BodyScope: A Wearable Acoustic Sensor for Activity Recognition

Koji Yatani[1,2] and Khai N. Truong[2]

[1]Microsoft Research Asia
Beijing, China
koji@microsoft.com

[2]University of Toronto
Toronto, ON Canada
khai@cs.toronto.edu

ABSTRACT

Accurate activity recognition enables the development of a variety of ubiquitous computing applications, such as context-aware systems, lifelogging, and personal health systems. Wearable sensing technologies can be used to gather data for activity recognition without requiring sensors to be installed in the infrastructure. However, the user may need to wear multiple sensors for accurate recognition of a larger number of different activities. We developed a wearable acoustic sensor, called BodyScope, to record the sounds produced in the user's throat area and classify them into user activities, such as eating, drinking, speaking, laughing, and coughing. The F-measure of the Support Vector Machine classification of 12 activities using only our BodyScope sensor was 79.5%. We also conducted a small-scale in-the-wild study, and found that BodyScope was able to identify four activities (eating, drinking, speaking, and laughing) at 71.5% accuracy.

Author Keywords

Activity recognition, wearable sensor, acoustic sensor, machine learning.

ACM Classification Keywords

H.5.2. [User interface]: Input devices and strategies; H.5.5. [Sound and music computing]: Signal analysis, synthesis, and processing.

General Terms

Human Factors

INTRODUCTION

The ability of designers to develop a variety of truly ubiquitous computing applications (*e.g.,* context-aware, lifelogging, and personal health systems) may depend on the existence of tools and techniques for continuously sensing user activities of interest. Wearable sensing technologies can be used to gather sensor data for activity recognition without relying on sensors to be installed in the infrastructure. Thus, researchers have explored how to perform activity recognition in a practical way using a variety of wearable sensors, such as location beacons, accelerometers, cameras, and physiological sensors. Despite the existence of these many different sensing technologies, there is still a wealth of activities that are not easily detected with a single sensor. Often, in order to accurately infer a large number of user activities, many sensors must be used in combination. But, users might be reluctant to carry or wear multiple sensors in real practice. Mobile phones could be a good platform for daily activity recognition [19, 20, 21]; however, they may not be always with the user [25], and therefore may miss some activities.

In this project, we focus on the sounds produced from different user activities that involve the user's mouth and throat. For instance, when we speak to someone, we generate vocal sounds. When we eat or drink, we produce chewing, sipping, and swallowing sounds [3]. We are interested in exploiting the sounds naturally produced from the user's mouth and throat area to recognize a wide variety of user activities, yet with only a single sensor.

We developed a wearable acoustic sensor for activity recognition, called BodyScope (Figure 1). We modified the Bluetooth headset to embed a microphone into one of its earpiece and then covered it with the chestpiece of a stethoscope. Because the uni-directional microphone points towards the user, this eliminates external noises. The chestpiece then amplifies the sounds produced inside the throat, enhancing features that differentiate the recorded audio for one user activity from another.

To evaluate its effectiveness, we conducted two user studies. Our laboratory study collecting 12 kinds of activities (*e.g.,* eating, drinking, speaking, laughing and coughing) with BodyScope reveals that the system was able to differentiate between the activities with Support Vector Machine classification at 79.5% accuracy (F-measure). Our small-scale in-the-wild study found that BodyScope was able to

Figure 1. The BodyScope prototype. It consists of a Bluetooth headset, a microphone (embedded in the headset), and the chestpiece of a stethoscope.

UbiComp' 12, Sep 5 – Sep 8, 2012, Pittsburgh, USA.

identify four activities (eating, drinking, speaking, and laughing) at 71.5% accuracy. Our results indicate the potential use of an acoustic wearable sensor recoding the sound produced around the throat area to recognize a wide variety of activities, and its integration into existing ubiquitous computing applications.

RELATED WORK

Activity Recognition with Wearable Sensors

Location sensors, such as a Global Positioning System (GPS) module are used commonly for activity recognition. Liao *et al.* examined a method for recognizing activities related to locations, such as working, and sleeping, based on GPS data [18]. Their method identifies and labels significant locations and activities from the GPS data using a conditionally-trained relational Markov network and tree representation. Their experiment showed that their method was more accurate than a simple threshold-based method.

An inertial sensor is another sensing technology used frequently in performing activity detection. Accelerometers, in particular, can be used for detecting the user's physical movements. Farringdon *et al*. [11] developed a wearable device with a built-in accelerometer that recognized the user's posture and movement (*e.g.*, sitting, standing, and walking) using a simple threshold-based method. Foerster *et al.* demonstrated that specific placements of two-axis accelerometers on four different areas of the body (on the chest, wrist, thigh, and lower leg) enabled the recognition of nine physical activities, including walking, sitting, talking, and cycling [12]. Using the 1-nearest neighbor algorithm, they were able to achieve 95.8% recognition accuracy in a laboratory setting. Bao and Intille also tested a similar method with five 2-axis accelerometers and classified 20 activities with a C4.5 decision tree at 70 – 90% accuracy [4].

Because cameras have become low-cost and small enough that the user can always wear them (*e.g.,* SenseCam [15]), computer vision can also be used to determine user activity. For example, Sundaram and Cuevas developed an activity recognition method that uses low-resolution images recorded by a wearable camera [30]. Their system recognized nine hand-related manipulations (*e.g.*, open the door, wipe, and eat/drink) at approximately 60% accuracy.

Philipose *et al.* employed radio frequency identification (RFID) tags to determine which objects the user interacts with, and used this information to infer the activity [27]. They developed a glove with an embedded RFID reader, and assumed that all objects of interest in a house are instrumented with RFID tags. Their system can detect when and which objects the user interacts with. By using dynamic Bayesian networks on the sequence of contact events, they were able to recognize 14 activities (*e.g.,* washing the hands, and preparing food or a drink) at 84% accuracy. Although this work demonstrated that the objects with which the user is interacting offers information useful for recognizing user activities, the installation of RFID tags to all objects around the user is not always feasible (*e.g.,* outside the home and in environments not under the control of the user).

Muscle movements can be used to infer different activities. Cheng *et al.* [8] developed a capacitive sensor that can be worn around the user's neck. This sensor can classify dietary activities and breathing as well as head movements. Our exploration can complement the prior work through investigating how useful sounds recorded from the user's throat area can be for gesture recognition.

Acoustic Sensors for Activity Recognition

In addition to the existing wearable sensors described above, acoustic sensors can also be used to detect user activities by recording and processing sound waves. Although various types of acoustic sensors applications have been built, we focus on technologies and systems using microphones to collect human-audible sound (20 – 20000 Hz).

Most related to activity recognition is computational auditory scene recognition (CASR) [26], which differs from computational auditory scene analysis (CASA) [6]. CASR aims to infer the context from the observed sound while CASA aims to extract the sound of interest. Clarkson *et al.* conducted a preliminary study on detecting scene changes based on the sound recorded by the microphone attached to the user's shoulder [10]. Their system could detect most scene changes, but generated many false positives. Peltonen *et al*. collected sound data from 17 scenes (*e.g.,* streets, parks, restaurants, and offices) for CASR [26], and achieved 63.4% classification accuracy with Gaussian Mixture Model.

CASR primarily examines the ambient sounds generated by different environments to infer contexts. But sounds produced by the user can be useful in inferring her activities. Chen *et al.* examined how accurately activities in a bathroom (*e.g.*, taking a shower, hand washing, and brushing teeth) could be recognized through the sound recorded by a microphone installed in the room [7]. They found that the accuracy for recognizing six activities was 84%. Lu *et al.* developed an activity recognition system using microphones embedded in mobile phones [19]. Two distinguishing features of their system are the use of multi-level classifiers for coarse and fine classification, and the integration of an unsupervised learning method to include newly found and user-specific sounds. The accuracy of the coarse classification for three kinds of sounds (ambient sound, music, and speech) was 92.3 %, and the average F-measure of the four events tested with their unsupervised classifier was 72.4 %. This system has also been extended to sensing systems using mobile phones by combining with other sensor data (*e.g.*, accelerometers and GPS) [20, 21].

Acoustic sensors can also be used to detect dietary activities. Amit *et al.* investigated the chewing sounds of different food with an inner-ear microphone [1]. Their study found that the system could distinguish chewing sounds with four kinds of food (chips, apples, pasta, and lettuce) at ~85% accuracy. Amit and Troster used EMG sensors and a microphone to further classify the human swallowing activities [2]. They found that their sensors could detect

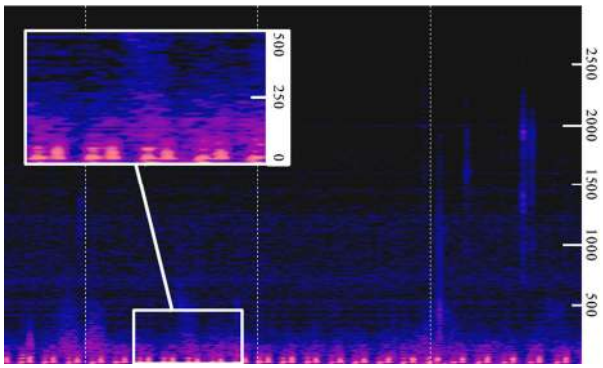

Figure 2. A sound spectrogram when the user is seated. The strongest signal from a heart sound is below 70 Hz. In this and later spectrograms, the interval of the two vertical dotted lines is 5 seconds, the vertical axis represents the frequency ranging from 0 to 3kHz, and the brighter color represents higher intensity.

swallowing events, and differentiate two levels of volumes and viscosities. They also integrated multiple sensors to recognize dietary activities in a more holistic manner [3]. We extend their work to explore how accurately a wider variety of activities can be recognized with a single acoustic sensor attached to the user's neck.

BODYSCOPE: A WEARABLE SOUND SENSOR

BodyScope is a wearable sensing system that records and classifies sounds produced in the user's mouth and throat by different activities. Our sensor prototype includes a modified Rocketfish Bluetooth headset with a uni-directional condenser microphone embedded into one of its earpieces so that it is pointed inwards the user's body (see Figure 1). We covered the microphone with a windscreen, and attached the chestpiece of a stethoscope over the windscreen. This has previously been demonstrated to be effective for recording sounds around the neck [3, 23].

The chestpiece is designed to be positioned on the side of the neck to amplify the sounds produced inside the throat and minimize audio from external sources. Through pilot studies, we observed that placement of the chestpiece at or near the larynx interfered with some activities, such as speaking and drinking. Thus, we designed the BodyScope device to be worn on the side of the neck, near the carotid artery. The device sends the sound to a computer or mobile phone via Bluetooth.

USER ACTIVITIES DETECTED THROUGH SOUND

Sitting and Deep Breath

A weak yet noticeable signal appears periodically even when the user is simply seated still (Figure 2). This sound likely results from blood pumping through the carotid artery because each cycle consists of two parts, which correspond to the sounds caused by the opening and closing of atrioventricular valves [14]. Although these sounds could be useful information for monitoring health status, the signal is weak and the isolation of the vascular sound seemingly becomes difficult when the sensor is recording

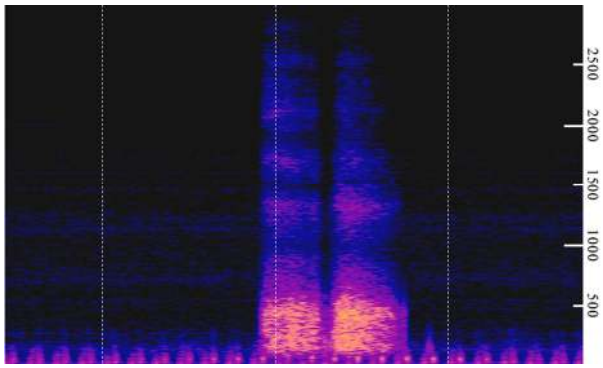

Figure 3. A sound spectrogram when the user is taking a deep breath. The highest signal from a deep breath appears from 100 to 500 Hz.

sounds caused by other activities. We do not analyze the heart sound further in this work.

Often, changes from a typical breath to a deep breath may indicate that the user is engaging in a physical activity. When the user takes a deep breath, a large volume of air goes through the throat. It produces a stronger signal than a normal breath, which distinguishes it from normal breathing. Figure 3 shows that there are two large changes in the signal. The first and second changes in the signal are produced by inhaling and exhaling air, respectively.

Eating

Detection of food intake may enable close monitoring of the user's dietary behavior. Figure 4 shows the spectrograms of when the user is eating a crunchy cookie (left) and a piece of soft bread (right). Regardless of the materials, the chewing sound appears approximately every 700 milliseconds. This is in line with the findings in the field of dental research [22], showing that the average frequency of chewing is about 1.25 Hz. The swallowing sounds also appear as relatively larger sounds at the end of the sequence.

When eating soft bread, the chewing sound is slightly weaker than when eating a cookie. This indicates that we might be able to distinguish whether the user is eating crunchy or soft food, in addition to recognizing her eating activity. Future work on this problem would involve using BodyScope to further determine the specific type of food being consumed.

Drinking

Related to eating, fluid intake is another important dietary and health-related activity. BodyScope detects the gulping sound of drinking, and can differentiate this sound from that of eating and swallowing. Figure 5 shows the spectrograms of the sound generated from drinking a regular beverage (cold water) and a warm beverage (hot tea). When the user takes a gulp, a fairly loud sound is caused. As seen in Figure 5, the gulping sound reaches up to 1500 Hz. We also observed that the gulping sound is generally stronger than the swallowing sound.

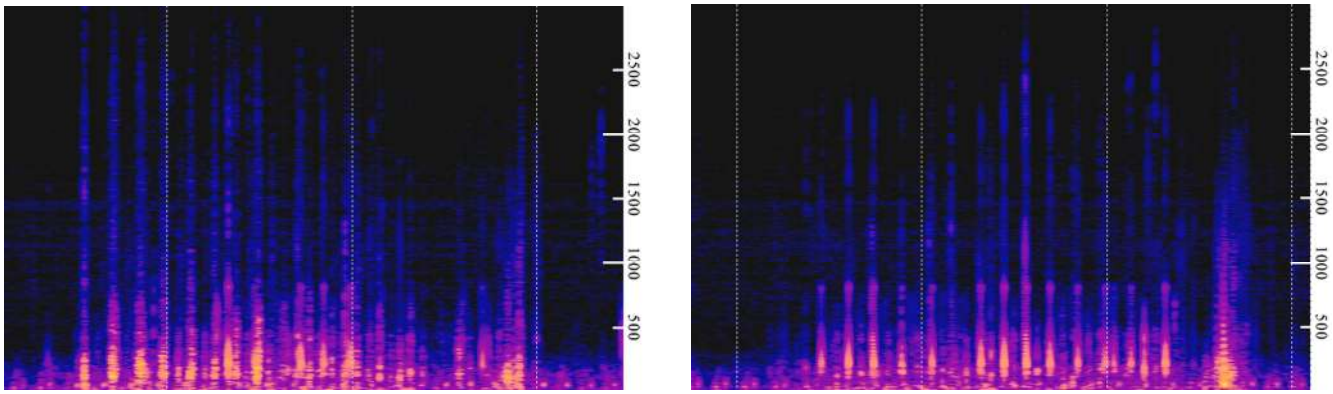

Figure 4. Sound spectrograms when the user is eating food (a cookie in the left figure and a piece of bread in the right). The chewing sound appears approximately every 700 milliseconds. When the user is eating a soft material (bread in this case), the power of the sound becomes slightly weaker. The swallowing sound also appears at the end of the sequence.

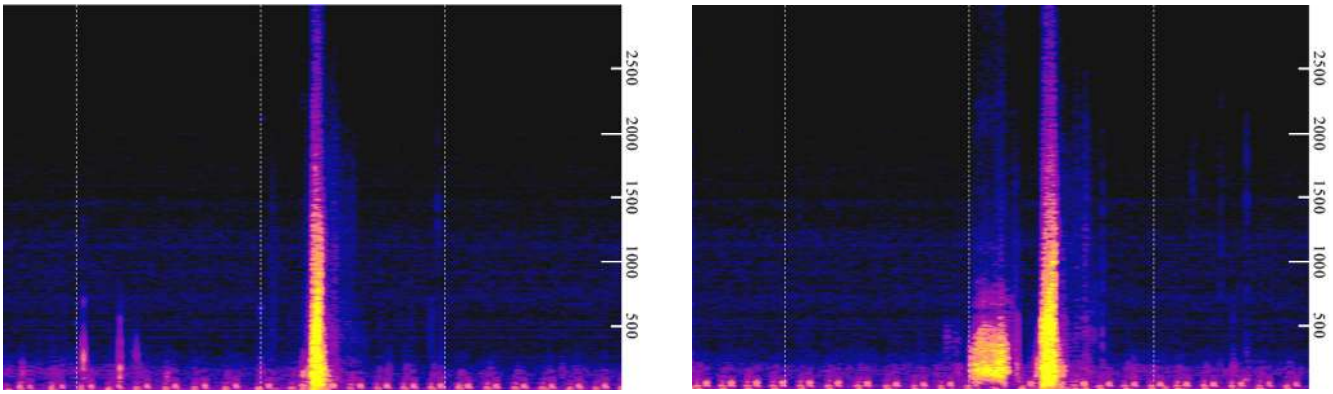

Figure 5. Sound spectrograms when the user is drinking (cold water in the left figure and hot tea in the right). Both spectrograms shows the moment of the gulp, but the sound of the sip also appears when the user is drinking hot tea. The gulping sound reaches up to 1500 Hz.

BodyScope can also sense the sound of sipping, for example, preceding a gulp of hot tea (see Figure 5). The presence of sipping may be a good indicator of whether the user is drinking cold or hot beverages. However, the user also could sip when drinking cold beverages or when drinking out of a different container. Thus, we focused only on examining whether drinking with and without a sip are distinguishable rather than detecting whether the user is drinking cold or warm beverages.

Speaking and Whispering

Identification of speaking and whispering is the first step towards determining the volume, tone, and connotations of speech. Figure 6 shows spectrograms of the sound when the user is speaking normally and whispering. With speaking, we can see clear harmonics in the sound, which differentiate the human voice from the sounds caused by eating and drinking. On the other hand, whispering sounds are generated mainly by a turbulent noise in and above the larynx, but not vibration of the vocal folds [32]. Thus, the sound intensity decreases and the harmonics disappear in whispering. Nakajima *et al.* attempted to recognize words when the user speaks and whispers using wearable microphones [23]. Our work differs in that we are primarily interested in being able to automatically separate whispers from normal speech for systems like Nakajima *et al.*'s work, and differentiate these two vocal activities from other activities, such as eating and drinking.

Whistling

Whistling can be an indicator of mood, attempts to get someone's attention, or simply a subconscious habit. Figure 7 shows a spectrogram when the user is whistling outwards. We clearly can see the melody of the whistle between 500 Hz and 1500 Hz. In addition to the melody, BodyScope can sense the flow of air between 100 Hz and 500 Hz. We observed a similar spectrogram for inward whistling.

Laughing, Sighing, and Coughing

Laughing, sighing and coughing are important activities that can provide clues for inferring mental and physical wellbeing. Audio-based detection of laughing [28, 31], coughing [17], and sighing [7] has been explored separately in the past. Here we focus primarily on how accurately BodyScope can distinguish these activities.

Because laughing aloud is a form of vocalization, the intensity is noticeably high (see Figure 8, left). Although BodyScope can collect the sound of a chuckle, its signal is significantly weaker than the signal for laughing aloud. Furthermore, the power distribution of the chuckling sound often becomes very similar to the sound signal of a deep breath.

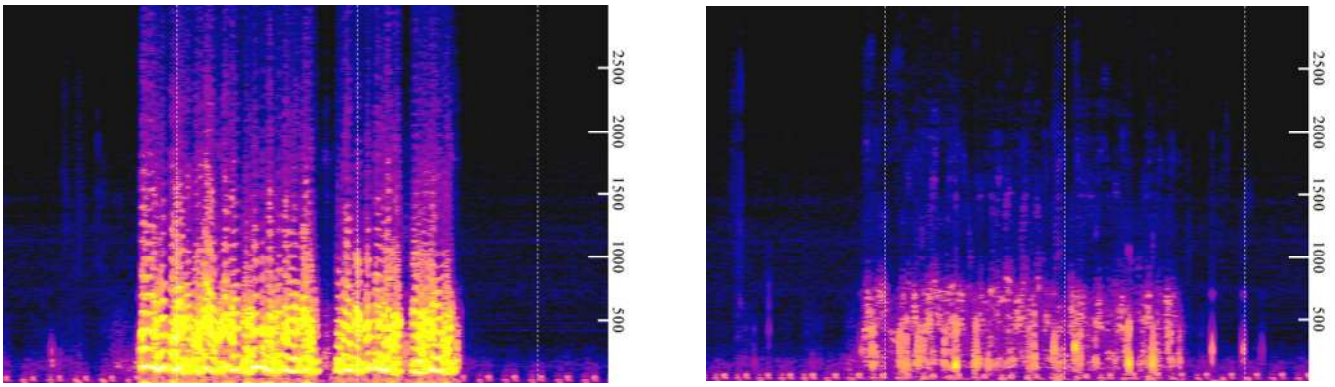

Figure 6. Sound spectrograms when the user is speaking normally and whispering the same short phrases (left: speaking, right: whispering). Harmonics which characterize the human voice are shown in speaking while the intensity decreases and the harmonics disappear in whispering.

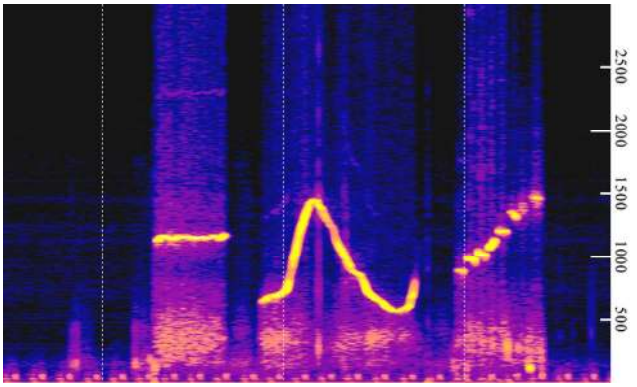

Figure 7. A spectrogram of the whistling sound. A similar spectrogram is observed when the user is whistling inwards. While the melody appears between 500 and 1500 Hz, the flow of the air appears between 100 and 500 Hz.

Sighing and coughing are particularly challenging to recognize accurately. The sound pattern for sighing (Figure 8: center) is somewhat similar to the sound of drinking hot beverages (Figure 5: right). But the first part of sighing (inhaling the air) is shorter than sipping when drinking hot beverages. A cough causes a large sound (see Figure 8: right) with a spectrum similar to that of drinking cold beverages (Figure 5: left), laughing, and sighing.

CLASSIFICATION TECHNIQUE

Sample Length, Sampling Rate and Frame Size

The length of sounds we are interested in varies from less than 1 second (*e.g.*, gulping or coughing) to over 10 seconds (*e.g.*, eating or speaking). We set the length of a sample sound to 5 seconds to capture the most important characteristics of the sound.

The sampling rate must be sufficiently high so that we do not lose important characteristics of the sound. Our observations of different sounds revealed that the power of a signal is distributed mostly below 5000 Hz. Thus, we set the sampling rate to 22050 Hz, meaning our system covers up to 11025 Hz (refer to as ω_0).

We defined the frame size for the Fast Fourier Transform (FFT) operation to be 4096 samples, 186 milliseconds long, without any overlap. We did not include the last 164 milliseconds of the sample in calculations. Each frame was Hamming-windowed by $\omega = 0.54 - 0.46 * \cos(2\pi n/4096)$.

Features

Based on the previous work we reviewed, we decided to use the following three domain features for our machine-learning classification: time, frequency and cepstral. Due to the space limit, we omit their mathematical definitions, but note that they are available in [13, 19, 29].

Time-domain Feature

We used the zero-crossing rate (ZCR) as a time-domain feature. ZCR is the rate of sign-changes along a signal. ZCR has been used for audio-based systems, such as speech recognition and audio classification. For example, it is used as a feature for differentiating voiced and unvoiced sounds (*i.e.*, voiced and unvoiced sounds tend to have low and high ZCRs, respectively) [19].

Frequency-domain Features

We used the following five frequency-domain features. To calculate these features and cepstral features (below), the sound data was pre-processed with FFT. We calculate the average and standard deviation across all the frames for each feature and use them in our classification.

- *Total Spectrum Power*: We used the logarithm of the summed spectrum power [13].
- *Subband Powers*: This represents the summed power signal in logarithmetically-divided bands [13]. We set five sub-bands for our classification: $[0, \omega_0/16]$, $[\omega_0/16, \omega_0/8]$, $[\omega_0/8, \omega_0/4]$, $[\omega_0/4, \omega_0/2]$, and $[\omega_0/2, \omega_0]$.
- *Brightness*: The brightness corresponds to the frequency centroid and represents the balancing point of the spectrum [19].
- *Spectral Rolloff*: The spectral rolloff represents the skewness of the spectral distribution. We use a 93% threshold similarly to [19].
- *Spectral Flux*: The spectral flux is defined as the L2-norm of the spectral amplitude difference of two adjacent frames [19]. This represents how drastically the sound is changing between frames.

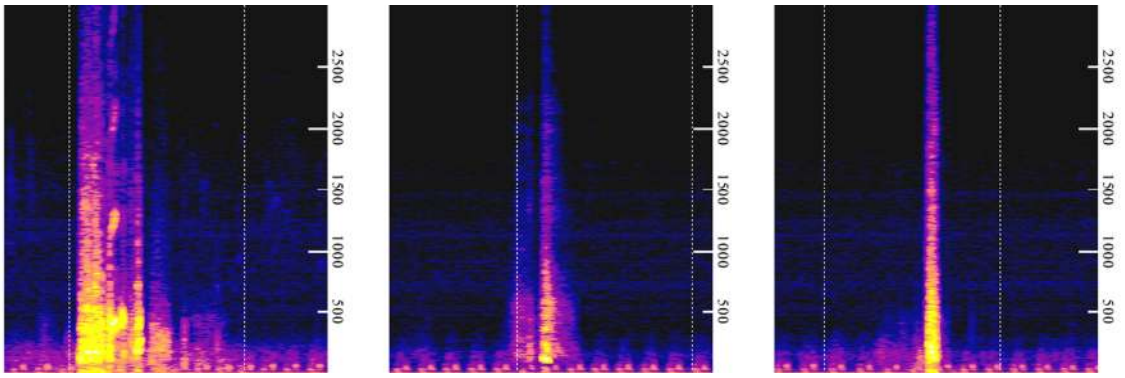

Figure 8. Spectrograms of other non-verbal sounds (left: laughing; center: sighing; right: coughing).

Cepstral Features

Mel-frequency cepstral coefficients (MFCCs) are commonly used for audio and speech recognition [29]. They were extracted by applying the discrete cosine transform to the log-scaled outputs of the FFT coefficients filtered by a triangular band-pass filter bank. We used 20 filters and calculated the first 12 coefficients for our classification.

Algorithms

We used Support Vector Machine (SVM) for our classification. SVM is a well-known machine learning technique that is used in a variety of applications, including audio classification [13]. We used Libsvm [9] to implement our classifier. One problem with SVM is that it can only do a binary classification. Thus, when there are more than two classes to classify, the classification must be divided into multiple binary classifications. We used the "one-against-one" strategy for our classification, which is already implemented in Libsvm. This strategy constructs $k(k\text{-}1)/2$ classifiers for k classes and conducts pairwise comparisons. Each pairwise comparison votes for the predicted class, and the class with the most votes is predicted after all pairwise comparisons are completed. The mathematical details of a SVM and different strategies for a multi-class classification are available at [5, 16].

We used the Radial Basic Function (RBF) as a kernel function for SVM. The penalty parameter and scale parameter for RBF were determined through a grid search using 5-fold cross validations. They were fixed during the classification test (explained later).

We also used Naïve Bayes and 5-nearest neighbor (5-NN) techniques for the comparison. These techniques are less computationally expensive than SVM [5]. Thus, these methods may be used when the system needs to recognize user activities in real time. We implemented these classifiers using Statistical Pattern Recognition Toolbox for Matlab [33]. We used a Gaussian Mixture Model for our Naïve Bayes classifier to accommodate the continuous input. We used Euclidean distance without a weight for the nearest neighbor algorithm.

Hidden Markov Model (HMM) is another method commonly used for audio-based pattern classification, like speech recognition, because it can accommodate sequential data [5]. By aggregating each feature across time, we can use SVM, Naïve Bayes, and 5-NN to handle the sequential data.

LABORATORY EVALUATION

We conducted a laboratory study to examine the accuracy of activity recognition with BodyScope. Ten participants (9 male and 1 female, all in their 20s or 30s) used the same prototype (Figure 1). All participants were in good health at the time of the experiment, and had no history of major diseases in their circulatory, respiratory, and gastrointestinal system. They were also all able to whistle.

Data Gathering Procedure

During the experiment, the participants were asked to wear the BodyScope sensor around the neck as shown in Figure 1. The experimenter asked the participants to perform one of the twelve activities discussed above and recorded the sound from the BodyScope sensor. We collected ten samples per activity per participant. No data processing on the recorded sounds was done during the experiment.

For each of the following twelve activities, participants were instructed as follows:

- *Seated*: The participants were asked to sit comfortably and breathe normally. Additionally, they were requested not to intentionally produce any other kind of sound.
- *Deep breath*: The participants were asked to take a deep breath. The samples gathered contain sounds caused by both an inhale and exhale.
- *Eating cookies*: The participants were asked to eat crunchy cookies. Because we did not intend to distinguish between chewing and swallowing in this project, some of the gathered samples included the participant's bite into the cookie and swallowing.
- *Eating bread*: The participants were asked to eat a piece of soft bread. Again, some samples included the sound of biting into the bread and swallowing, similar to *Eating cookies*. The two eating classes were intended to represent the cases in which a user is eating crunchy and soft food.

- *Drinking*: The participants were asked to take a gulp of room-temperature water. The gathered sample each contained one gulping sound.
- *Drinking with a sip*: The participants were asked to take a sip and gulp of hot tea. We controlled the temperature of the tea so that the participants could drink it in this manner comfortably. All the samples contained both the sipping and gulping sounds. For both *Drinking* and *Drinking with a sip*, the participants used the same mugs.
- *Speaking*: The participants were asked to read aloud sentences from reading materials we prepared. We gathered samples while they were reading. These samples may contain short pauses and longer stops.
- *Whispering:* The participants were asked to whisper sentences from the same reading materials used in *Speaking*, and we sampled the sound. Similar to *Speaking*, some samples contained short pauses and longer stops.
- *Whistling*: The participants were asked to continuously whistle. Because the whistling skills of the participants varied, the participants were allowed to whistle melodies of their choice. As a result, we gathered whistling sound samples ranging from a single note to pop music tunes.
- *Laughing:* We prepared a 15-minute long comedy clip and asked the participants to watch it as they normally would. We offered the participants a headphone to listen to the audio from the clip. We recorded all the sounds while the participants watched the clip. All participants found the clip funny and laughed aloud to it. After the experiment, a member of the research team manually extracted parts of the recorded sound in which the participants were laughing aloud.
- *Sighing*: The participants were asked to sigh. Each sampled sound contained one sigh.
- *Coughing*: The participants were asked to cough. Each sampled sound contained one cough.

The study took about an hour. None of the collected sound data overlapped with each other, and all were stored as WAV files. In total, we collected 1200 samples.

Classification Procedure

After gathering the sound data, we conducted a classification test using the three classifiers. The system calculated the features mentioned in the previous section for each sound sample, and normalized the values of all the features across all samples.

For training and testing, we decided to use two following protocols:

- *Leave-one-participant-out* cross validation: We used the data gathered from nine of the participants for training, and used the data from the other participant for testing. This was repeated such that each participant's data were used once as the validation data set.

	Leave-one-participant-out			*Leave-one-sample-per-participant-out*		
	PR	**RE**	**F**	**PR**	**RE**	**F**
Bayes	47.0%	45.7%	46.3%	72.3%	71.2%	72.2%
5-NN	43.5%	43.2%	43.3%	75.3%	75.1%	75.2%
SVM	50.2%	49.1%	49.6%	79.6%	79.4%	79.5%

Table 1. The classification accuracies (the average values across the classes) with three machine learning techniques (PR: Precision; RE: Recall; F: F-measure).

- *Leave-one-sample-per-participant-out* cross validation: We reserved one sample for one class from each participant as a test case, and used the rest of the samples for training. Similarly, this was repeated such that each sample from each participant's data set was used once as the validation data set.

The primary difference between these two strategies is whether the training data contains data gathered from the participant to be tested. In this sense, *Leave-one-participant-out* trains the classifier in a user-independent manner while the other considers user-dependency. We believe that the comparison of these two protocols will illustrate what level of user-dependency in the data we have collected. We calculated the classification accuracy for each round of the cross validations.

We also considered training the classifier for each participant. However, the size of the samples which could be used for training was limited (nine samples per class). Thus, we did not execute this training protocol in this study.

RESULTS OF THE LABORATORY EVALUATION

We calculated the overall precision, recall, and F-measure (the harmonic mean of the precision and recall). In all techniques, the F-measure with the *Leave-one-sample-per-participant-out* protocol was about 25 – 30% higher than one with the *Leave-one-participant-out* protocol (see Table 1). SVM outperformed the other two techniques in our classification. Because we found that the error distributions were similar across the three techniques, we focus on the results from the SVM classification.

The F-measure with the *Leave-one-participant-out* protocol was 49.6%. With the *Leave-one-sample-per-participant-out* protocol, the F-measure reached 79.5%. This result indicates that the classifier should be trained for each user if the samples for training are abundant.

Table 2 and 3 show the confusion matrices under the two protocols. There are a few noticeable differences between them beyond the overall accuracy. For instance, with *Leave-one-sample-per-participant-out* protocol, the classification accuracies for *Deep breath* and *Sighing* improved greatly. *Whispering* is another class whose accuracy was improved largely. These activities may be more user-dependent than other activities. We will discuss this in the next section.

Actual Activities \ Prediction	Seated	Deep breath	Eating (Cookie)	Eating (Bread)	Drinking	Drinking (with a sip)	Speaking	Whispering	Whistling	Laughing	Sighing	Coughing	Recall [%]
Seated	61	9	1	2	20	4	0	0	0	0	3	0	**61.0**
Deep breath	2	15	9	7	7	2	0	16	4	15	21	2	**15.0**
Eating (Cookie)	0	2	56	20	2	4	0	9	0	4	2	1	**56.0**
Eating (Bread)	2	4	27	51	5	2	0	0	1	3	1	4	**51.0**
Drinking	8	8	4	3	35	16	0	0	1	2	20	3	**35.0**
Drinking (with a sip)	3	10	17	10	33	9	0	3	0	6	6	3	**9.0**
Speaking	0	0	0	0	0	0	90	4	0	3	0	3	**90.0**
Whispering	0	11	5	0	2	1	20	53	2	3	0	3	**53.0**
Whistling	0	1	0	0	0	2	0	1	96	0	0	0	**96.0**
Laughing	1	14	4	1	6	4	8	7	1	46	3	5	**46.0**
Sighing	7	21	5	11	10	0	1	0	0	8	28	5	**28.0**
Coughing	4	3	2	3	5	3	2	0	1	11	4	62	**62.0**
Precision [%]	**69.3**	**15.3**	**43.1**	**50.5**	**27.8**	**15.8**	**75.0**	**56.4**	**90.6**	**45.5**	**31.8**	**68.1**	

Table 2. The confusion matrix of the classification with the *Leave-one-participant-out* protocol.

Actual Activities \ Prediction	Seated	Deep breath	Eating (Cookie)	Eating (Bread)	Drinking	Drinking (with a sip)	Speaking	Whispering	Whistling	Laughing	Sighing	Coughing	Recall [%]
Seated	94	0	0	0	4	1	0	0	0	1	0	0	**94.0**
Deep breath	0	79	0	2	3	2	0	5	0	4	2	3	**79.0**
Eating (Cookie)	0	1	81	7	3	6	0	1	0	1	0	0	**81.0**
Eating (Bread)	0	1	8	80	4	5	0	0	0	0	1	1	**80.0**
Drinking	0	5	3	1	78	5	0	1	0	2	5	0	**78.0**
Drinking (with a sip)	2	2	10	5	14	60	0	2	0	1	2	2	**60.0**
Speaking	0	0	0	0	0	0	97	0	0	2	0	1	**97.0**
Whispering	0	6	2	0	0	0	4	82	0	5	0	1	**82.0**
Whistling	0	1	0	0	0	0	0	1	98	0	0	0	**98.0**
Laughing	0	7	3	0	5	4	4	6	0	64	2	5	**64.0**
Sighing	2	10	0	4	6	1	0	1	0	6	66	4	**66.0**
Coughing	4	4	1	1	1	2	0	1	0	4	8	74	**74.0**
Precision [%]	**92.2**	**68.1**	**75.0**	**80.0**	**66.1**	**69.8**	**92.4**	**82.0**	**100**	**71.1**	**76.7**	**81.3**	

Table 3. The confusion matrix of the classification with the *Leave-one-sample-per-participant-out* protocol. The accuracy was improved by approximately 30 % over the *Leave-one-participant-out* protocol.

LABORATORY STUDY DISSCUSSION

The classification accuracies with the two protocols were significantly better than that of the random classification. This indicates that a single wearable acoustic sensor can reasonably recognize various activities by using the sounds recorded from the user's throat area.

The classification results with the *Leave-one-sample-per-participant-out* protocol were approximately 30% better than ones with the *Leave-one-participant-out* protocol. During the study, the sounds produced by the activities we are interested in often varied by participant. For example, whispering by one of our participants was similar to speaking with a low volume in another, and we observed harmonics in the spectrograms for the whispering sound samples. This implies that the classifier for BodyScope performs better when it is trained for each user. BodyScope would be used by an individual in many applications, and personalizing the classifier is plausible.

Distinguishing the two eating and two drinking classes from each other was not as accurate as we expected. We observed that some participants produced only weak sounds when sipping and gulping because they took only a small amount of tea. We also noticed that some participants produced fairly strong sounds even when eating soft bread. Thus, accuracies for these classes were not high under the *Leave-one-participant-out* protocol. However, we believe that personalizing the classifier could mitigate this problem as the accuracies were improved under the *Leave-one-sample-per-participant-out* protocol.

Decreasing the activity granularity would help to increase the classification accuracy. For instance, we can combine *Eating cookies* and *Eating bread*, and *Drinking* and *Drinking with a sip* as *Eating* and *Drinking*, respectively. This would result in 87.5% and 78.5% F-measure for *Eating* and *Drinking* under the *Leave-one-sample-per-participant-out* protocol. Future applications using the

		Prediction				Recall [%]
		Eating	Drinking	Speaking	Laughing	
Actual Activities	Eating	157	11	11	0	**87.8**
	Drinking	19	33	7	9	**56.0**
	Speaking	16	10	498	7	**93.8**
	Laughing	1	0	25	14	**35.0**
Precision [%]		**81.3**	**61.1**	**92.1**	**66.7**	

Table 4. The confusion matrix of the SVM classification in our small-scale in-the-wile study.

		Prediction				Recall [%]
		Eating	Drinking	Speaking	Laughing	
Actual Activities	Eating	125	49	4	1	**69.8**
	Drinking	17	36	6	0	**61.0**
	Speaking	58	40	352	81	**66.3**
	Laughing	1	1	16	22	**55.0**
Precision [%]		**62.2**	**28.6**	**93.1**	**21.2**	

Table 5. The confusion matrix of the Naïve Bayes classification in our small-scale in-the-wile study.

BodyScope need to consider what level of granularity is necessary so that BodyScope can perform at its best.

The distribution of the classification confusions is also interesting to discuss. The confusions are minimized with the *Leave-one-sample-per-participant-out* protocol, but the patterns of confusions between the two protocols are similar. As we expected, the two eating classes and two drinking classes were confused with each other in both protocols. *Laughing* and *Sighing* were confused for most of the other activities, whereas *Seated* was mainly confused for *Drinking*. The three non-verbal activities seem to be hard to distinguish very accurately, but our system could still distinguish them at 72.0% F-measure accuracy under the user-dependent training.

SMALL-SCALE IN-THE-WILD STUDY

We also conducted a small-scale in-the-wild study to examine how well BodyScope can detect human activities in a realistic setting. We recruited another 5 participants (3 male and 2 female) for this study. We asked them to wear the BodyScope sensor for as much of the day and any portions of it in which they felt comfortable to do so. The sound from the sensor was recorded onto a mobile phone which the experimenters also gave the participants. In order to know the ground truth, we also asked the participants to wear another mobile phone around the neck. The camera of this phone faced away from the user's body to record the user's context as SenseCam does [15]. We focused on four activities (eating, drinking, speaking, and laughing) in this study because they were observed frequently and are representative of the classes we covered in the laboratory study (eating and drinking as a food-consumption activity, and speaking and laughing as a social activity). The participants were told to behave while wearing the sensor as they would normally do. We collected the data of 64 minutes long on average.

After the data collection, we analyzed and labeled the recorded sound. We split the sound data into 5-second WAV files, and marked the files which we considered that were associated with one of the activities based on the audio files and pictures recorded in the mobile phone. We then extracted the features, and trained the classifiers for each participant as we did in the laboratory study.

Table 4 and 5 show the results of the classification using SVM and Naïve Bayes. The overall F-measure was 71.5% and 56.5% with the SVM and Naïve Bayes classifier, respectively. Again, SVM outperformed Naïve Bayes, but the results show several differences in classification errors between the two techniques. Particularly, Naïve Bayes showed a better recall rate for laughing activities. But generally, the accuracy for the drinking and laughing activities was relatively low. As seen in Table 4 and 5, the number of the instances for these two activities was small; thus, more data may be necessary to examine the true classification accuracy for these activities.

We also note that the eating activities were identified fairly accurately. This indicates that BodyScope can enhance existing systems related to food consumption activities. For example, Noronha *et al*. developed Platemate, which allows the user to analyze her food consumption by taking a picture of the food with a mobile phone and getting food annotations through a crowdsourcing system [24]. With BodyScope, a wearable camera (like SenseCam [15]) automatically can identify the moments when the user is eating, and perform analysis on her food consumption through Platemate. In summary, this small-scale in-the-wild study shows that BodyScope can classify some of the activities we explored in this work even in a realistic setting with high accuracy, and demonstrates its potential to enhance lifelogging systems.

CONCLUSIONS AND FUTURE WORK

We presented BodyScope, an acoustic sensor wearable around the neck designed to recognize a variety of user activities. BodyScope is able to detect different sounds generated from different user activities, such as speaking, eating, drinking, and laughing. Our laboratory study shows that BodyScope classified the user's twelve activities at 79.5% F-measure accuracy. We also conducted a small-scale in-the-wild study, and found that BodyScope was able to identify four activities (eating, drinking, speaking, and laughing) at 71.5% F-measure accuracy.

We believe the BodyScope is able to sense additional activities (*e.g.*, smoking and sneezing) that were not examined in this paper. We will explore the use of unsupervised learning methods like Lu *et al.*'s work [19] to automatically find such activities. We also plan to revise the BodyScope hardware. We will investigate how the

BodyScope could be integrated into accessories worn around the user's neck.

REFERENCES

1. Amit, O., Stager, M., Lukowicz, P., Troster, G. Analysis of chewing sounds for dietary monitoring. In *Proc. Ubicomp 2005*, Springer (2005), 56-72.
2. Amit, O., Troster, G. Methods for detection and classification of normal swallowing from muscle activation and sound, In *Proc. PHC 2006*, IEEE (2006), 1-10.
3. Amit, O., Troster, G. On-body sensing solutions for automatic dietary monitoring. *IEEE Pervasive Computing 8*, 2 (2009), 62-70.
4. Bao L., Intille S. S. Activity recognition from user-annotated acceleration data. In *Proc. Pervasive 2004*, Springer (2004), 1-17.
5. Bishop, C. M. *Pattern recognition and machine learning*. Springer Science+Business Media, LLC, New York, NY, USA, 2006.
6. Brown G. J., Cooke M. Computational auditory scene analysis. *Computer speech and language 8* (1994), 297-336.
7. Chen, J., Kam, A., Zhang, J., Liu, N., Shue, L. Bathroom activity monitoring based on sound. In *Proc. Pervasive 2005*, Springer (2005), 47-61.
8. Cheng, J., Amft, O., Lukowicz, P. Active capacitive sensing: exploring a new wearable sensing modality for activity recognition. In *Proc. Pervasive 2010*, Springer (2010), 319-336.
9. Chung, C-C., Lin, C-J. Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/livsvm
10. Clarkson, B., Sawhney, N., Pentland A. Auditory context awareness via wearable computing. In *Proc. PUI 1998*.
11. Farringdon, J., Moore, A. J., Tilbury, N., Church, J., Biemond, P. D.: Wearable sensor badge and sensor jacket for context awareness. In *Proc. ISWC 1999*, IEEE (1999) 107-113.
12. Foerster, F., Smeja, M., Fahrenberg J. Detection of posture and motion by accelerometry: a validation in ambulatory monitoring. *Computers in Human Behavior, 15*, 5 (1999), 571-583.
13. Guo, G., Li, S. Z. Content-based audio classification and retrieval by support vector machine. *IEEE Transactions on Neural Networks 14*, 1 (2003), 209-215.
14. Guyton, A. C., Hall, J. E. *Textbook of medical physiology*. Elsevier Inc., Philadelphia, PA, USA, 2006.
15. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K. SenseCam: a retrospective memory aid. In *Proc. Ubicomp 2006*, Springer (2006), 177-193.
16. Hsu C-W., Lin, C-J. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks 13* (2002), 415-425.
17. Larson, E. C., Lee, T., Liu, S., Rosenfeld, M., Patel, S. N. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proc. UbiComp 2011*, ACM (2011), 375-384.
18. Liao, L., Fox, D., Kautz, H. Location-based activity recognition. In *Proc. NIPS 2005*, MIT Press (2005), 787-794.
19. Lu, H., Pan, W., Lane, N. D., Choudhury, T., Campbell, A. T. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proc. MobiSys 2009*, ACM (2009), 165-178.
20. Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., Campbell, A. T. The Jigsaw continuous sensing engine for mobile phone applications. In *Proc. SenSys 2010*, ACM (2010), 71-84.
21. Milizzo, E., Papandrea, M., Lane, N. D., Lu, H., Campbell, A. T. Pocket, bag, hand, etc. - automatically detecting phone context through discovery. In *Proc. PhoneSense 2010*, 21-25.
22. Morimoto, T., Inoue, T., Nakamura, T., Kawamura, Y. Frequency-dependent modulation of rhythmic human jaw movements. *Journal of Dental Research 63*, 11 (1984), 1310-1314.
23. Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *Proc. ICASSP 2003*, IEEE (2003), 708-711.
24. Noronha, J., Hysen, E., Zhang, H., Gajos, K. Z. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proc. UIST 2011*, ACM (2011), 1-12.
25. Patel, S. N., Kientz, J. A., Hayes, G. R., Bhat, S., Abowd, G. D. Farther than you may think: an empirical investigation of the proximity of users to their mobile phones. In *Proc. UbiComp 2006,* Springer (2006), 123-140.
26. Peltomen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T. Computational auditory scene recognition. In *Proc. ICASSP 2001*, IEEE (2001), 1941-1944.
27. Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., Hahnel, D. Inferring activities from interactions with objects. *IEEE Pervasive Computing 3*, 4 (2004), 50-57.
28. Schuller, B., Weninger, F. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization, In *Proc. ICASSP 2010*, IEEE (2010), 5054-5057.
29. Sigurdsson, S., Petersen, K. B., Lehn-Schioler, T. Mel frequency cepstral coefficients: an evaluation of robustness of mp3 encoded music. In *Proc. ISMIR 2006*.
30. Sundaram, S., Cuevas, W. W. M. High level activity recognition using low resolution wearable vision. In *Proc. CVPRW 2009*, IEEE (2009), 25-32.
31. Truong, K. P., Leeuwen, D. A. Automatic detection of laughter. In *Proc. Interspeech 2005*, 485-488.
32. Tsunoda, K., Niimi, S., Hirose, H. The roles of the posterior cricoarytenoid and thyropharyngeus muscles in whispered speech. *Folia Phoniatrica et Logopaedica 46* (1994), 139-151.
33. Statistical Pattern Recognition Toolbox for Matlab, http://cmp.felk.cvut.cz/cmp/software/stprtool/.