christensen, and E. Møller, "Blood flow in human mandibular elevators at rest and during controlled biting," *Archives of oral biology*, vol. 22, no. 8-9, pp. 539–543, 1977.

[22] S. Arberet, M. Lemay, P. Renevey, J. Sola, O. Grossenbacher, D. Andries, C. Sartori, and M. Bertschi, "Photoplethysmography-based ambulatory heartbeat monitoring embedded into a dedicated bracelet," in *Computing in Cardiology Conference (CinC), 2013*. IEEE, 2013, pp. 935–938.

[23] M. Lemay, M. Bertchi, J. Sola, P. Renevey, J. Parak, and I. Korhonen, "Applicarion of optical heart rate monitoring," in *Wearable Sensors: Fundamentals, Implementation and Applications*, E. Sazonov and M. R. Neuman, Eds. Elsevier, 2014.

[24] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925–1932, 1999.

[25] C. L. Nikias and J. M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal processing magazine*, vol. 10, no. 3, pp. 10–37, 1993.

[26] M. Altini, J. Penders, R. Vullers, and O. Amft, "Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 219–226, 2015.

[27] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2002.

[28] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1253–1260, 2014.

Vasileios Papapanagiotou is a PhD student in the Multimedia Understanding Group (MUG) of the Information Processing Laboratory (IPL), Department of Electrical and Computers Engineering, Aristotle University of Thessaloniki (AUTh). He received his diploma from the same Department in 2013. His research interests are digital signal processing, wearable sensors, behavioural monitoring and analysis, supervised and semi-supervised machine learning, and concept-based image retrieval.

Christos Diou received his diploma and PhD from the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki, in 2004 and 2010 respectively. Since 2011 he has been working as a Research Associate at the Information Processing Laboratory of the Aristotle University of Thessaloniki in the areas of machine learning and signal processing, for applications such as multimedia analysis and retrieval, analysis of human eating and physical activity behavior, analysis of electrical energy consumption behavior of small-scale consumers as well as machine learning for computer security applications.

Lingchuan Zhou received the B.Sc. degree in electrical engineering from Xian Jiaotong University, Xian, China, in 1998, and the Master and Ph.D. degrees in microelectronics from Louis Pasteur University, Strasbourg, France, in 2001 and 2008, respectively. He worked as R&D engineer at Ela Medical (Sorin Group), Paris, France from 2001 to 2002. He worked as R&D engineer at SCHILLER Médical, Wissembourg, France and then at SCHILLER AG, Baar, Switzerland from 2004 to 2009. Dr. Zhou joined CSEM SA, Neuchâtel, Switzerland, in 2010 and works as Senior Engineer at the Systems Division. His current development activities include monitoring of physiological parameters, wearable devices and eHealth.

Janet van den Boer followed the BSc and MSc program Nutrition and Health at Wageningen University (The Netherlands). In 2013 she graduated and obtained the MSc degree. Currently she is a Ph.D. student at the Sensory Science and Eating Behaviour chair group at the Division of Human Nutrition at Wageningen University. Her main research areas include studying the relation between eating style and food intake.

Dr. ir. Monica Mars (1976) holds a position as assistant professor within the group of Sensory Science and Eating Behaviour of the section of Human Nutrition at Wageningen University and Research (WUR) in the Netherlands. Her scientific interest is on the relation between eating behaviour and weight management. This includes effects of food properties, chewing behaviour and eating rate on food intake. The development and evaluation of smart tools for monitoring and steering eating behaviour is a main topic. She was WorkPackage leader of WP6 Evaluation studies of the SPLENDID project.

Monica Mars received her MSc (1999) and PhD (2004) degree in Nutrition and Health at Wageningen University. She worked as a postdoc and lecturer at WUR and was partially detached to the Top Institute Food and Nutrition. The last years she has been working at Wageningen University, first as researcher and lecturer, and currently as an assistant professor. She has over 40 peer reviewed papers in scientific journals. She supervises several PhD students and is involved in a number of BSc and MSc courses on sensory research and eating behaviour within the curriculum of Nutrition and Health at Wageningen University.

Dr. Anastasios Delopoulos was born in Athens, Greece, in 1964. He graduated from the Department of Electrical Engineering of the National Technical University of Athens (NTUA) in 1987, received the M.Sc. from the University of Virginia in 1990 and the Ph.D. degree from NTUA in 1993. From 1995 till 2001 he was a senior researcher in the Institute of Communication and Computer Systems of NTUA. Since 2001 he is with the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki where he serves as an associate professor. His research interests lie in the areas of machine learning, signal and multimedia processing and computer vision. On the applied domain he works in the areas of multimedia retrieval, biomedical engineering and behaviourali informatics. He is the (co)author of more than 80 journal and conference scientific papers. He has participated in 21 European and National R&D projects related to application of his research to entertainment, culture, education and health sectors. Dr. Delopoulos is a member of the Technical Chamber of Greece and the IEEE.