Inferring Meal Eating Activities in Real World Settings from Ambient Sounds: A Feasibility Study

Edison Thomaz, Cheng Zhang, Irfan Essa, Gregory D. Abowd

{ethomaz, chengzhang, irfan, abowd}@gatech.edu
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Dietary self-monitoring has been shown to be an effective method for weight-loss, but it remains an onerous task despite recent advances in food journaling systems. Semi-automated food journaling can reduce the effort of logging, but often requires that eating activities be detected automatically. In this work we describe results from a feasibility study conducted in-the-wild where eating activities were inferred from ambient sounds captured with a wrist-mounted device; twenty participants wore the device during one day for an average of 5 hours while performing normal everyday activities. Our system was able to identify meal eating with an F-score of 79.8% in a person-dependent evaluation, and with 86.6% accuracy in a person-independent evaluation. Our approach is intended to be practical, leveraging off-the-shelf devices with audio sensing capabilities in contrast to systems for automated dietary assessment based on specialized sensors.

Author Keywords

Activity recognition; Food journaling; Dietary intake; Automated dietary assessment; Ambient sound; Acoustic sensor; Sound classification; Machine learning

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Food journaling has been demonstrated to encourage individual to adopt healthier eating habits since the self-reflection that occurs when people keep track of what they eat plays an important role in behavior change [3]. Unfortunately, despite its benefits, food journaling remains a difficult undertaking; individuals must remember to log meals and snacks throughout the day, and then manually record eating activities on a food diary, a tedious and time-consuming task.

Over the years researchers have suggested various approaches towards automatically recording eating activities. Although promising, they have often required individuals to wear specialized devices such as swallow-detecting neck collars and other sensors that have made many of these systems impractical for everyday use. Additionally, most approaches have focused on fully automating the food tracking process, a direction that is not only highly challenging from a technical perspective, but also misguided from a behavior one since it eliminates the self-reflective benefits brought forth by journaling.

An emerging form of behavior journaling called *semi-automated journaling* aims to minimize the effort associated with manual logging while still keeping individuals directly involved in the activity of interest. One of the keystones required for the implementation of semi-automated or automated food journaling systems is identifying *when* individuals are engaged in an eating activity. Inferring the moment or time window when individuals are consuming food supports a number of scenarios that ultimately help individuals reflect on their diet. For instance, the recognition that eating is taking place could automatically trigger a reminder to capture a relevant food photo. Moreover, if eating moments can be recognized in real time, adaptive systems supporting just-in-time dietary interventions can be realized.

To address the challenge of automatic eating activity detection, we present a system that identifies meal eating moments from ambient sounds using acoustic sensors. Our aim is to improve the practicality of current approaches associated with food journaling by leveraging devices available and in use by individuals. This approach, referred to as opportunistic sensing [11], contrasts to methods that require more specialized forms of sensing modalities (*e.g.,* electromyography for swallow detection). Microphones are simple sensors and virtually ubiquitous; they are guaranteed to be present in mobile handsets across the board, from top of the line smartphones to more basic feature phones. Additionally, audio data is contextually very rich, and has been successfully used in health-focused applications [9].

The two contributions of this work are (1) a practical system for the recognition of meal eating activities in the wild from ambient sounds and (2) a system evaluation using over 100 hours of audio collected *in-the-wild* from 20 participants.

IUI'15, March 29–April 01, 2015, Atlanta, GA, USA
Copyright 2015 ACM 978-1-4503-3306-1/15/03$15.00
http://dx.doi.org/10.1145/2678025.2701405

RELATED WORK

Efforts focused on eating recognition date back to the 1980s when researchers tried to detect chews and swallows using oral sensors in order to measure the palatability and satiating value of foods [21]. Ongoing research work in this area ranges from the use of crowdsourcing techniques [15], wearables [8, 4], and instrumented objects [6]. Sound is a contextually-rich source of information that can be easily recorded using one of the simplest and most ubiquitous sensors; a microphone. Hence, a large body of work at the intersection of acoustic sensing and activity recognition has emerged over the last decade [22, 17, 10].

One of the most explored applications of sound-based activity recognition has been dietary intake tracking, realized through wearable devices. Sazonov *et al.* proposed a system for monitoring swallowing and chewing through the combination of a piezoelectric strain gauge positioned below the ear and a small microphone located over the laryngopharynx [18]. A promising and comprehensive approach to automated dietary monitoring was proposed by Amft *et al.* [1]. It involves having individuals wear sensors in the wrists, head and neck and automatically detect food intake gestures, chewing, and swallowing from accelerometer and acoustic sensor data.

More recently, Yatani and Truong presented BodyScope, a wearable acoustic sensor attached to the user's neck [24]. The system was able to recognize twelve activities at 79.5% F-measure accuracy in a lab study and four activities (eating, drinking, speaking, and laughing) in a in-the-wild study at 71.5% F-measure accuracy.

There are two elements that distinguish our work from previous initiatives around audio-based dietary monitoring. Firstly, our implementation is meant to be practical; our system recognizes eating moments from ambient sounds without the need for specialized sensors. Instead, our approach leverages devices such as smartphones, that individuals have already adopted into their lives. Secondly, the feasibility of our system was tested in-the-wild with twenty participants. Efforts like BodyScope [24] were also evaluated in real world conditions, but in smaller studies.

IMPLEMENTATION

Our system was designed to learn to recognize sounds that are associated with eating activities, such as the background noise in a restaurant environment, and the softer but highly distinguishable sounds generated by the mouth when chewing and biting. This sound identification task presents two technical challenges: the extraction of information-rich features from ambient audio collected with a microphone, and the design of a binary classifier with the ability to distinguish eating sounds from non-eating sounds from audio features.

Practicality was of utmost priority in the design of our system, therefore it does not rely on any specialized sensors. The implementation we propose could run on a smartphone device and was evaluated on the wrist in an effort to simulate a smart watch device or some other wearable piece of technology designed for everyday use.

Audio Frames and Features

Audio was recorded at a sample rate of 11,025Hz (16 bits per sample), and audio frames with size 50ms were extracted using a Hanning-filtered sliding window with an overlap of 50% (block size=552, step size=276). This audio frame size is larger than what is typically chosen for speech recognition applications but adequate to capture environmental sounds.

We extracted 50 features from each frame, using the Python-based Yaafe tool [13]. Based on previous work that also attempted to recognize human activities from audio [10, 17], we chose the following time and frequency domain features: Zero-Crossing Rate [19], Loudness [14], Energy, Envelope Shape Statistics, LPC [12], LSF [2, 20], Spectral Flatness, Spectral Flux, Spectral Rolloff [19], Spectral Shape Statistics [5], and Spectral Variation.

Clustering and Classification

Because many ambient sounds that characterize eating activities are often much longer than a single audio frame, we clustered 400 consecutive frames and calculated the mean and variance of each feature across these frames (Figure 1). This step also reduced feature "noise" that could be introduced if we had accounted for the acoustic characteristics of every single audio frame. For clustering, we applied a sliding window over the audio frame stream, also with 50% overlap. This resulted in a frame cluster vector of size 100 (mean and variance of 50 features). We chose 400 frames for each cluster because that is equivalent to a total of 10 seconds of audio, a duration that can encapsulate sounds of interest that are both short (e.g., the clicking sound of utensils hitting plates or bowls), and long (e.g., background noise in a restaurant). We performed classification with the Random Forest classifier available in the Scikit-learn Python package [16].

USER STUDY

To evaluate our system, we conducted an IRB-approved in-the-wild study, where we recruited participants and examined how our system performed when classifying ambient sounds collected in the real-world, as individuals performed their normal everyday activities. We recruited 21 participants (15 males and 6 females) between the ages of 21 and 55 through our social network, word-of-mouth, flyers and mailing lists. For joining the study, they received $20 as compensation. Participants included students, research scientists, designers, entrepreneurs and other professionals.

The study lasted between 4 and 7 hours on a single day; for 17 participants, the study began in the morning sometime between 8AM and 11AM and ended between 3PM and 4PM, while for 3 participants it began between 4PM and 7PM and ended before 10PM. This time period was enough to guarantee that all study participants had at least one meal (lunch or dinner).

Subjects wore an audio recording device on the wrist. We chose this placement for the collection of ambient sounds because we anticipate that smart watch-type devices will become popular in the near future. It is very likely that these devices will be capable of recording and even analyzing audio, despite their compact size.

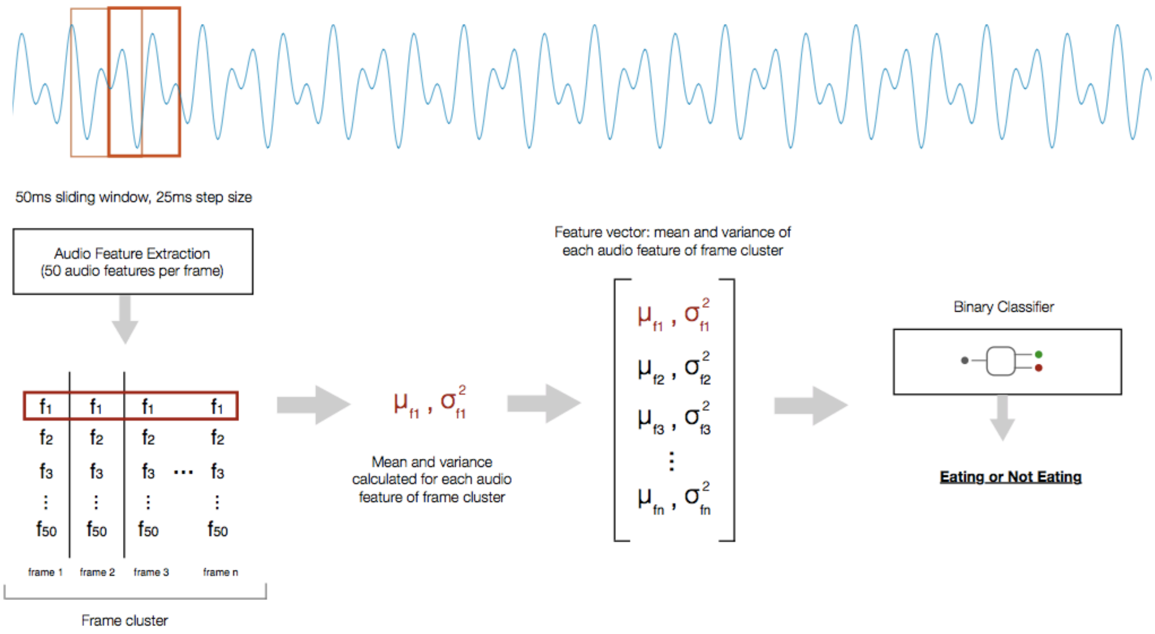
50ms sliding window, 25ms step size
Audio Feature Extraction
(50 audio features per frame)
f1 f1 f1 f1
f2 f2 f2 f2
f3 f3 f3 ··· f3
f50 f50 f50 f50
frame 1
frame 2
frame 3
frame n
Frame cluster
μf1 , σ²f1
Mean and variance calculated for each audio feature of frame cluster
Feature vector: mean and variance of each audio feature of frame cluster
μf1 , σ²f1
μf2 , σ²f2
μf3 , σ²f3
μfn , σ²fn
Binary Classifier
Eating or Not Eating

Figure 1. The audio processing pipeline consists of audio framing, audio feature extraction, frame clustering, frame clustering, and classification.

The audio recorder registered sounds continuously throughout the study. At the end of the study, participants were given the opportunity to review their audio file, and delete any audio segment that they did not want to share with us. After this initial step, we performed a walkthrough of the 4-7 hour study period with participants using the Day Reconstruction Method (DRM) [7]. At the end of this process, we were able to discover when individuals ate during the study interval and segmented and labeled their audio clips accordingly.

Ground Truth

To obtain ambient audio ground truth for the eating activities, we asked participants to recall their activities for the day and list them in order, indicating an estimated beginning and end time for each activity. This activity list in chronological order allowed us to discover if and when the participant had a meal. To make sure that time periods indicated by participants were in fact eating activities, two of the authors coded the audio files independently after agreeing on a guideline and then compared results. Disagreements beyond a range of 5 minutes at the beginning or end of an eating activity audio segment were discussed; there were 5 disagreements in total. The final set of ground truth data for each participant included the audio clip referring to the reported eating activity, and another clip with all the audio except for the eating activity segment. As expected, the eating activity audio clip was always much shorter in duration than the audio clip of non-eating activities.

RESULTS

To reiterate, our high-level goal is to develop and evaluate a *practical* approach to detect when meals are being consumed in the wild. In this work, the primary performance metric we wished to assess was whether the system could identify meal eating activities from ambient sounds. This assessment was driven by collecting data in real situations and learning models from the data to test our approach.

We evaluated our models using a person-dependent technique and reported results in terms of precision, recall and F-score metrics (Table 1); we performed 10-fold cross-validation on each study participant's data and then averaged the results across all participants to obtain an overall result. For comparison, we tested three different classifiers: Support Vector Machines (SVM), Nearest Neighbors (n=5), and Random Forest. The Random Forest classifier proved to be vastly superior to the other two classifiers, yielding an F-score of 79.8%. As a means of comparison, this result is equivalent to what Yatani et al. achieved with BodyScope [24]. On one hand, BodyScope was able to recognize multiple activities. On the other hand, our system does not require any specialized sensor, and can run in any off-the-shelf device that is capable of recording and processing audio, such as smartphones and smart watches.

A LOPO (leave-one-participant-out) cross-validation resulted in an F-score of 28.7%, suggesting that this approach would greatly benefit from personalization. It is important to note that F-measures below 50% are not uncommon in LOPO evaluations, particularly in the context of free-living studies [24].

DISCUSSION

Our ambient audio dataset included meal eating activities in a wide variety of contexts. Participants ate alone and with friends; they ate at home, at work, at school and in the classroom. Although desirable, this level of variety in the data made the classification task particularly challenging.

Classifier	Precision	Recall	F-score
SVM	47.5%	50.5%	48.9%
5-NN	53.3%	51.9%	51.4%
Random Forest	**89.6%**	**76.3%**	**79.8%**

Table 1. Person-dependent, 10-fold cross-validation results for each classified we evaluated. The Random Forest classifier performed significantly better that the SVM and Nearest Neighbors classifiers.

One factor that hampered the classifier's ability to identify meal eating was the short duration of meal events, which were shorter than 12 minutes in some cases. This resulted in a small number of frame clusters for the classifier to examine, and a misclassification proved very costly. Another difficulty was that some of the participants had their meals while performing other activities such as attending a class or working in the computer, which were not labeled as meal eating activities. It is likely that additional examples would help with activity class separation in this case. Finally, classifying meal-eating in quiet environments, such as one's office or home, has obvious challenges. This suggests a design rationale for training the classifier while emphasizing the specific characteristics of different sounds environments (e.g. home, school, restaurant).

Despite these difficulties, it is worth noting that it would have been impractical to evaluate our system in a controlled lab setting, since it would have been devoid of most of the natural environmental sounds that individuals are enveloped in when in real world settings and conditions.

Ground Truth

Estimating ground truth from the audio files proved to be a challenging undertaking. Individuals were asked to recall the exact time they had meals, but often could not do so accurately. In some cases, finding this segment proved particularly difficult, especially when the length of the meal was under 10 minutes. Moreover, while in some audio clips it was possible to hear that participants were eating or were in a restaurant environment, in other clips this was not clear at all. For instance, participants P9 and P14 ate in a classroom or classroom-like environment, whose sounds could not be easily identified as those that are characteristic of an eating activity. In these situations we had to rely on subtle cues, such as the sound of a food container coming out of a brown bag.

Another difficulty we faced in obtaining ground truth had to do with the characterization of an eating activity. Some participants had hour-long lunches, where they chatted with friends extensively before, during and after the meal. On the other hand, some participants had very short meals, eating uninterruptedly for 10 or 15 minutes. In the case of the long lunch, a question might be raised as to whether the whole meal event should be labelled as "eating" or only the period when individuals were actively eating.

Data Collection

Although our feasibility study represents a large ecologically-valid data collection effort, it is limited in two important ways. First of all, since participants joined the study for 4-6 hours in a single day, ambient audio data was recorded for only one meal of their day. For most participants the recorded meal was lunch. The system was evaluated on a per-participant basis through cross-validation, but having just one example of a meal eating activity per participant lowers the confidence that our results generalize over several days. In the future, we plan to address this weakness by collecting data for multiple days per participant. Additionally, the lack of multi-day audio data makes it unlikely that our system's capability to infer eating activities generalizes across individuals. Although we plan to evaluate our system using a person-independent metric in the future, we believe that most applications and interfaces built on top of our implementation will be personalized (e.g., a just-in-time intervention tailored to address an individual's specific challenges).

Secondly, snacking behavior was not the focus of this study. The duration of data collection per day combined with the times when the study began and ended precluded us from capturing ambient audio around snack-eating activities. However, there is no question that snacking is a highly relevant behavior, and we plan to improve our study design and techniques to account for it in the future. Having said this, a few of the meal eating activities logged in our feasibility study were shorter than 10 minutes, which more closely matches snack eating duration than a "traditional" meal eating duration. The truth is that there is a great deal of ambiguity when it comes to characterizing an eating activity as meal eating versus snack eating.

One of the key issues in audio-based activity recognition is privacy. Understandably, most people object to the recording and analysis of audio of their everyday lives, particularly if it is done completely autonomously and without human input. In our implementation we did not address this challenge, although techniques for protecting privacy in audio streams, and conversational speech in particular, have been proposed [23].

CONCLUSION

Based on our results, and despite the limitations of our study, it is clear that acoustic sensing represents a promising opportunity. Our system was able to identify meal eating with 89.6% precision and 76.3% recall in a person-dependent evaluation. Although our focus in this work is on the binary presence of eating moments in an audio stream, there are many other dimensions of eating that are relevant from a diet and behavior change perspective. With audio, it might be possible to determine whether individuals are eating alone or with friends, and whether they are eating while working (e.g. typing in a computer) or watching television. We hope to extend our audio-based activity classification platform in the future to capture these additional contextual parameters.

ACKNOWLEDGEMENTS

We would like to thank Thomas Ploetz for his feedback. This work was supported by the Intel Science and Technology Center for Pervasive Computing (ISTC-PC), and by the National Institutes of Health under award 1U54EB020404-01.

REFERENCES

1. Amft, O., and Tröster, G. Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine 42*, 2 (Feb. 2008), 121–136.

2. Bäckström, T., and Magi, C. Properties of line spectrum pair polynomials—A review. *Signal Processing 86*, 11 (Nov. 2006), 3286–3298.

3. Burke, L. E., Wang, J., and Sevick, M. A. Self-Monitoring in Weight Loss: A Systematic Review of the Literature. *YJADA 111*, 1 (Jan. 2011), 92–102.

4. Dong, Y., Scisco, J., Wilson, M., Muth, E., and Hoover, A. Detecting periods of eating during free living by tracking wrist motion. *IEEE Journal of Biomedical Health Informatics* (Sept. 2013).

5. Gillet, O., and Richard, G. Automatic transcription of drum loops. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE (2004), iv–269–iv–272.

6. Kadomura, A., Li, C.-Y., Tsukada, K., Chu, H.-H., and Siio, I. Persuasive technology to improve eating behavior using a sensor-embedded fork. In *the 2014 ACM International Joint Conference*, ACM Press (New York, New York, USA, 2014), 319–329.

7. Kahneman, D., Krueger, A. B., Schkade, D. A., and Schwarz, N. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* (2004).

8. Kalantarian, H., Alshurafa, N., and Sarrafzadeh, M. A Wearable Nutrition Monitoring System. In *Wearable and Implantable Body Sensor Networks (BSN), 2014 11th International Conference on* (2014), 75–80.

9. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. StressSense: detecting stress in unconstrained acoustic environments using smartphones. In *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM Request Permissions (Sept. 2012).

10. Lu, H., Pan, W., Lane, N., Choudhury, T., and Campbell, A. SoundSense: scalable sound sensing for people-centric applications on mobile phones. *Proceedings of the 7th international conference on Mobile systems, applications, and services* (2009), 165–178.

11. Lukowicz, P., Pentland, A. S., and Ferscha, A. From Context Awareness to Socially Aware Computing. *IEEE pervasive computing 11*, 1 (2012), 32–40.

12. Makhoul, J. Linear prediction: A tutorial review. *Proceedings of the IEEE 63*, 4 (Apr. 1975), 561–580.

13. Mathieu, B., Essid, S., Fillon, T., Prado, J., and Richard, G. *YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software*. In *proceedings of the 11th ISMIR conference, 2010* (Sept. 2010).

14. Moore, B. C. J., Glasberg, B. R., and Baer, T. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *Journal of the Audio Engineering Society 45*, 4 (1997), 224–240.

15. Noronha, J., Hysen, E., Zhang, H., and Gajos, K. Z. Platemate: crowdsourcing nutritional analysis from food photographs. *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011), 1–12.

16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

17. Rossi, M., Feese, S., Amft, O., Braune, N., Martis, S., and Tröster, G. AmbientSense: A real-time ambient sound recognition system for smartphones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on* (2013), 230–235.

18. Sazonov, E., Schuckers, S., Lopez-Meyer, P., Makeyev, O., Sazonova, N., Melanson, E. L., and Neuman, M. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. *Physiological Measurement 29*, 5 (Apr. 2008), 525–541.

19. Scheirer, E., and Slaney, M. Construction and evaluation of a robust multifeature speech/music discriminator. *IEEE Internation Conference on Acoustics, Speech and Signal Processing, p.1331-1334, 1997. 2* (1997), 1331–1334.

20. Schussler, H. A stability theorem for discrete systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing 24*, 1 (Feb. 1976), 87–89.

21. Stellar, E., and Shrager, E. E. Chews and swallows and the microstructure of eating. *The American journal of clinical nutrition 42*, 5 (1985), 973–982.

22. Ward, J. A., Lukowicz, P., Tröster, G., and Starner, T. E. Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 10 (2006), 1553–1567.

23. Wyatt, D., Choudhury, T., and Bilmes, J. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. *Proceedings of Interspeech* (2007), 586–589.

24. Yatani, K., and Truong, K. N. BodyScope: a wearable acoustic sensor for activity recognition. *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (2012), 341–350.