2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)

FamilyLog: A Mobile System for Monitoring Family Mealtime Activities

Chongguang Bi*, Guoliang Xing*, Tian Hao†, Jina Huh‡, Wei Peng*, Mengyan Ma*
*Michigan State University, †IBM Research, ‡ University of California, San Diego
bichongg@msu.edu, glxing@cse.msu.edu, thao@us.ibm.com, jinahuh@ucsd.edu, pengwei@msu.edu, mamengya@msu.edu

***Abstract*—Research has shown that family mealtime plays a critical role in establishing good relationships among family members and maintaining their physical and mental health. In particular, regularly eating dinner as a family significantly reduces prevalence of obesity. However, American families with children spend only 1 hour on family meals while three hours watching TV on an average work day. Fine-grained activity-logging is proven effective for increasing self-awareness and motivating people to modify their life styles for improved wellness. This paper presents FamilyLog – a practical system to log family mealtime activities using smartphones and smartwatches. FamilyLog automatically detects and logs details of activities during the mealtime, including occurrence and duration of meal, conversations, participants, TV viewing etc., in an unobtrusive manner. Based on the sensor data collected from real families, we carefully design robust yet lightweight signal features from a set of complex activities during the meal, including clattering sound, arm gestures of eating, human voice, TV sound, etc. Moreover, FamilyLog opportunistically fuses data from built-in sensors of multiple mobile devices available in a family through an HMM-based classifier. To evaluate the real-world performance of FamilyLog, we perform extensive experiments that consist of 77 days of sensor data from 37 subjects in 8 families with children. Our results show that FamilyLog can detect those events with high accuracy across different families and home environments.**

I. Introduction

Research has shown that the family mealtime plays a critical role in establishing good relationships among family members and maintaining their physical and mental health [5][9][7]. In addition to the implications for family health, fine-grained analysis of family mealtime enables important studies in sociology and home economy. For instance, research has showed that the amount of shared time (including conversation and eating) between spouses and between parents and children have strong links with family income, mother's employment status, ages of children, and geographic location (urban or rural) [6][13][12]. However, according to a national survey in 2014, American families with children on an average work day spend almost 3 hours watching TV accounting for more than half of the leisure and sport time, while only 1 hour for family meal [24].

It is shown that activity logging is a very effective approach to improving the self-awareness and motivating people to modify their behaviors toward a healthy lifestyle [11]. Unfortunately, to date, there has been no unobtrusive and convenient methods to log family meals and related activities. Some of the available methods for family activity monitoring rely on video-taping [8], which not only incurs considerable installation/analysis costs, but also raises privacy concerns. There has been a number of studies on activity recognition using personal wearables and smartphones [19] [33]. However, as we argue in this paper, detecting the activity of individual family members separately is insufficient for studying family communications, e.g., due to the fact that young children are usually not allowed to carry personal devices.

This paper presents FamilyLog – the first practical system to log family mealtime activities using smartphones and smartwatches. FamilyLog uses the built-in accelerometer and microphone of the smartphone/smartwatch to detect mealtime activities that are closely related to family wellness, including occurrence, duration, and participants of the family meal as well as conversations and TV viewing during the meal. By providing a detailed record of the family mealtime activities, FamilyLog empowers family members to actively engage in making positive changes to improve family wellness, e.g., preventing child obesity.

The design of FamilyLog faces several challenges such as the significant interference from various noises in the home. Moreover, uploading sensor data to the cloud is often undesirable due to the privacy concerns. To address these challenges, we carefully design several lightweight acoustic and motion features based on in-depth analysis of data sets from multiple families. Furthermore, FamilyLog employs novel HMM-based sensor fusion techniques to opportunistically leverage multiple built-in sensor modalities of mobile devices available in a family, which maximizes the spatiotemporal sensing coverage and achieves robust sensing accuracy across different homes. They can also shorten the system training period by incorporating one-time user input such as the typical time/frequency of family dinners. We have evaluated FamilyLog with extensive experiments involving 8 families with children (one or two week recording in each family) and total 251 hours of sensor data collected over 77 days. Our results show the effectiveness of FamilyLog in family activity detection (with average 88.7% precision and 93.3% recall for meal detection, and 97.8% precision and 92.8% recall for the participant identification) across different families and home environments. The long-term, fine-grained family activity history provided by FamilyLog makes it possible to analyze communication patterns/anomalies and improve family life styles.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

II. Related Work

The studies by American Academy of Pediatrics have shown that, healthy family meals are not only helpful in establishing good relationships among family members, but also critical for the proper development of children's physical and mental health [5][9][7][14]. In order to monitor family mealtime activities, several systems are designed to detect the usage of electrical appliances based on the electromagnetic interference and ambient sensors [26][15][18]. However, these systems can only detect the activities that involve substantial appliance usage. Recently, activity monitoring using mobile devices has received significant attention. Several systems are designed to detect food and drink intakes. For example, [19] presents the design of a fork with sensing abilities to help track and improve user's eating behaviors. In [33], the authors propose an approach of profiling user's gesture while eating using motion sensors on smartwatches. However, these systems are focused on tracking eating behavior of individuals, and are not suitable for detecting family mealtime activities, which may involve children without wearing any devices, and conversations among family members. Moreover, some mobile health systems are designed based on off-the-shelf smartphones to monitor human activities, such as sleep quality [16] or physical activities [28]. Several recent studies are focused on user experiences with mobile health systems such as privacy concerns [1] and sharing behaviors [27]. However, these efforts are not concerned with studying family meals or group activities.

Acoustic event recognition algorithms have been widely adopted in smartphone-based activity monitoring systems. *Auditeur* [23] is designed as a mobile-cloud service platform to allow client's smartphone to recognize various sound events such as car honks or dog barking. $SoundNet$ associates environmental sounds with words or concepts in natural languages to infer activities [22]. Recent work shows that the eating activity can also be detected by the acoustic features [34]. However, this work does not pinpoint main features for detecting family meals. It requires a large amount of data, and employs complex signal processing and machine learning methods, which raise burden of the implementation on mobile devices.

In order to detect the participants in the conversation, *Crowd++* [35] counts the number of speakers using MFCC (Mel-frequency cepstral coefficient) [29] features. Row mean vector of spectrogram [20] is a simple but effective method for speaker recognition by comparing the Euclidean distance of the energy distributional features. However, voice recognition during a family meal is more challenging due to the presence of significant noise and requires new techniques.

III. Motivation and Requirements

A national survey shows that American families with children spend only 1 hour on family meal on a typical work day [24]. Moreover, it is shown that TV viewing during the meal significantly increases the energy intake [2]. Based on the datasets we collected from 8 families, over 60% of the family meals are accompanied by concurrent TV viewing. In addition to the occurrence, duration, and frequency of family meals, the conversations during a family meal are also important as they constitute a significant portion of communications between family members during a day. Analyzing the conversation during a family meal is also important for culture studies [4]. Moreover, it is shown that, by reviewing detailed activity logs, people are motivated to modify their behaviors toward a healthy lifestyle [11][6][13][12].

There has been a number of studies on personal activity recognition using wearables and smartphones [19] [33]. However, we argue that detecting the activity of individual family members separately is insufficient. First, the existing solutions typically require the mobile device (smartphone or wearable) to be carried by the user. As a result, they cannot be applied to detect many activities of young children who are usually not allowed to carry personal devices. Second, many people do not carry smartphone or wear watch constantly at home, making it difficult to monitor one's activity continuously. Moreover, detecting each individual's behavior is often unnecessary or significantly more challenging when she/he is participating in a group activity. For instance, detecting whether a particular family member is eating based on sound is more difficult when the family is having dinner together due to the higher level of ambient noise.

FamilyLog is designed to be an unobtrusive system that helps users keep track of their family mealtime activities. It employs the built-in accelerometer and microphone of smartphones and smartwatches to detect various information and activities related to a family meal. Specifically, FamilyLog is designed to meet the following requirements: 1) Since FamilyLog needs to operate in parallel with family mealtime activities. It must to be unobtrusive to use. It should minimize the burden on the user, e.g., without requiring the users to carry extra devices, and should not interfere with the users' daily activities by any means. 2) FamilyLog needs to monitor the details of family meals, including their start/end time, participants, and possible TV viewing, in a robust fashion, i.e., across different users, smartphones, smartwatches and households. 3) Since family meals involve privacy sensitive activities such as family conversation, the privacy of the family needs to be strictly protected. For example, the system should process the collected sensor samples on the fly and only keep the results, instead of storing or transmitting any raw data, which may contain sensitive information such as contents of the conversations. The sensing algorithms we develop can accurately classify a number of important contextual features of activities such as arm gestures from wearables, eating sounds, environmental noise, conversations, etc. As a result, in the future, these algorithms can be adapted and used as building blocks to detect a wide range of family activities such as parties, family meetings, gaming etc.

IV. System Design

FamilyLog detects family meals by using the built-in sensors of mobile devices, namely microphone on smart-

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

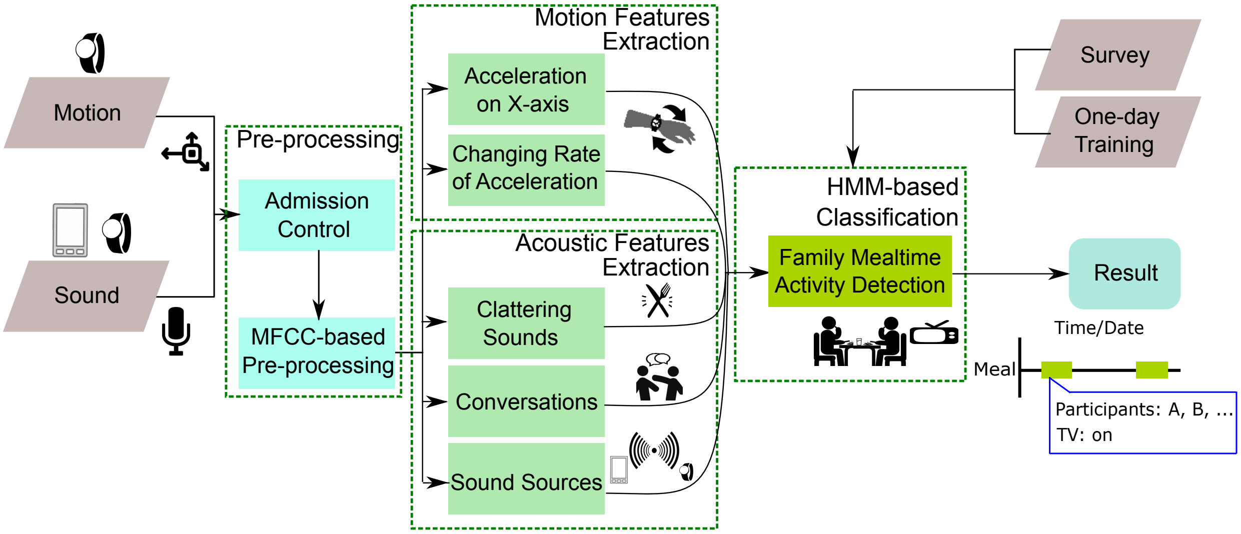

Fig. 1. System overview

phone/tablet and both microphone and accelerometer on smartwatch[1]. However, FamilyLog is designed to leverage these sensing modalities in an opportunistic manner depending on the availability of mobile devices in a home. In particular, FamilyLog may achieve satisfactory sensing performance even with a single smartphone when it is placed in the proximity of family activities (see Section V). When multiple devices are available, FamilyLog runs separately on each individual device and fuses the detection results to achieve better performance and extended coverage.

As shown in Fig.1, FamilyLog consists of four components: *pre-processing*, *acoustic feature extraction*, *motion feature extraction*, and *HMM-based activity classification*.

In pre-processing, sensors are sampled at certain rate and the samples are framed. A frame is discarded if it only contains noise which is indicated by low variance. Otherwise, each acoustic frame is processed to extract energy features using filters based on *Mel-frequency cepstrum coefficients* (MFCC). In the acoustic and motion feature extraction components, FamilyLog groups data frames ($50ms$ by default) into a detection window ($3min$ by default), and extracts a set of distinct features for each window. Specifically, FamilyLog extracts gesture-related motion features such as the average X-axis acceleration and changing rate, and acoustic features to detect the clattering sounds and the human voice.

To detect activities from extracted features, FamilyLog adopts a HMM-based (*Hidden Markov Model*) classifier. Compared with several commonly used classifiers like *Support Vector Machine* (SVM) that are only applicable to discrete event detection, HMM can naturally capture the temporal pattern of family activities by incorporating continuous sensor input. The HMM classifier is trained by a combination of short period of sensor data, e.g., a one-day family activities labeled by users, and some general knowledge of family meals which can be obtained from a one-time user input or a brief survey with simple questions such as “how much time does your weekday dinner usually take?”.

A. Pre-processing

The primary objective of pre-processing is to reduce unnecessary computation and prepare data for feature extraction. Specifically, it consists of the following three components.

First, FamilyLog reduces the unnecessary computation by discarding detection windows that likely contain only environmental noises (e.g., noise of appliances). Specifically, the noise detection is achieved by first calculating the *root mean square* (RMS) (i.e., the volume of signal) for each frame, and then computing the variance of RMS of all the frames within each window. A key observation is that a window with low RMS variance only contains ambient noise. Similarly, FamilyLog discards the motion data with low RMS variance, which typically indicates a stationary smartwatch not worn by the user.

Second, to increase computational efficiency, FamilyLog represents acoustic data with MFCC-based features, which will be used in later feature extraction. For each frame, FamilyLog first calculates its energy spectrum from $80Hz$ to $8kHz$ with the *Fast Fourier Transform* (FFT) [32]. Then the resulting spectrum is transformed into 21 energy channels by applying Mel Filters [21][25][30]. The energy of channel i will be represented as e_i hereafter.

Third, to preserve power, FamilyLog turns on sensor sampling only when the device is home, which can be determined by the system location. Moreover, as an optional feature, FamilyLog can start the the sensor sampling of a new detection window probabilistically based on the percentage of historical noise frames in a predefined time window. We note that this strategy may turn off sampling falsely when noise appears in a burst within an event of interest. In the future, we will take into account the feedback from event detection component and reduce the sensor sampling when no activity is detected.

[1] Most off-the-shelf smartwatches ship with microphone for voice control and making calls.

B. Feature Extraction

FamilyLog identifies the occurrence of the family meals by several key characteristics, based on sounds and gestures associated with dining and whether the family members are currently in close proximity to one another. Specifically, we use the following features to characterize the family meals. The first feature is the clattering sound caused by clashes between tableware. This is because the clattering sound is the most distinctive acoustic characteristic of family dining activity, regardless of other dynamics, such as the type of food and variation of tableware. The second feature is the gesture of the users captured by smartwatches. When the user is holding food or using tableware, the arm of the user often exhibits a certain pattern of movements. The third feature is the human voice, i.e. the conversation between family members, which implies that the family members are near each other.

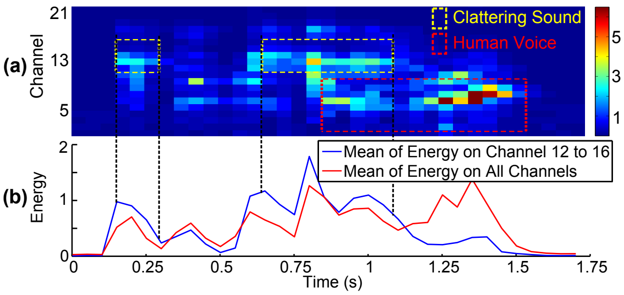

Fig. 2. An example of clattering sound detection in a typical family meal scenario. (a) shows the energy on 21 channels over time, where clattering sound and human voice are marked with rectangles. (b) shows the comparison between $\overline{e}_{12-16}$ and $\overline{e}_{all}$ for the same sound clip.

1) Clattering Sound: To infer family meal events, FamilyLog calculates the occurrences and frequency of clattering sound within a detection window. It looks for an energy peak from channel 12 to 16 (associated with frequency ranging from $1-4kHz$) for each $50ms$ frame. Specifically, for each frame, it computes $\overline{e}_{all}$, the average energy over all channels, and $\overline{e}_{12-16}$, the average energy across channel 12 to 16. The feature associated with clattering sound is calculated as $r = \overline{e}_{12-16}/\overline{e}_{all}$. For example, Fig.2 shows an example of clattering sound detection in a typical family meal scenario. Fig.2(a) shows the energy on 21 channels over time, and Fig.2(b) shows the corresponding $\overline{e}_{12-16}$ and $\overline{e}_{all}$. We can see that one occurrence of clattering sound may result in several continuous clattering frames with higher $\overline{e}_{12-16}$, even when the clattering sound and human voice overlapped around 1 second. Therefore, comparing $\overline{e}_{12-16}$ and $\overline{e}_{all}$ is a simple and effective way of detecting clattering sound in typical family meal scenarios. After obtaining r for each acoustic frame, FamilyLog calculates $E[N_{clattering}]$ which represents the expectation of amount of clattering sound contained in a detection window. Specifically, $E[N_{clattering}]$ is calculated as the sum of $P(clattering|r)$ which is preset in the system and generated using the data collected from 5 families. Fig.3 shows an example of clattering sound detection based on the real data set collected in a home. We can see that all family meal windows contain large numbers of clattering frames. The clash of other objects such as keys and coins can also produce a similar sound. Different from clattering frames of dining activity, such false alarms are usually isolated and not likely to occur in a burst.

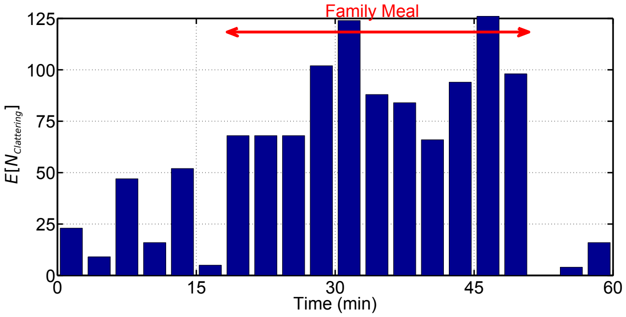

Fig. 3. An example of family meal detection. Each bar represents the expected number of frames containing clattering sound in a detection window.

2) Arm Gesture: When smartwatch is available, FamilyLog also extracts motion-based features that characterizes dining behavior, which include the acceleration on the X-axis ($\overline{Acc_x}$) and the changing rate of the acceleration ($\overline{R_c}$). The X-axis acceleration is sensitive to various arm movements, as its direction is always parallel to the user's arm. Therefore, it can be used as a simple and effective feature for inferring arm gesture while avoiding the overhead of data processing on the other two dimensions. Specifically, FamilyLog samples the built-in motion sensor on smartwatch and calculates two features for each frame. The X-axis acceleration is directly read from the accelerometer. The changing rate between two frames can be computed as the angle between two acceleration vectors from them. Since the acceleration is mostly corresponded to the gravity, the angle describes how much the orientation of the watch face is turned along with the user's action. For a detection window, $\overline{Acc_x}$ is calculated as the average acceleration on X-axis for all frames, and $\overline{R_c}$ is calculated by the average changing rate of all neighboring frames. Fig.4 shows three typical activities and the motion features. We can see that the arm gesture and the movements of wrist during meal show distinct distributions.

C. Conversation and TV Viewing Detection

1) Human Voice Identification: An important acoustic feature for the detection is the conversation, which identifies human speech, as well as the family members who participate in it. Among all the family communications, the family meal is typically accompanied by a considerable amount of conversations. The speaker recognition technique presented in [10] shows that pronunciation of vowels is a identical characteristic of human. However, maintaining a database for voice of each family member is costly for mobile devices. Here, row mean vector of spectrogram [20] provides an effective and efficient approach to recognize speakers by measuring Euclidean distance of energy distribution on frequency domain. Specifically, the family members are required to register their voice to FamilyLog by reading a short sentence. For each frame, FamilyLog compares the vector from MFCC-based processing with the ones obtained during training, and calculates the

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

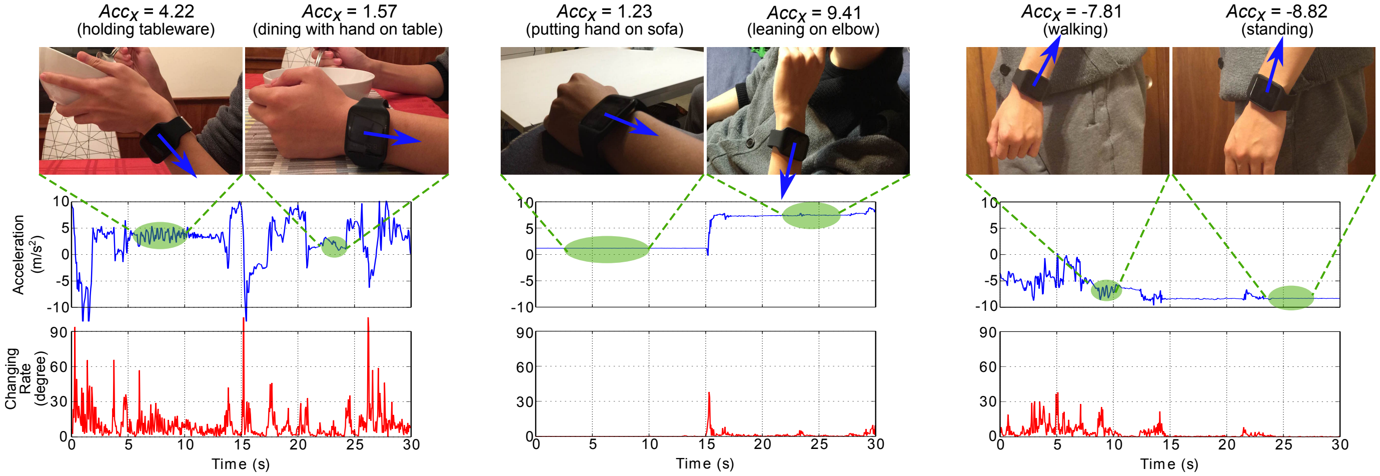

Fig. 4. Examples of typical activities and related motion features. The left column shows a dining scenario ($\overline{Acc_x} = 2.69m/s^2$, $\overline{R}_c = 10.41°$). The center column shows a TV viewing scenario ($\overline{Acc_x} = 4.25m/s^2$, $\overline{R}_c = 0.85°$). The right column shows a walking/standing scenario ($\overline{Acc_x} = -6.98m/s^2$, $\overline{R}_c = 3.19°$). The upper row shows the ground truth at some moments during these activities, and the arrows in these photos indicate the direction of X-axis. The acceleration on X-axis is shown for each photo. The center row shows the acceleration on X-axis in each frame. The lower row shows the changing rate of acceleration in each frame.

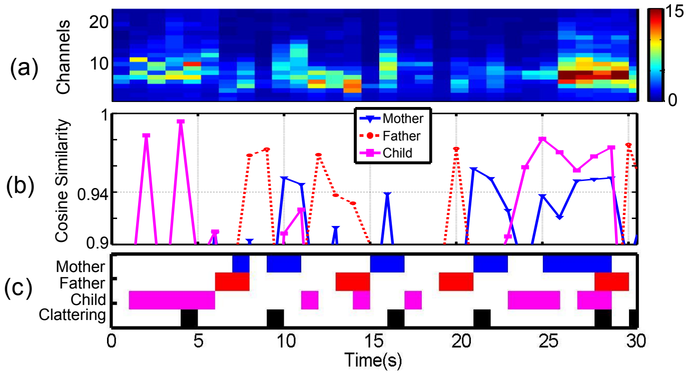

Fig. 5. An example of conversation detection during a typical dining scenario.

probability that the frame contains voice of at least one family member by cosine similarity, represented as $P(voice|\boldsymbol{E})$, where $\boldsymbol{E}$ is the energy distribution in the frame, as shown in Fig.5. In a detection window, FamilyLog sums $P(voice|\boldsymbol{E})$ for each frame to extract $E[N_{voice}]$, representing the expection of number of frames that contains family members' voice.

2) Localization-based TV Viewing Detection: TV is a sound source with fixed location, whose volume usually stays within a limited range. However, the clattering sound and the conversation during the mealtime come from multiple sound sources. We can focus on detecting the number of sound sources to find the existence of the TV, and seperate TV sound from the clattering sound or the human voice. FamilyLog employs a novel approach based on *Interaural Level Difference* (ILD) [3] that fuses acoustic features captured by different devices (features are exchanged on cloud servers, local wi-fi or bluetooth connections) to determine the sound sources. In this section we only focus on the fusion algorithm for two devices although it can be extended to more generic scenarios. Specifically, the process of feature fusion consists of two steps: similarity check and sound source detection. In the first step, it figures out whether two devices are at home and near each other by examining the similarity between sound captured by two devices. We define the detection windows that cover the same period of time on two different devices as the binaural detection windows. The similarity between binaural detection windows A and B can be calculated as follows:

$$C(A,B) = \frac{\sum_{i=1}^{l} cos(\boldsymbol{E}(A,i), \boldsymbol{E}(B,i))}{l} \quad (1)$$

where vector $\boldsymbol{E}(X,i)$ is the energy distribution for frame i in detection window X. FamilyLog only proceeds to conduct sound source detection if $C(A,B)$ is above a threshold, indicating the two devices are in proximity to one another. The sound source detection aims to detect the number of sound sources in binaural detection windows. A key observation is that if all the acoustic signal originates from a single sound source, it is more likely caused by TV. In contrast, if the acoustic signal originates from multiple sound sources, it is more likely to be caused by human activities other than TV. The method we use to detect sound sources is based on acoustic localization by ILD . Specifically, if the acoustic signal is from a single source and captured by two receivers, it satisfies $V_1/V_2 = d_1^2/d_2^2 = \Delta_V$, where V_1 and V_2 are

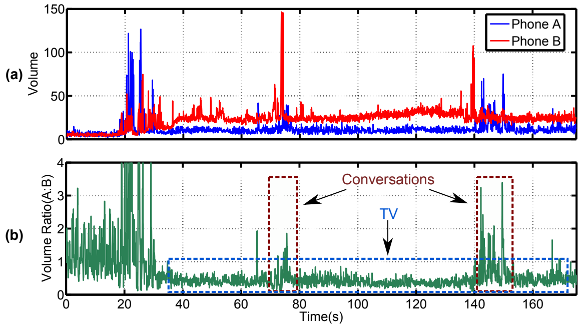

Fig. 6. An example of TV viewing with conversation. (a) shows captured volume by two smartphones. (b) shows the volume ratio between corresponding frames.

volumes received by receivers and d_1 and d_2 are distances between receivers and sound source, calculated by the RMS. This equation can be applied to compute the relative distances between the sound source and the devices. In indoor scenarios, Δ_V may be impacted by various factors (e.g., echoes and obstacles), but its coefficient of variation is limited when d_1 and d_2 are fixed. To detect whether the acoustic signals come from the same source, we define *Coefficient of Variation of Volume Ratio per Frame* ($CV(A, B)$) in binaural detection windows A and B as:

$$\begin{aligned} CV(A,B) &= \frac{\sigma(\Delta_V(A,B))}{\mu(\Delta_V(A,B))} \\ \Delta_V(A,B) &= \left\{ \frac{V_{A,i}}{V_{B,i}}, i \in [1,l] \right\} \end{aligned} \tag{2}$$

Here, the volume of frame i in detection window X is represented by $V_{X,i}$, $\mu(\Delta_V(A,B))$ is the mean of volume ratios between A and B, and $\sigma(\Delta_V(A,B))$ is the standard deviation of volume ratios. $CV(A,B)$ thus is the ratio of the standard deviation to the mean. The lower $CV(A,B)$ is, the more likely the acoustic signals come from a single source. Fig.6 shows an example of how to detect sound sources by volume ratio. In the first 20 seconds, phone B is carried by user from the dining table to the sofa. TV is turned on at the 30th second. During the 70th-75th second and the 140th-150th second, the subjects talk to each other. We can see that when the frames only contain TV sound, volume ratio is relatively stable. In contrast, as conversation involves multiple sound sources, the variance of the volume ratio is significantly increased.

By detecting the sound source with multiple devices, the accuracy of the detection of family meals can be improved in several challenging scenarios. Although TV programs that contain similar sound as family meal or conversation may be misclassified, the frames contain clattering sound and conversation still come from a single source, and they will be more likely from TV than family activities.

TV sound during the family meal can be separated from "foreground" sounds (clattering, conversation, etc.) by extracting low-energy frames, i.e. the frames that have a RMS less than the average RMS in a detection window. To detect whether TV is on during the family meal, we can check the volume of sound from all low-energy frames, and whether the acoustic signal is probably from a single sound source. If the TV is on, the continous sound from TV will rise the volume of low-energy frames, and $CV(A,B)$ of all low-energy frames will have a relatively low value, indicating the sound comes from a single sound source with fixed location.

D. HMM-based Classification

Similar to speech and gesture recognition, the family meal detection involves identifying a temporal pattern rather than detecting discrete events. We design the classifier of FamilyLog based on HMM, where we treat extracted features as observations, and the family event contained in each detection window as hidden state. Therefore, the primary goal of the our HMM-based classification is to recover the family events overtime using the features extracted from a sequence of detection windows.

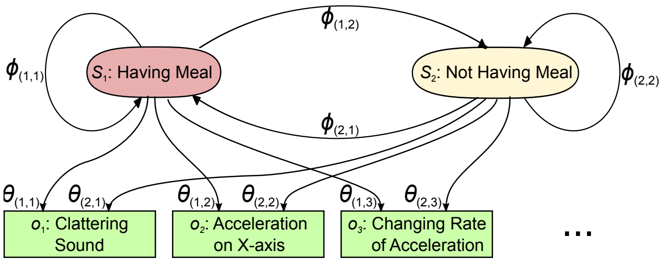

Fig. 7. The Hidden Markov Model of one family communication activity

Fig.7 shows the HMM-based classifier for family meal as an example. We can see that in this case, the state is either "having meal" or "not having meal", and the observations includes four features extracted from each detection window. The transition probabilities between two states are simply generated based on a simple survey conduced before using the system. The emission parameters associating states with observations are calculated using the one-day training data. Therefore, we formally define the our HMM-based classification as follows:

$$\arg\max_X P(X|\lambda, O) \tag{3}$$

where X is a sequence of states; λ represents the transition probabilities Φ and emission parameters Θ of the HMM; O is a sequence of observations. The output of the classifier is a sequence of states that maximize the likelihood, which can be calculated by using the *Viterbi algorithm*.

1) Transition Probabilities: The set of transition probabilities Φ contains four entries $\{\phi_{(1,1)}, \phi_{(1,2)}, \phi_{(2,1)}, \phi_{(2,2)}\}$. According to the definition of our HMM, when the activity is not occurring, we only need to know the probability of its occurrence in next detection window. On the other side, while the activity is currently occurring, we only need to know the probability of whether it continues in the next detection window. Therefore, two models are enough to describe all the transition probabilities, which are the probability distribution of one activity's occurrence related to time/date, and the probability distribution of its duration.

The probability distributions can be estimated based on one-time user answers to questions like "What's the typical frequency and duration of your weekday family meals?". Alternatively, they can be derived from historical detection results. To improve the accuracy of such an approach, FamilyLog presents intuitive system UIs that allow users to rate previous detection results. The characteristics of family meal, including time, duration, and frequency are often highly dependent on the day of the week. Therefore, FamilyLog generates different models for the weekdays and weekends.

The transition probabilities of our HMM can be read directly on generated models. When applying the Viterbi algorithm, the transition probability from S_2 to S_1 is equal to the probability of the activity's occurrence; the transition probability from S_1 to S_1 is equal to the probability of the activity's continuance to the next 3 minutes. Because the transition probabilities are dependent on previous states (due to the influence of activity's duration), we need to adjust the structure of our HMM to ensure Viterbi algorithm runs properly. As shown in Fig.8, all transition probabilities are independent of previous states, at the price of increased memory usage. In practice, we run Viterbi algorithm for 40 states in this HMM (i.e. 40 detection windows) to ensure that any activities within 2 hours can be captured.

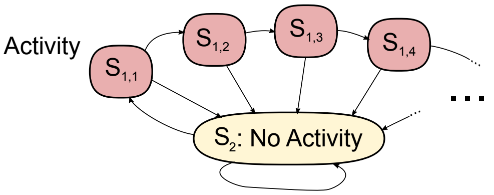

Fig. 8. The adjusted HMM structure. An activity is divided into multiple states as $S_{1,k}$, indicating that the activity already lasts for k detection windows. Here $S_{1,k}$ can only transit to $S_{1,k+1}$ or S_2, corresponding to the cases where the activity continues to the next detection window or stops, respectively.

2) Emission Parameters: The set of emission parameters Θ contains entries as $\theta_{(S,\boldsymbol{o})}$, which describes the probability to observe the observation $\boldsymbol{o}$ in state S. The observation $\boldsymbol{o}$ within a detection window is represented as a vector of features, i.e. $\boldsymbol{o} =< o_1, o_2, o_3, ... >$, where o_i corresponds to a feature related to the activity. For the detection of the family meals, the features are shown in Table.I.

TABLE I
FEATURES FOR THE FAMILY MEAL DETECTION

Term	Description
$E[N_{clattering}]$	The expectation of number of frames containing clattering sound
$E[N_{voice}]$	The expectation of number of frames containing the family members' voice
$\overline{Acc_x}$	The average acceleration on X-axis
$\overline{R_c}$	The changing rate of acceleration
$CV(A,B)$	The coefficient of variation of volume ratio per frame in binaural detection windows A and B

The HMM classifier is trained by a period of sensor data. Typically, at least a whole day is required to fully cover the communication of family. After training, we apply Gaussian KDE to calculate the PDF, which is represented as $p(\boldsymbol{o}|S)$, corresponding to the observations associated to each state. Therefore, the emission parameter is defined as $\theta_{(S,\boldsymbol{o})} = p(\boldsymbol{o}|S)$ for the HMM with multiple continuous observations [17].

V. Performance Evaluation

In order to evaluate the performance of FamilyLog, we have collected 77 days of data from 37 subjects in 8 families (details shown in Table II). The procedure of the data collection has been approved by the *Institutional Review Boards* (IRB) at the Michigan State University. The period of data collection was one or two weeks for each family. We intentionally chose families with young children for this study because family routine analysis has important implications for children's health. Our results also showed that small children often sometimes presented challenges to event detection due to the excessive noise they make at home.

We provided each family one or multiple devices. An app pre-installed on the devices continuously records audio and motion unless the device is taken out of home. The app runs automatically, but the subjects can manually start/end the recording on any device. The devices can be carried with the subjects or left somewhere in the house, depending on their habits. We also offer them the opportunity to review the recording and delete the part of recording that raises privacy concerns. We adopted two methods to obtain the ground truth, which include an interview with the family members immediately after data collection is finished, and listening to the recordings to manually label family activities.

A. Micro-scale Routine Analysis

To evaluate the performance of our HMM-based classifier, we compare our classification result with the result classified by the *Support Vector Machine* (SVM), which recognizes the family meals only based on features in individual detection windows rather than considering their temporal nature. The overall performance of FamilyLog and its comparison with SVM will be discussed later in SectionV-B.

Fig.9 shows the detection results along with the ground truth of the data from 5 days in Family 4. We can see that the family usually has dinner around 7-8 pm for about an hour, except for day 5, which is Friday, when they started dinner at around 8 pm for about 20 minutes. Compared with the ground truth, we can see that FamilyLog is accurate in detecting most of the meals. In day 3 and 5, the SVM classifier yields a few misclassifications due to the interferences caused by TV viewing. However, FamilyLog's HMM-based classifier is able to avoid such false negative errors. Furthermore, by taking into account the temporal nature of family routine activities, HMM is able to minimize the short false negative and false positive classification results.

B. Evaluation of Meal Detection

In this section, we investigate the overall performance of FamilyLog in detecting family meals. For each individual family, the HMM-based classifier is trained using the information from the survey and data labeled by the subjects collected in the first day. We use the precision and recall as the metrics for this evaluation. Specifically, the precision is defined as the ratio of the number of true-positive windows to the total number of windows. The recall is defined as the number of true-positive windows divided by the total number of windows detected as family meals. The true negatives are not considered, because most of the windows containing no activities are able to be detected, and they have been discarded. In addition, we also present the evaluation result after making certain relaxation (e.g., $\pm 3min$) on the start/end time. Note that our design objective will not be affected by minor errors in start/end time, as long as the the system is able to accurately identify the occurrences of the family meals.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

TABLE II
FAMILIES THAT PARTICIPATED IN THE EXPERIMENT

Family	Children	Phone	Smartwatch	Data	Family Meals
	(Ages in Years)			(Weeks)	(Number of Times)
1	1 daughter(5)	Nexus 4	N/A	1	4
2	1 daughter(4)	Nexus 4	N/A	1	6
3	2 daughters(5, 8),2 sons(1, 3)	Nexus 4	Sony Smartwatch 3	2	9
4	3 sons (1, 3, 5)	Nexus 3	Sony Smartwatch 3	2	16
5	2 sons (3, 5)	Moto G	N/A	1	5
6	2 daughters(1,3),1 son(7)	Moto G2 × 2	Sony Smartwatch 3 × 2	2	22
7	2 daughters(3,11),2 sons(7,13)	Moto G2 × 2	N/A	1	10
8	3 daughters(7,10,18)	Moto G2 × 2	Sony Smartwatch 3	1	6

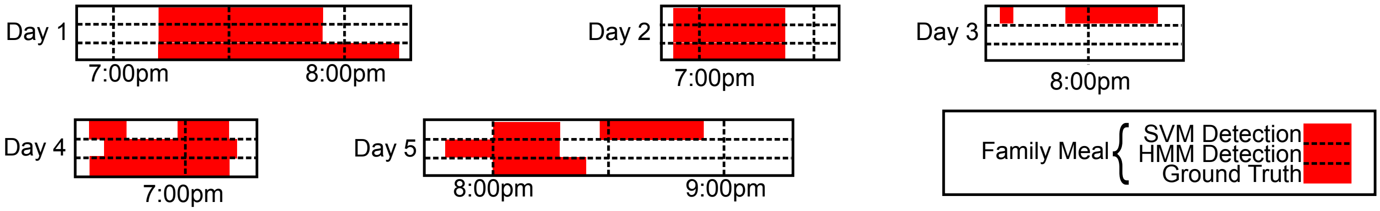

Fig. 9. Detected family meals based on data collected from family 4 during 5 days.

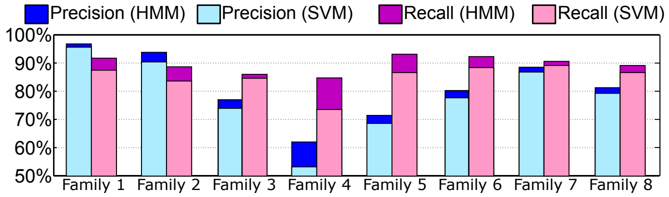

Fig. 10. Overall accuracy of family meal detection in detection windows. The average precision and recall of FamilyLog are 80.7% and 89.5%, respectively.

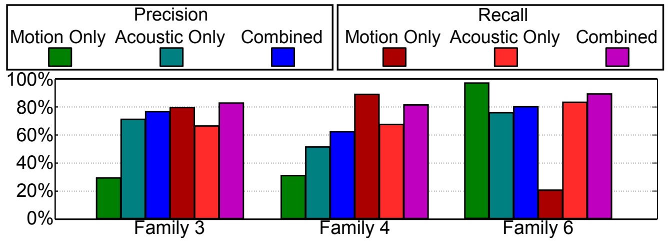

Fig. 11. Accuracy of family meal detection using only motion or acoustic data.

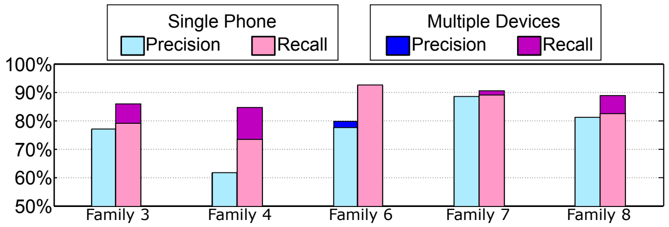

Fig. 12. Accuracy of family meal detection by a single phone or all the available devices.

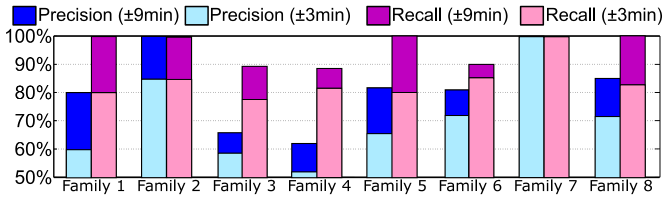

Fig. 13. Accuracy of detection for each occurrence of family meal by relaxing the start/end time by $\pm 3min$ and $\pm 9min$.

The evaluation result of family meal detection is shown in Fig.10. Our HMM-based classifier outperforms SVM by 6.82% on average in recall. This is primarily because HMM is more effective in correcting isolated false negatives. We can also observe that FamilyLog achieves an overall precision of 81.1%, with the highest being 91.1% for family 1 and lowest being 62% for family 4. We found that the two major causes of the relatively low precision in family 4 and 5 are the high pitch voice from children and music, which have similar acoustic features as clattering sound during meal. However, since these sounds usually have a short duration, FamilyLog is able to correct a considerable amount of the resulting false positives.

For the detection that is only based on the motion data from the smartwatch or the acoustic data, the accuracy is shown in Fig.11 for Family 3, 4, and 6. For the Family 3 and 4, the precision is very low when only motion data is used. The reason is that the motion data for the eating action can be very similar to some activities like reading or writing, especially when the smartwatch is wear on the non-dominant hand. On the other side, the recall is relatively high, because most of the family meals are able to be correctly detected by the motion data. Moreover, the smartwatch in Family 6 is rarely worn when they are at home, and the detection based on the motion data is not always reliable. Generally, the features from the acoustic data contribute the major part of the detection, and the motion data can assist the detection in some special cases. For example, depending on the food, the clattering sound may be weak for a family meal, but the detection result can still be correct due to the conversation between the family members and the eating action.

Fig.12 shows a comparison of the accuracy of the detection by a single smartphone or all the available devices in a family. If FamilyLog only runs on a single device, the sound source detection will be unavailable. This happens in Family 6, where the sound from a TV program about cooking

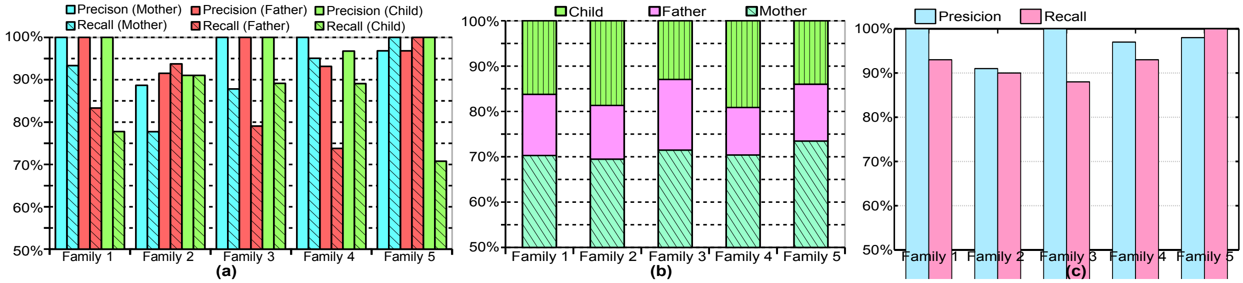

Fig. 14. Evaluation of participant detection. (a) shows accuracy of speaker recognition. (b) shows each member's overall proportion in family conversation. (c) shows accuracy of overall participant detection for each family. The average precision and recall are 97.8% and 92.8%, respectively.

is wrongly detected as a meal without knowing the sound sources. Furthermore, during a family meal, if one device is left far away from the dining table but another device is near, it is possible that the meal can only be detected by one device. By combining the results from all the available devices, FamilyLog is less likely to miss a family meal than only relying on one of them.

Fig.13 shows the detection accuracy of the occurrence of each family meal. We can see that FamilyLog rarely fails to detect an occurrence of family meals with the $9min$-relaxation on the starting/ending time, achieving 88.7% precision and 93.3% recall on average. The detection error in each family meal's duration is about 4 minutes on average.

C. Participant and TV Detection

The human voice serves as a clue for family meals, and is also unique feature of a participant. With the permissions from the Family 1-5, we listened to the raw acoustic data provided by them, and manually labeled the family members who have talked during each family meal. For each family, we only focus on the mother, the father, and one selected child aged between 5-12. We count the number of detection windows that FamilyLog can correctly detect all the participants, and calculate the precision and recall. Fig.14(c) shows that, FamilyLog achieves high accuracy in participant detection across different families, with the average precision and recall being 97.8% and 92.8%, respectively. This means most of the detection windows yields correct results for participant detection. This also ensures a high accuracy for detecting all participants for each occurrence of the family meal.

Fig.14(a) shows the participant identification result for each family member. One key observation is that the overall recognition accuracy of other family members are better than that of fathers. This is mainly due to the fact that father usually speaks with short sentences or only phases, which are more difficult to detect. Fig.14(b) shows the proportion for each family member in overall conversation. We can see that father speaks less frequently that other family members, which is also consistent with the findings from social behavior studies [31]. Another observation is that the child from family 5 has a relatively low recall. This is mainly because he often speaks with different tones, thus the voice is difficult to identify using the signature extracted from his training data.

TABLE III
TV VIEWING DURING FAMILY MEALS

Total: 78		Detected		Not available
		TV is on	TV is off	
True	TV is on	34	4	15
	TV is off	3	22	

Table III shows the result of detection of TV viewing during the family meals. Some of the detections are unavailable, because only one device is used for the experiment, while the sound source detection requires at least two different devices.The accuracy is 91.8% precision and 89.5% recall. The errors are mainly caused by the noise, or the long distance between the TV and the devices. It can be seen that the TV is turned on during over 60% of the family meals in our experiment.

VI. Conclusion

In this paper we present the design and implementation of FamilyLog – a practical system to log family mealtime activities using off-the-shelf smartwatches and smartphones. It uses the built-in accelerometers on the smartwatches and microphones of all mobile devices to detect family mealtime activities that are closely related to the family wellness, including occurrence and duration of meal, conversations, participants, and TV viewing. The design of FamilyLog addresses several challenges such as the significant interferences from various noises in the home. We carefully analyze the sensor data collected from real families and design the signal features for HMM-based activity classification, which are robust against various noises and can be computed efficiently on mobile devices. We have evaluated FamilyLog with extensive experiments involving 8 families with children (at least one-week recording in each family). Our results show that FamilyLog is effective in logging details of family meals across different families and home environments.

Acknowledgments

This work was supported in part by U.S. National Science Foundation under grant IIS1521722.

References

[1] S. Avancha, A. Baxi, and D. Kotz. Privacy in mobile technology for personal healthcare. *ACM Comput. Surv.*, 45(1):3:1–3:54, Dec. 2012.

[2] N. Bellissimo, P. B. Pencharz, S. G. Thomas, and G. H. Anderson. Effect of television viewing at mealtime on food intake after a glucose preload in boys. *Pediatric Research*, 61(6):745–749, 2007.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

[3] S. T. Birchfield and R. Gangishetty. Acoustic localization by interaural level difference. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 4, pages iv–1109. IEEE, 2005.

[4] S. Blum-Kulka. *Dinner talk: Cultural patterns of sociability and socialization in family discourse*. Routledge, 1997.

[5] W. T. Boyce, E. W. Jensen, J. C. Cassel, A. M. Collier, A. H. Smith, and C. T. Ramey. Influence of life events and family routines on childhood respiratory tract illness. *Pediatrics*, 60(4):609–615, 1977.

[6] W. K. Bryant and C. D. Zick. An examination of parent-child shared time. *Journal of Marriage and Family*, 58(1):pp. 227–237.

[7] S. A. Denham. Relationships between family rituals, family routines, and health. *Journal of Family Nursing*, 9(3):305–330, 2003.

[8] T. J. Dishion, S. E. Nelson, and K. Kavanagh. The family check-up with high-risk young adolescents: Preventing early-onset substance use by parent monitoring. *Behavior Therapy*, 34(4):553–571, 2003.

[9] C. J. Dunst, D. Hamby, C. M. Trivette, M. Raab, and M. B. Bruder. Everyday family and community life and children's naturally occurring learning opportunities. *Journal of Early Intervention*, 23(3):151–164, 2000.

[10] N. Fakotakis, A. Tsopanoglou, and G. Kokkinakis. Text-independent speaker recognition based on vowel spotting. In *Digital Processing of Signals in Communications, 1991., Sixth International Conference on*, pages 272–277, Sep 1991.

[11] B. H. Fiese, A. Hammons, and D. Grigsby-Toussaint. Family mealtimes: a contextual approach to understanding childhood obesity. *Economics & Human Biology*, 10(4):365–374, 2012.

[12] K. R. Ginsburg et al. The importance of play in promoting healthy child development and maintaining strong parent-child bonds. *Pediatrics*, 119(1):182–191, 2007.

[13] K. P. Goebel and C. B. Hennon. Mother's time on meal preparation, expenditures for meals away from home, and shared meals: Effects of mother's employment and age of younger child. *Home Economics Research Journal*, 12(2):169–188, 1983.

[14] L. Greening, L. Stoppelbein, C. Konishi, S. S. Jordan, and G. Moll. Child routines and youths adherence to treatment for type 1 diabetes. *Journal of Pediatric Psychology*, 32(4):437–447, 2007.

[15] S. Gupta, M. S. Reynolds, and S. N. Patel. Electrisense: Single-point sensing using emi for electrical event detection and classification in the home. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, Ubicomp '10, pages 139–148, New York, NY, USA, 2010. ACM.

[16] T. Hao, G. Xing, and G. Zhou. isleep: Unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 4:1–4:14, New York, NY, USA, 2013. ACM.

[17] A. Honkela. *Nonlinear switching state-space models*. PhD thesis, Citeseer, 2001.

[18] Z. Huang and T. Zhu. Sbd: A signature-based detection for activities of appliances: Poster abstract. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, BuildSys '14, pages 222–223, New York, NY, USA, 2014. ACM.

[19] A. Kadomura, C.-Y. Li, Y.-C. Chen, H.-H. Chu, K. Tsukada, and I. Siio. Sensing fork and persuasive game for improving eating behavior. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 71–74. ACM, 2013.

[20] H. B. Kekre, A. Athawale, and M. Desai. Speaker identification using row mean vector of spectrogram. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, ICWET '11, pages 171–174, New York, NY, USA, 2011. ACM.

[21] S. Kopparapu and M. Laxminarayana. Choice of mel filter bank in computing MFCC of a resampled speech. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, pages 121–124, May 2010.

[22] X. Ma, C. Fellbaum, and P. R. Cook. Soundnet: Investigating a language composed of environmental sounds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1945–1954, New York, NY, USA, 2010. ACM.

[23] S. Nirjon, R. F. Dickerson, P. Asare, Q. Li, D. Hong, J. A. Stankovic, P. Hu, G. Shen, and X. Jiang. Auditeur: A mobile-cloud service platform for acoustic event detection on smartphones. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 403–416, New York, NY, USA, 2013. ACM.

[24] U. B. of Labor Statistics. *American Time Use Survey*. 2015.

[25] D. O'Shaughnessy. *Speech communication: human and machine*. Addison-Wesley series in electrical engineering: digital signal processing. Universities Press (India) Pvt. Limited, 1987.

[26] D. E. Phillips, R. Tan, M.-M. Moazzami, G. Xing, J. Chen, and D. K. Y. Yau. Supero: A sensor system for unsupervised residential power usage monitoring. *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 0:66–75, 2013.

[27] A. Prasad, J. Sorber, T. Stablein, D. Anthony, and D. Kotz. Understanding sharing preferences and behavior for mhealth devices. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, pages 117–128, New York, NY, USA, 2012. ACM.

[28] M. Rabbi, S. Ali, T. Choudhury, and E. Berke. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 385–394, New York, NY, USA, 2011. ACM.

[29] M. Sahidullah and G. Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication*, 54(4):543 – 565, 2012.

[30] S. S. Stevens. A Scale for the Measurement of the Psychological Magnitude Pitch. *Acoustical Society of America Journal*, 8:185, 1937.

[31] D. Tannen. *You just don't understand: Women and men in conversation*. Virago London, 1991.

[32] D. Tharini and J. Kumar. 21 band 1/3-octave filter bank for digital hearing aids. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, pages 353–358, March 2012.

[33] E. Thomaz, I. Essa, and G. D. Abowd. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1029–1040. ACM, 2015.

[34] E. Thomaz, C. Zhang, I. Essa, and G. D. Abowd. Inferring meal eating activities in real world settings from ambient sounds: A feasibility study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 427–431, New York, NY, USA, 2015. ACM.

[35] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, and B. Firner. Crowd++: Unsupervised speaker count with smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 43–52, New York, NY, USA, 2013. ACM.