A Smartwatch-Based System for Audio-Based Monitoring of Dietary Habits

Haik Kalantarian, Majid Sarrafzadeh
Wireless Health Institute, Department of Computer Science
University of California, Los Angeles
Email: {kalantarian, majid}@cs.ucla.edu

***Abstract*—In recent years, smartwatches have emerged as a viable platform for a variety of medical and health-related applications. In addition to the benefits of a stable hardware platform, these devices have a significant advantage over other wrist-worn devices, in that user acceptance of watches are much higher than other custom hardware solutions. In this paper, we describe the development of an Android application on a Samsung smartwatch device for evaluating eating habits using a microphone and various signal processing techniques. Though other works on acoustic monitoring of food habits have been conducted, the varied arm movement during eating creates a unique set of challenges that our work attempts to address. Evaluation results confirm the efficacy of our technique; classification was performed between apple and potato chip bites, water swallows, and talking, with an F-Measure of 94.5% based on 200 collected samples. *Index Terms*—smartwatch; nutrition; spectrogram**

I. Introduction

There is little doubt that obesity is associated with various negative health outcomes such as an increased risk for stroke, diabetes, various cancers, heart disease, and other conditions. In 2008, medical costs associated with obesity were estimated to exceed $147 billion, with over one-third of adults in the United States estimated to be obese [1]. The two major contributors to weight gain are an inactive lifestyle and poor diet. Though the former has been addressed by many wearable devices in recent years both in research and the consumer electronics field, few works exist on automatic detection of dietary habits in an inconspicuous form-factor [2][3][4]. Instead, characterization of an individual's eating habits is possible through manual record keeping such as food diaries, 24-hour recalls, and food frequency questionnaires. However, these approaches suffer from low accuracy, high user burden, and low rates of long-term compliance. Wireless health-monitoring technologies have the potential to promote healthy behavior and address the ultimate goal of enabling better lifestyle choices.

In recent years, several electronic devices have been proposed for monitoring dietary habits. However, most works attempt to characterize eating from patterns in chewing and swallow counts, and very few attempt to identify the nutritive properties of the foods themselves. Therefore, a fundamental question in the field of electronic food monitoring is the validity of chew and swallow counts as a heuristic for estimation of Caloric intake. A very recent work by Fontana et al. [5] addresses this issue by comparing several different techniques for estimation of Caloric intake: weighed food records (gold standard), diet diaries, and electronic sensor-based measurements of chews and swallows. The conclusion of this study was that chew and swallow counts were more closely affiliated to the gold standard measurement than self-reporting methods and photographic food records.

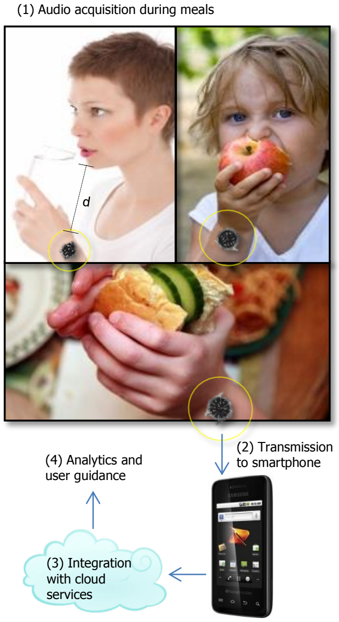
(1) Audio acquisition during meals
d
(2) Transmission to smartphone
(4) Analytics and user guidance
(3) Integration with cloud services

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

Many prior works address the problem of nutrition monitoring by processing audio signals associated with ingestion [6][7]. Typically, these systems use a throat microphone for recognizing deglutition (swallows), or using time-frequency decomposition techniques, such as Wavelet Packet Decomposition (WPD) or Spectrogram Analysis to extract distinctive features, and either classify between different food groups or recognize anomalies in swallow patterns. While many of these works are novel from a perspective of algorithmic techniques, they generally propose custom hardware solutions or bulky non-standard equipment which are of limited use outside of clinical environments.

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

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

The paper is structured as follows. In Section II, we present related work. In Section III, we describe the system architecture. Section IV presents an overview of the proposed algorithms for classification. Section V describes the experimental procedure, followed by results in Section VI. Section VII concludes the paper.

II. Related Work

Many works have proposed detecting food intake using static microphone placement, generally on the throat. For example, the work in [8] uses acoustic data acquired from a small microphone placed near the bottom of the throat. Their system is coupled with a strain gauge placed near the ear. Other works attempt to characterize and address swallow disorders in seniors, such as dysphagia [9].

In the work by Amft et al. in [10], authors analyze bite weight and classify food acoustically from an earpad-mounted sensor. In [7], the authors present a similar earpad-based sensor design to monitor chewing sounds. Food grouping analysis revealed three significant clusters of food: wet and loud, dry and loud, soft and quiet. An overall recognition accuracy of over 86.6% was achieved. A more recent study using support vector machines have been able to reach swallow detection accuracies of up to 84.7% in an in-lab setting [11]. These devices are mounted very high in the upper trachea, near the laryngopharynx.

In [12], Pler et al. proposed a system geared towards patients living in ambient assisted living conditions and used miniature electret microphones which were integrated into a hearing aid case, and placed in the ear canal. In [13], the authors are able to achieve a food detection accuracy of 79% using hidden Markov models based on data acquired from microphones in the ear canal.

III. System Architecture

Our proposed system does not require any custom hardware: the Android application runs on Samsung Galaxy Gear smartwatch running Android 4.2.1. This phone features an 800 MHz ARM-based processor, 512 MB of RAM, and a 320x320 pixel 1.6 inch display. The device also supports transfer of data using the Bluetooth LE protocol, and can be configured to access the Internet using Bluetooth tethering with compatible smartphones.

Data was recorded using the Samsung Galaxy Gear microphone in MPEG-4 Part 14 (m4a) format at a rate of 96 kbps, as prior research has shown that the spectral energy for many common foods is between 0-10 kHz, with highest amplitude ranges between 1 and 2 kHz for water [14][15].

IV. Algorithm Design

A. Frequency-Domain Evaluation: Liquids

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

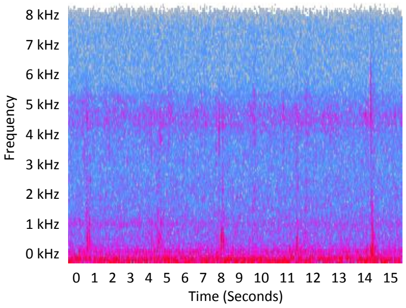

Fig. 2. A spectrogram of an audio clip consisting of five swallows, generated with a window of size 1024 samples. There is a visible change in the spectral density at points corresponding with swallows as shown above.

Figure 2 shows a spectrogram corresponding with an audio clip consisting of five water swallows acquired from the smartwatch. This spectrogram is generated with the Short-Time Fourier Transform, and shows changes in the frequency distribution over time [6]. Figure 3 shows a more detailed comparison between a brief interval of noise (1s) and a water swallow. Generally speaking, the data of interest is between 600 Hz and 1 kHz, as shown by the deviation between the signals at this time, and confirmed by the spectrogram shown in Figure 2. We conclude that analysis of this frequency range is critical for classification of liquid swallows. This observation is confirmed by Figure 4, which shows the transformation of an audio signal corresponding with ten swallows. The top waveform is the original, while the bottom is the post-processed filter output in which noise is substantially reduced. This is achieved by band-pass filtering the audio data with cutoffs of 600 Hz and 1 kHz and a rolloff of 48 dB- meaning

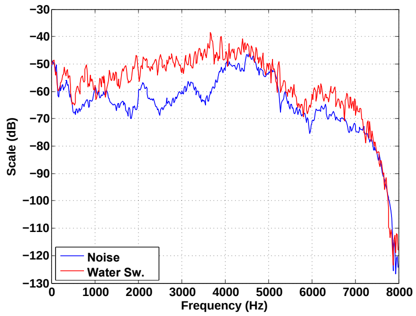

Fig. 3. Frequency distribution of a water swallow vs. silence (noise). This graph reveals that the frequency range between 500 Hz and 1000 Hz is the point of interest.

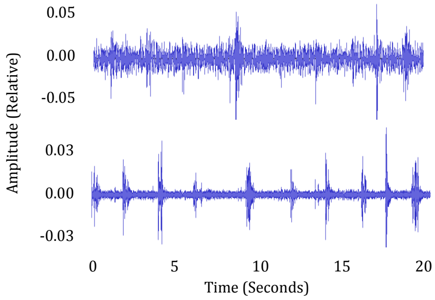

Fig. 4. Post-processing of the audio signal corresponding with water can dramatically improve signal-to-noise ratio. The top shows the original waveform. The bottom shows the waveform after a bandpass filter is applied.

the amplitude decreases by 48 dB for each octave outside the filter threshold.

While the resulting signal clearly shows the swallows, marked by pronounced peaks, this technique is not very generalizable to other foods besides water, because the data is pre-processed. In the case of the frequency distribution of a one second window around the initial bite of a potato chip, compared to an equal period of chewing, the amplitude of the bite signal is greater from 600 Hz to 4 kHz. However, the pattern is not as distinctive as for liquids, and may certainly vary between individuals with different eating styles. A more generalizable approach is described in the next subsection.

B. Generalizable Feature Extraction

Detection of eating habits is somewhat different than that of liquid consumption, as the smartwatch will not be near the throat during a swallow. Therefore, in these cases we attempt to identify when an individual bites into a food item rather than chewing. The smartwatch platform is particularly well suited for this application because the microphone will be nearest to the sound source during the times at which the signal is of interest. The proposed model must be flexible to identify biting and swallowing for many different foods and drinks, between individuals with varying eating styles.

TABLE I. Partial List of openSMILE Speech Features from [16]

Speech-Related Features		
Signal Energy	Loudness	Mel/Bark/Octave Spectra
MFCC	PLP-CC	Pitch
Voice Quality	Formants	LPC
Line Spectral Pairs	Spectral Shape	CENS and CHROMA

TABLE II. Partial List of openSMILE Statistical Features from [16]

Speech-Related Features		
Means	Extremes	Moments
Segments	Samples	Peaks
Zero Crossings	Quadratic Regression	Percentiles
Duration	Onset	DCT Coefficient

openSMILE [17] is a feature extraction tool intended for producing large audio feature sets. This tool is capable of various audio signal processing operations such as applying window functions, fast-Fourier transforms,finite impulse response filterbanks, autocorrelation, and cepstrum. In addition to these techniques, openSMILE is capable of extracting various speech related features and statistical features. A partial list of extracted features is shown in Tables I and II, respectively. After data is collected from a variety of subjects eating several foods, classifiers can be used to identify strong features that are accurate predictors of swallows and bites for various foods, while reducing the dimensionality by eliminating redundant or weakly correlated features.

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

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

V. Experimental Procedure

A. Data Collection for Recognition

A total of ten subjects were used for data collection, with ages ranging from 22 to 35 in order to develop a model for identifying swallows. The subjects included eight males and two females. Each subject was asked to eat the following foods, in order: three apple slices with at least two bites per slice, one 8 oz. glass of room-temperature water, and one bag of potato chips. The moments at which the food was bitten into (or swallowed as in the case of the water) were manually annotated.

```
RecordAudio(Buffer)
if Buffer.Utilization = $\beta$ then
    d = Buffer[1:$\beta$];
    f = ExtractFeatures(D);
    s = {Water, Talk, Apple, Chips, Other};
    c = Classify(F,s);
Counter$_c$++;
if $c \neq$ Other then
    PauseRecording($\tau$)
```

Fig. 5. Simplified Classification Scheme

Data was manually extracted from the audio recordings at a later time. Regardless of the food or activity type, each sample was exactly 0.25 seconds in length, and the peak of the wave amplitude was not necessary centered in the window. In some cases, such as during the biting of an apple, one quarter of a second was not sufficient to capture the entire bite.

Subjects were also asked to read a brief passage from a Wikipedia article, with no particular instruction about the rate at which they should read. The data was then automatically split into 0.25 second audio fragments using an audio processing program. Therefore, some samples were collected between phrases, and were relatively silent.

B. *Smartwatch Feedback: A Survey*

Before the system development phase, we had several important questions about how individuals feel about smartwatches. As described previously, a wearable device must have both high accuracy, and high rates of user adherence for the subject to reach his or her intended goals. Furthermore, we proposed several questions about which hand a subject prefers to wear a watch. For example, our experimental evaluation requires that subjects wear a watch on the same hand with which they typically eat food such as chips or raise a glass of water.

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

VI. Results and Discussion

A. *Audio Classification*

Results for classification between apples, chips, water, and speaking are shown in Table IV based on 50 unprocessed samples collected from each of these foods, using the Random Forest classifier [18] with 6555 extracted features from each sample. The Random Forest classifier consisted of 100 trees, each constructed using 13 random features, and was validated using 10-fold cross validation. This particular classifier was chosen for its high accuracy in our experimentation- several other classifiers performed poorly in comparison. A total of 189 instances were classified correctly (94.5%) while the remaining 11 (5.5%) were classified incorrectly. The weighted average precision, recall, and F-Measure was 94.6%, 94.5%, and 94.5% respectively. Classification of water and speaking were particularly accurate, with only one incorrect classification. The majority of classification errors were between apples and potato chips.

B. *Feature Extraction*

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

The top ten features are listed in Table III. The first feature is the skewness of the logarithmic signal energy, in which skewness is defined as the asymmetry of the variable in comparison with a normal probability distribution [19]. More formally, skewness is defined in below, where μ_i is the ith central moment about the mean.

$$\gamma_1 = \frac{\mu_3}{{\mu_2}^{3/2}} \tag{1}$$

The ith moment M$_i$ moment of a discrete function f(x) defined on an interval [a,b] can be generalized as:

$$M_i = \sum_{x=a}^{b} x^i f(x) \tag{2}$$

To calculate the moment about the mean for a probability density function, it is necessary to first calculate the mean *m*. The ith moment about the mean can be represented as:

$$\bar{M}_s = \frac{\sum_{i=a}^{b}(x_i - m)^s}{b-a} \tag{3}$$

Therefore, for a probability density function f(x), the first moment about the mean is always zero (with s = 1), while the second moment is the variance. The third central moment is defined as skewness such that a distribution skewed to the right has a positive value, while one shifted towards the left has a negative skewness.

The second most highly correlated feature is the mean peak distribution, which is defined as the mean distance between peaks for the logarithmic representation of the signal energy. The third feature is the number of non-zero values of the normalized log-energy signal.

Features 4-10, preceded by MFCC, are Mel-Frequency Cepstral Coefficients [20], which represent the spectral characteristics of the signal. A cepstrum is the result of the Inverse Fourier Transform of the logarithm of a signal spectrum. Mel-Frequency Cepstral Coefficients are based on the mel scale, which is a perpetual scale of pitches judged by listeners to be equidistant from one another [21]. The relationship between the frequency and mel scales is logarithmic, and can be defined by the following formula (though other variations exist) [21]:

$$MEL(f) = 2595 \cdot log_{10}(1 + \frac{f}{700}) \tag{4}$$

Assume we attempt to obtain the MFCC of a 0.25 second audio clip. The first step is to compute the Discrete Fourier Transform of the window, Y(k), from which we can obtain the power spectral density (PSD), P(k) using the following formula in which W is the window size:

$$P(k) = \frac{1}{W}|Y(k)|^2 \tag{5}$$

Next, we must obtain an estimate of the energies of different frequency ranges in the signal. However, the human ear can discern differences in frequency at low frequency ranges with a much higher resolution than at higher ranges, due to the physical properties of the cochlea. Therefore, a Mel Filterbank [22] is applied to the signal, which consists of N partially overlapping triangular window functions in the frequency domain. At higher frequencies, the triangular filters are wider, because we are less concerned with small variations in energy in these frequency ranges. [23]. Generally, N is a value between 20 and 40, with each window in the filterbank equally spaced in the Mel domain, which ranges from 300 Hz to 8000 Hz for speech-processing applications.

A dot product is computed between the filterbank and vector P(k), which yields N intermediary coefficients- one for each triangle window function in the filterbank. Because humans do not perceive loudness on a linear scale, the logarithm is calculated for all N coefficients. Finally, the Discrete Cosine Transform [24] (DCT) of the log powers is applied in order to decorrelate the energies of the overlapping filterbank energies. The resulting coefficients are used to extract statistical features as shown in Table III.

C. Smartwatch Feedback: A Survey

Figure 6 provides several pertinent questions from the survey. From the total sample of 221 respondents, 86% claimed to be right handed, 12% right-handed, and the remaining responded that they were 'unsure' or the question was 'not applicable'. In the following question, a total of 76% of respondents stated that they generally would wear a watch on their left hand, with an additional 19% who preferred to wear the watch on their right. The remaining 5% of those surveyed expressed no preference.

TABLE III. Partial List of Selected Features

Rank	Feature Name
1	pcm_LOGenergy_sma_skewness
2	pcm_LOGenergy_sma_meanPeakDist
3	pcm_LOGenergy_sma_nnz
4	mfcc_sma[0]_quartile3
5	mfcc_sma[0]_meanPeakDist
6	mfcc_sma[0]_nnz
7	mfcc_sma[1]_quartile2
8	mfcc_sma[1]_meanPeakDist
9	mfcc_sma[1]_peakMean
10	mfcc_sma[1]_amean

TABLE IV. Audio: Confusion Matrix (Random Forest)

	Predicted Outcome				
Swallow Type	Apple	Chips	Water	Talk	Recall
Apple	46	4	0	0	92%
Chips	6	44	0	0	88%
Water	1	0	49	0	98%
Talk	0	0	0	50	100%
Precision	86.7%	89.7%	100%	100%	

The next question asked respondents how they felt about wearing watches in general. Most individuals stated that they always wear a watch (38%). However, 23% claimed that they preferred not to wear a watch, 24% stated that they would not mind, and 14% stated that they like to wear a watch. Only 1% of individuals claimed that they would not consider wearing a watch. However, another survey question revealed that those who drank water out of a glass would use their primary hand to lift the cup for their mouth (69%), rather than the secondary hand on which the watch is worn (20%) with a remaining 10% claiming to be unsure. Clearly, this would pose a challenge to detection of liquid consumption.

The next question asked respondents if they would be willing to wear a watch on the opposite hand to which they are accustomed. The results were quite promising, with 40% of respondents answering 'maybe', 32% answering 'yes', and 28% answering 'no'. It appears that enough individuals are willing to change which hand they wear their watch, to make detection of most eating habits possible if the algorithm settings are customized to their personal habits.

VII. Conclusion and Future Work

This paper presents a novel approach to detecting ingestion of foods and liquids, using a Samsung smartwatch for identification of bites and swallows from acoustic signals. We conclude that the smartwatch platform is a strong choice for non-invasive evaluation of eating habits, and the versatility and comfort of the watch platform is a substantial advantage over existing schemes that rely on custom hardware solutions. This paper also presents a survey of users about smartwatch usage which confirms that a substantial portion of individuals would be willing to wear a watch on the hand with which they primarily eat.

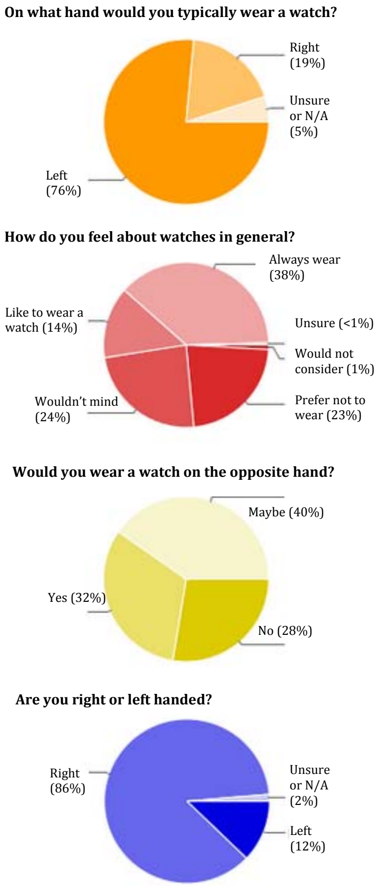

Fig. 6. Partial survey results are shown above.

In future works, we would like to automatically detect the hand on which the watch is being worn, and modify the classification thresholds accordingly in order to improve classification accuracy. This is necessary because the magnitude of the signal will vary if the watch is not worn on the same hand used to pick up an item of food. We would also like to explore the integration of audio-based detection of eating with inertial sensors for gesture recognition. Because smartwatches include an accelerometer and gyroscope, detection of eating-related motions coupled with audio data can improve our ability to characterize a meal.

Acknowledgment

This work is funded by the National Science Foundation AIR Option 1: Award Number 1312310

References

[1] "Centers for disease control and prevention. annualmedical spending attributable to obesity: Payer-and service-specific estimates," http://www.cdc.gov/obesity/data/adult.html, accessed: 2015-02.

[2] P. S. Freedson, E. Melanson, and J. Sirard, "Calibration of the Computer Science and Applications, Inc. accelerometer," *Med Sci Sports Exerc*, vol. 30, no. 5, pp. 777–781, May 1998.

[3] P. S. Freedson, K. Lyden, S. Kozey-Keadle, and J. Staudenmayer, "Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample," *J. Appl. Physiol.*, vol. 111, no. 6, pp. 1804–1812, Dec 2011.

[4] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 21, 2012.

[5] J. Fontana, J. Higgins, S. Schuckers, and E. Sazonov, "Energy intake estimation from counts of chews and swallows," *Journal of Appetite*, no. 85, pp. 14–21, 215.

[6] H. Kalantarian, N. Alshurafa, M. Pourhomayoun, S. Sarin, T. Le, and M. Sarrafzadeh, "Spectrogram-based audio classification of nutrition intake," in *IEEE EMBS Healthcare Innovations & Point of Care Technologies (HIPT)*, 2014.

[7] O. Amft, "A wearable earpad sensor for chewing monitoring," in *Sensors, 2010 IEEE*, Nov 2010, pp. 222–227.

[8] 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.

[9] M. Nagae and K. Suzuki, "A neck mounted interface for sensing the swallowing activity based on swallowing sound," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, Aug 2011, pp. 5224–5227.

[10] O. Amft, M. Kusserow, and G. Troster, "Bite weight prediction from acoustic recognition of chewing." *IEEE Trans. Biomed. Engineering*, vol. 56, no. 6, pp. 1663–1672, 2009.

[11] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE Trans Biomed Eng*, vol. 57, no. 3, pp. 626–633, Mar 2010.

[12] S. Passler and W. 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*, May 2011, pp. 266–269.

[13] W. Pler S and F. W. M, "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food." *Physiological Measurement*, pp. 1073–1093, jun 2012.

[14] D. Brochetti, M. Penfield, and S. Burchfield, "Speech analysis techniques: A potential model for the study of mastication sounds," *Journal of Texture Studies*, vol. 23, no. 2, pp. 111–138, 1992. [Online]. Available: http://dx.doi.org/10.1111/j.1745-4603.1992.tb00515.x

[15] W. E. HI, A. E. Deibel, C. T. Glembin, and E. Munday, "Analysis of food crushing sounds during mastication: Frequency-time studies1," *Journal of Texture Studies*, vol. 19, no. 1, pp. 27–38, 1988. [Online]. Available: http://dx.doi.org/10.1111/j.1745-4603.1988.tb00922.x

[16] "Opensmile faq," http://www.audeering.com/research/opensmile, accessed: 2015-02.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874246

[18] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[19] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 66–78, April 2011.

[20] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 93–96.

[21] D. O'Shaughnessy, *Speech communication: human and machine*. Addison-Wesley, 1987.

[22] B. Milner and X. Shao, "Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model." in *INTERSPEECH*. Citeseer, 2002.

[23] "Mel frequency cepstral coefficient (mfcc) tutorial," http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[24] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *Computers, IEEE Transactions on*, vol. 100, no. 1, pp. 90–93, 1974.