Ubiquitous Computing and Communication Journal
(ISSN 1992-8424)

AUTOMATIC LIFE-LOGGING: A NOVEL APPROACH TO SENSE REAL-WORLD ACTIVITIES BY ENVIRONMENTAL SOUND CUES AND COMMON SENSE

Mostafa M. A. Shaikh, Md. Khademul Islam Molla and Keikichi Hirose
Dept. of Information and Communication Engineering
The University of Tokyo, Tokyo, Japan
E-mail: mostafa_masum@ieee.org, {molla, hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

People can stay connected to their loving ones ubiquitously that they care about by sharing awareness information in a passive way. Only very recently the detection of real-world activities has been attempted by processing multiple sensors data along with inference logic for real-world activities. This paper proposes to infer human activity from environmental sound cues and common sense knowledge, which is an inexpensive alternative to the traditional sensors. Because of their ubiquity, the mobile phones or hand-held devices (HHD) are ideal channels to achieve a seamless integration between the physical and virtual worlds. Therefore, a prototype is proposed here to log daily events by HHD based application to infer activities from environmental sound cues. To the best of our knowledge, this system pioneers the use of environmental sound based activity recognition in mobile computing to reflect one's real-world activity for ambient communication.

Keywords: ambient communication, life logging, acoustic event detection, activity recognition, sound cues, auditory scene analysis.

1 INTRODUCTION

Although speech is the most informative acoustic event, other kind of sounds may also carry useful information regarding the surrounding environment. In fact, in that environment the human activity is reflected in a rich variety of acoustic events, either produced naturally or by the human body or by the objects manipulated or interacted by humans. Consequently, detection or classification of acoustic events may help to detect and describe the human and social activity that takes place in the environment. For example: Jingling sound of cooking utensils (like cooking pan, spoon, knife etc.) may lead to infer someone's cooking activity, vehicle passing sound may lead to infer that someone is on the road, etc. Additionally, the robustness of automatic speech recognition systems may be increased by a previous detection of the non-speech sounds lying in the captured signals.

Many sources of information for sensing the environment as well as activity are available [1][2][3]. In this paper, we consider two objectives namely, sound-based context awareness, where the decision is based merely on the available acoustic information at the surrounding environment of the user and automatic life-logging, where the detected sound context infers an activity to be logged along with temporal information. Acoustic Event Detection (AED) is a recent sub-area of computational auditory scene analysis [4] that deals with the first objective. AED processes acoustic signals and converts those into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources. Life-log is a chronological list of activities performed by the user with respect to time. Such a list might indicate the user's well-being or abnormality according to the consideration of the person's self assessment or by someone else who cares about the person (e.g., relatives or care-givers). Therefore we apply the concept of AED to perform automatic generation of life-log. This life-log can be transmitted autonomously as a simple text message to someone else with the notion of ambient communication.

In this paper, we describe a listening test made to facilitate the direct comparison of the system's performance to that of human subjects. A forced choice test with identical test samples and reference classes for the subjects and the system is used. The second main concern in this paper is to evaluate how acceptable the automatic generation of life-log is. Since we are dealing with a highly varying acoustic material where practically any imaginable sounds can occur, we have limited our scope in terms of location and the activities to recognize at a particular location. It is most likely that the acoustic models we are using are not able to sufficiently model the observation statistics. Therefore, we propose using

discriminative training instead of conventional maximum likelihood training.

The paper is organized as follows: Section 2 explains the motivation of the envisioned application. Section 3 reviews the background studies related to this research. Our approach, in terms of system architecture and description of the system components is explained in Section 4. Section 5 discusses about several development challenges and our response towards those. Section 6 explains the experimental setup, the results obtained by the detection and classification system as well as user evaluations. Conclusions are presented in Section 7.

2 MOTIVATION

Life logs include people's activities performed in specific locations at a specific time and it can be collected from various sources. We envision that with the proliferation of computing power of hand held devices (HHD), availability of the Internet connectivity and improvements in communication technologies ambient communication will find a universal place at our daily life and allow us to create a vivid and intelligent online social network. Let's consider the following scenario.

Scenario 1: Rahman family (Mr. and Mrs. Rahman) lives in Khulna, one of the metropolitan cities of Bangladesh. They have three sons living overseas, one in Texas, another in Ottawa and the youngest one in Bonn of Germany. The Rahman family is now at their age of over 50 and Mr. Rahman had a massive heart operation last year. Mrs. Rahman is also ailing from several sicknesses like diabetics, high blood pressure, etc. The three sons are always worried regarding the well beings of their parents and consequently they often talk to their parents over the phones to know their whereabouts. Though calling to Bangladesh from USA, Canada, and Germany is relatively cheaper now-a-days than before, but having a phone conversation with their parents is not always possible due to various reasons, for example, due to inconvenience in time differences (e.g., when it is 10 am in Khulna it is 11:00 pm in Texas, 12:00 am in Ottawa and 6:00 am in Bonn) that is, when the sons have convenient time to call, their parents are usually sleeping or resting and vise versa. But they are often worried to know at least how their parents are doing everyday. Therefore, let's imagine that Rahman family has internet connectivity at their house and installed an inexpensive system capable of doing the following that may provide mental peace to their sons by providing ambient communications through the internet. The system makes automatic life logging of daily activities by detecting and recognizing sound cues from their surrounding environments and sends email message(s) to their sons reporting their daily life-sketch. In this case, an example email message containing life log for a particular day as following might be very relieving to the sons. *"Your parents woke up at 7:30 am today and they had breakfast around 8:15 in the morning. They watched TV several times in the day. Went out of home for two times and walked in the roads and parks. They took lunch and dinner at around 2 PM and 8 PM. Your mother went to toilet for 5 times and father went to toilet 6 times in a day. They had communicated with each other or other people by talking. It seems they are doing fine".*

3 BACKGROUND

A number of researchers have investigated to infer activities of daily living (ADL). In [5] authors have successfully used cameras and a bracelet to infer hand washing. The authors of [6] used radio-frequency-identification (RFID) tags functionally as contact switches to infer when users took medication. The system discussed in [7] used contact switches, temperature switches, and pressure sensors to infer meal preparation. Authors of [8] used cameras to infer meal preparation. In [9] authors used motion and contact sensors, combined with a custom-built medication pad, to get rough inference on meal preparation, toileting, taking medication, and up-and-around transference. A custom wearable computer with accelerometers, temperature sensors, and conductivity sensors to infer activity level is used in [10]. Author of [11] used 13 sensors to infer home energy use, focusing on the heating-use activity. Motion detectors to infer rough location were used in [12]. Several sensors like motion sensors, pressure pads, door latch sensors, and toilet flush sensors to infer behavior are reported in the system described in [13]. The authors [1] have described monitoring bathroom activities based on sound. The system [2] utilized RFID tags to detect objects and thereby inference of activities is done from the interaction with the detected objects. The research on MIT's *house_n* project [14] places a single type of object-based adhesive sensor in structurally unmodified homes and sensor readings are later analyzed for various applications—kitchen design, context sampling, and potentially ADL monitoring. All of these systems have a commonality that they perform high-level inference from low-level by coarse sensor data reporting and analyses. Some have added special pieces of hardware to help performance improvement, but progress toward accurate ADL detection has nevertheless been slow. Only a few researchers have reported the results of any preliminary user testing [5][9][12][13]. The level of inference using sensors has often been limited—for example, reporting only that a person entered the living room and spent time there. Moreover, as an example, research aiming to detect hand washing or tooth brushing have had nearly no synergy, each

using its own set of idiosyncratic sensors and algorithms on those sensors. Furthermore a home deployment kit designed to support all these ADLs would be a mass of incompatible and non-communicative widgets. Our approach instead focuses on a general inference engine and infers activities from the sound cues that are likely to be produced either naturally or from the interactions with objects. Thus we can use our system for many ADLs.

The idea of a "life-log" or a personal digital archive is a notion that can be traced back at least 60 years [15]. Since then a variety of modern projects have spawned such as the *Remembrance Agent* [16], *the Familiar* [17][18], *myLifeBits* [19], *Memories for Life* [20] and *What Was I Thinking* [21]. In [22] the authors evaluate the user's context in real time and then use variables like current location, activity, and social interaction to predict moments of interest. Audio and video recordings using a wearable device can then be triggered specifically at those times, resulting in more interest per recording. Some previous examples of this approach are the Familiar and *iSensed* systems [17,18,22] which structure multimedia on the fly; the *eyeBlog* system [23] which records video each time eye contact is established; and the *SenseCam* [24] which records images and sound whenever there's a significant change in the user's environment or the user's movement. Life log includes people's experiences which are collected from various sensors and stored in mass storage device. It is used to support user's memory and satisfy user's needs for personal information. If he wants to inform other people of his experience, he can easily share his experience with them by means of providing his life log.

To collect life log (e.g., GPS based location, SMS, call, charging, MP3, photos taken, images viewed, and weather information, etc) smart phones (e.g., iPhone 3G) are usually used. Smart phone is a mobile device that includes color LCD screen, mass storage, large memory, and communicative function by using Wi-Fi, Bluetooth, and infrared. It also has a variety of software such as scheduler, address book, media player, and e-book. Mika Raento developed a framework for collecting contexts from smart phone [25], which collects GSM Cell ID, Bluetooth, GPS data, phone data, SMS data, and media information that are transmitted to the server. The contexts could be provided for other contents as additional information. Panu *et al.* collect log data from mobile devices, and extracts features by pre-processing the log data [26]. The mobile device uses GPS log, microphone, temperature, moisture, and light sensor. *MyLifeBits* Project is one of the implementations of personal record database system [19]. Personal information is collected by PC, *SenseCam* and so on, and stored in database server with relationships among personal information. However, user faces difficulties to explore and search contents because of large amount of personal data. KeyGraph-based mobile contents management system was suggested to manage user's information in mobile device, which extracted important information using KeyGraph algorithm and provided searching or exploring contents [27]. The problem of the system is using only log data. If analysis and inference of the data was added to the system, it would give better performance.

Our work differs from others in three key ways. First, we utilize environmental sounds cues to get the idea regarding the interactions with objects or environment instead of sensor or camera data. Thus we can identify a large set of objects like spoons, toothbrushes, plates etc. Second, due to simple use of microphone to capture environmental sound we can also infer outdoor environments like on the road, in a park, in a train station etc. that previous research was limited to perform. Thirdly, our model is easy to incorporate new a set of activities for further needs by just adding more appropriately annotated sound clips and re-training of the HMM based recognizer.

4 OUR APPROACH

Our approach of logging daily events is to detect activities of daily living (e.g., laughing, talking, traveling, cooking, sleeping, etc.) and situational aspects of the person (e.g., inside a train, at a park, at home, at school, etc.).The system infers human-activity from environmental and object-interaction related sound cues as well as common sense knowledge. Initially we have collected 114 types of acoustic sounds that are usually produced during object interaction (e.g., cooking pan jingling sound while cooking) or by the environment itself (e.g., bus/car passing sound while on a road) or by a deliberate action of a person (e.g., laughing, speaking). Such kind of sounds serve as the underlying clues to infer a person's activity and one's surroundings with the help of common sense knowledge (e.g., while the system identifies cooking pan's jingling and chopping board sound as consecutive cues and system's local time indicates evening then from common sense database the system infers this activity as 'cooking').

4.1 System Architecture

Figure 1 serves as the top-level pipelined architecture of the system as following. Because of their ubiquity we plan to use HHD (e.g., portable computer or smart phone) to deploy this application that will capture environmental sound at some intervals to be processed. According to Figure 1, a mixed signal is passed through a robust signal processing and sound classes are detected by trained HMM classifiers. Based on the detected sounds and common sense knowledge regarding human activity,

object interaction, ontology of human life (e.g., daily life of a student, or a salary man etc.) and temporal information (e.g., morning, noon etc.) are applied to infer both the activity and the surrounding of the person. This information is then stored in the log of activities.

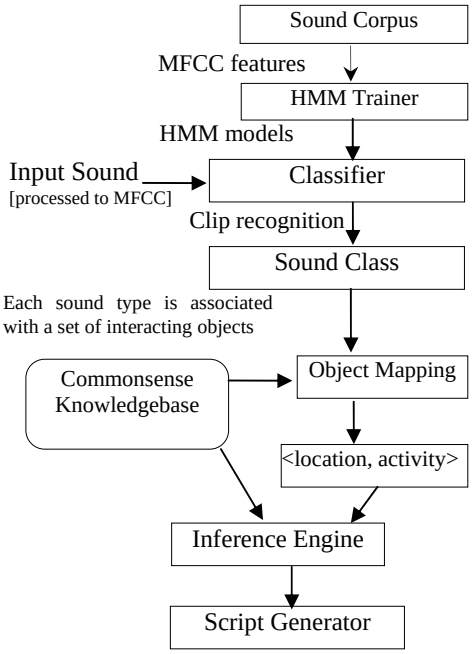

Figure 1: The System's Architecture

4.2 Description of System Components

In this section the system components are described briefly.

4.2.1 Sound Corpus

The patterns of sounds arising from activities occurring naturally or due to interaction with the objects are obviously a function of a many environmental variables like size and layout or the indoor environment, material of the floors and walls, type of objects (e.g., electrical or mechanical) and persistent ambient noise present in the environment etc. In is essential to install this system to the same culture and environment from where sound samples are acquired and proper training of the system is made. It is analogous to the practice adopted for speech recognition whereby the system is individually trained on each user for speaker dependent recognition because such environmental sounds may vary in different cultures and places. Therefore the sample sounds we have collected are from the different places of Tokyo city and Tokyo University. For clear audio-temporal delineation during system training, the sound capture for each activity of interest was carried out separately. A number of male and female subjects were used to collect the sounds of interest; each subject would typically go into the particular situation as depicted in Table I with the sound recording device and the generated sounds are recorded. We used the digital sound recorder of SANYO (model number: ICR-PS380RM) and signals were recorded as Stereo, 44.1 KHz, .Wav formatted files. It is important to note that in the generation of these sounds, associated 'background' sounds such as the ambient noise, rubbing of feet, friction with cloths, undressing, application of soap, etc., are being simultaneously recorded. The variability in the captured sounds of the each activity provides realistic input for system training, and increases the robustness and predictive power of the resultant classifier. Some sounds (e.g., water falling, vacuum cleaning machine sounds etc.) are generally loud and fairly consistent. There are samples that needed to sufficiently train the classification model due to a high degree of variability even for the same individual. For example, hands washing, drinking, eating, typing related sounds exhibited a high degree of variability. This required us to collect many more samples for such kind of activities related sounds to capture the diversity of the sounds.

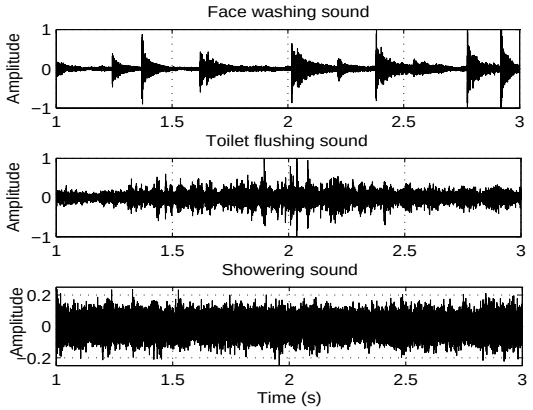

Figure 2: Waveforms of different water related sound types

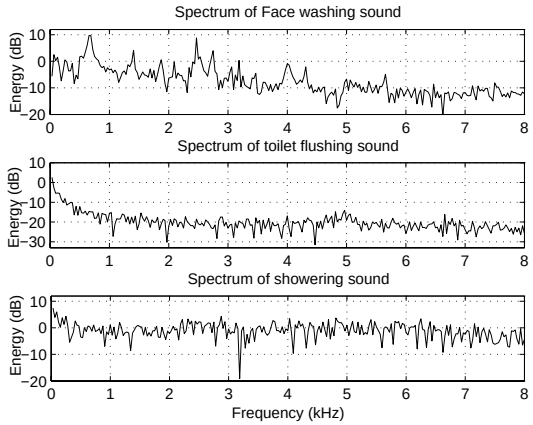

Figure 3: Spectrum of the sounds of Figure 2.

Table 1: Sample list of sound classes and mapped objects

Sound Class	Mapped Objects
vehicle_pass	road, bus, car, cycle, wind, people
TABLE I. toilet_flush	water, toilet, toilet-flush
TABLE II. n_shower	bathroom, shower, towel, water, body-wash
TABLE III. n_open_air_train	train, journey, people, announcement
TABLE IV. ood_eating	soba, noodle, people, chop stick, soup, rice, plate, cutleries, food, kitchen, restaurant
TABLE V. V_watching	TV, game, living room, news
TABLE VI. rain_enter_leave	people, train, announcement, station
TABLE VII. ater_basin	restroom, water basin, wash, hand soap
TABLE VIII. ater_sink	kitchen, sink, water, dish washer

The typical waveforms of the water related sounds of five different actions are shown in Figure 2. As can be seen, the water splashing for face washing (top), toilet flushing (middle) and shower water (bottom), the sounds as depicted in the waveforms has distinct patterns with different amplitudes. The Fourier spectra of those three type sounds are illustrated in Figure 3. According to the location and activities of our interest mentioned in Table I, we have collected 114 types of sounds. Each of the sound types has 15 samples of varying length from 10 second to 25 seconds.

4.2.2 HMM Training

An accurate and robust sound classifier is critical to the overall performance of the system. There are however many classifier approaches in the literature, e.g., those based on Hidden Markov Models (HMMs), Artificial Neural Networks (ANN), Dynamic Time Warping (DTW), Gaussian Mixture Models (GMM), etc. From these options, we have chosen an approach based on HMM as the model has a proven track record for many sound classification applications [28][29][31]. Another advantage is that it can be easily implemented using the HMM Tool Kit (HTK) [32] which is an open source Toolkit available on the Internet. It is also notifying that HTK was originally designed for speech recognition, which means we need to adapt the approach when applying in for environmental sounds of our interest. Each sound file, corresponding to a sample of a sound type, was processed in frames pre-emphasized and windowed by a Hamming window (25 ms) with an overlap of 50%. A feature vector consisting of a 39-order MFCC characterized each frame. We modeled each sound using a left-to-right forty-state continuous-density HMM without state skipping. Each HMM state was composed of two Gaussian mixture components. After a model initialization stage was done, all the HMM models were trained in eight iterative cycles.

4.2.3 Classification

It was obvious that simple frequency characterization would not be robust enough to produce good classification results. To find representative features, previous study [28] carried out an extensive comparative study on various transformation schemes, including the Fourier Transform (FT), Homomorphic Cepstral Coefficients (HCC), Short Time Fourier Transform (STFT), Fast Wavelet Transform (FWT), Continuous Wavelet Transform (CWT) and Mel-Frequency Cepstral Coefficient (MFCC). It was concluded that MFCC might be the best transformation for non-speech environmental sound recognition. A similar opinion was also articulated in [29][30]. These finding provide the essential motivation for us to use MFCC in extracting features for environmental sound classification. For classification, continuous HMM recognition is used. The grammar used is as follows: (<alarm_clock|ambulance|body_spray|chop_board|cycle_bell|dish_cleaning|egg_fry|face_wash|female_cough|toilet_flush|food_eating|gargle|hair_dry|in_open_air_train|insect_semi|in_shower|in_subway_train|in_train_station|liquid_stir|lunch_time_cafe|male_cough|male_male_Speak|male_sniff|meeting_talk|microwave_oven|paper_flip|phone_vibration|plate_clutter|plate_knife|silence|pan_spoon|tap_water|tooth_brush|train_enter_leave|tv_watching|urination|vacuum_cleaner|vehicle_pass|water_basin|water_sink>), which means that there is no predefined sequence for all the activities and each activity may be repeated at time at any sequence.

4.2.4 Sound Class

We have 114 types of sounds that are grouped into 40 sound classes to infer 17 kinds of activities. For example, we have sound samples named as the following types, “eating soba”, “food munching”, “belching”, “spoon plate friction”, which are grouped into one sound class called “food_eating”. Similarly, the sound samples named as, “news on TV”, “talk-show on TV”, “baseball game on TV”, “football game on TV” are grouped into “tv_watching” sound class. By such grouping of the samples into a particular sound class, we mean that the group of samples is used to train the classifier to recognize that particular sound class. Thus we have 40 sound classes.

4.2.5 Object Mapping

Each of the sound class is labeled with a list of objects that are usually conceptually connected with the produced sound. Table II shows some of the sound classes along with the associated list of objects related to a particular sound class. After the classifier detects an input sound sample, the “object mapping” module returns the list of associated object that belongs to the detected sound class. This list of

object is given to the commonsense knowledge module to facilitate activity inference.

4.2.6 Commonsense Knowledgebase

Once we get the list of objects involved in recognized sound class, we must define the object involvement probabilities with respect to the activities of our interest. Intuitively, these describe the probability of using the object in that activity state. For example, the activity "eating" always involves food, plate, people and water. Requiring humans to specify these probabilities is time consuming and difficult. Instead, the system automatically determines these probabilities utilizing ConceptNet [33], which is a large corpus of commonsense knowledge. If an activity name co-occurs often with some object name in human discourse, then the activity will likely involve the object in the physical world. Our approach is in the spirit of such manner while we use commonsense knowledgebase corpus to assign a probability value to the object pertaining to a sound class as a model of relatedness in human activity. Thus, for example, if the system detects that the sound cues represent frying pan, sink water, and chop board from consecutive input samples the developed common sense knowledge can infer a cooking activity.

Moreover we have also considered another kind of knowledge namely, ontology of daily-life that contains usual routines of peoples of different roles in terms of their presence at specific locations at specific time. In this initial prototyping three types of users are considered and based on empirical study the ontology of daily life (listed in Table III) of each user type is considered along with common sense knowledge base to perform legitimate inference. For example, if the system detects a sound cue related to eating activity during BN (before noon) time frame on a weekday and if the user is a service-holder, it doesn't log the event because the user is not happened to be present in a kitchen around that time. We also plan to provide the flexibility to personalize such ontology of daily life on the basis of each respective user.

4.2.7 Inference Engine

The system continuously listens to the environment but it records sounds for ten seconds with an interval of ten seconds pause between two recordings. Thus for a minute the system gets three sound clips of equally length (i.e., ten seconds) that serves as the input to the classifier to get three sound classes in one minute. Then object-mapping module provides a list of objects pertaining to the recognized sound classes. In this manner the system gathers a list of objects for every minute. The inference engine works with the list of objects that are gathered in every three minutes. This list of objects is then consulted with the *involvement probabilities* of activities stored in the commonsense knowledge base. An example of such inference is given in the sub-section "A Schematic Solution".

4.2.8 Script Generator

At specific time (e.g., midnight of each day) the whole day's activities are summarized and a daily report is automatically generated and archived in the hard disk for further reference. This report typically contains a consolidation of the frequency of occurrences of major activities of interest. A series of pre-prepared words are used and intelligently strung together to form simple sentences that conform to the basic rules of English grammar. An example of this is given as following:

Daily Report of Mr. Rahman's Activity, Dec 12, 2008
Mr. Rahman woke up at 8:20 am in the morning. He went to bathroom then. He had his breakfast at 9:38 am. He spent most of his time in living room. He watched TV. He had lunch at 2:13 PM. He talked with people. He went out and walked by the roads during evening. He didn't take shower today. At night he talked with someone.

The generated script is then automatically emailed to predefined email addresses. The archival of these daily reports may enable a caregiver or a doctor to review records very quickly and in the process, build a detailed understanding of the subject's daily behavioral patterns.

4.3 A Schematic Solution

The system continuously listens to the environment and captures sound regularly at 10 seconds of interval for 10 seconds of duration. Each captured sound clip is sent to a computer wirelessly where the clip is processed. Thus for a three-minute interval the system gets nine sound clips. These nine sound clips are considered to infer an activity at that moment. Each sound clip is processed by a HMM classifier that outputs a particular sound class. Each sound class actually represents some real-word objects out of which interaction the sound maybe produced. For example, Let's assume that the system receives the following nine sound clips to process and the HMM classifies the nine clips into the following five unique sound classes: "chop_board", "pan_spoon", "water_sink", "plate_clutter", and "plate_knife".

These five classes of sounds are pre-mapped as following,

- "chop_board" → {knife, chop board, kitchen}
- "pan_spoon" → {cooking pan, spoon, stirrer, kitchen}
- "water_sink" → {kitchen, sink, water, dish washer}
- "plate_clutter" → {cup-board, plate, dish washer, rack}
- "plate_knife" → {plate, knife, spoon, fork}

The list of interacting objects is dealt with common

Table 2: Ontology of daily life

Role	Weekday							Weekend						
	EM	M	BN	AN	E	N	LN	EM	M	BN	AN	E	N	LN
Graduate Student	H	H, T	H, U, K,T, R, Tr	U, G, T, P	H, U, R, Tr, K, T, G, P	H, U, R, Tr, K, T, P	H, K, T	H	H, T	H, K,T, R, Tr	G, T, P	H, R, Tr, K, T, G, P	H, R, Tr, K, T, P	H, K, T
Service-Holder	H	H, K, T, R, Tr	O, T	O, T	O, T, H, R, Tr, P	H, P, R, Tr, T	H	H	H, K, T	H, T, P, R, Tr	H, T, P, R, Tr, G	H, T, R, Tr, P	H, P, R, Tr, T	H
Elderly People	H, T	H, T, R, K	H, T, R, Tr, P	H, T, R, Tr, P	H, T, R, Tr, K, P	H	H,T	H, T	H, T, R, K	H, T, R, Tr, P	H, T, R, Tr, P	H, T, R, Tr, K, P	H	H,T

H: Home; O: Office; U: University; K: Kitchen; T: Toilet; G: Gym; R: Road; Tr: Train; P: Public place
EM: Early Morning [03:00 – 05:00], M: Morning [05:00 – 8:00], BN: Before Noon [8:00 – 12:00], AN: After Noon [12:00 – 17:00], E: Evening [17:00 – 20:00], N: Night [20:00 – 01:00], LN: Late Night [01:00 – 03:00]

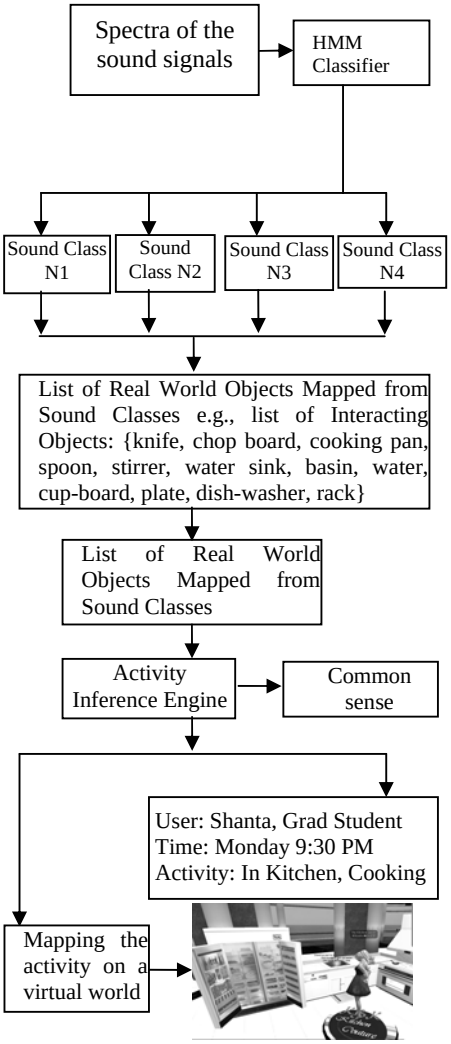

Figure 4: Different steps of the scheme solution

sense knowledge. The common sense knowledge is extracted from ConceptNet [33] and outputs a probabilistic value representing a relationship between an activity and object. In the above example the objects yield a maximum probability of having a relationship with "cooking" activity and the near candidates are "eating", "drinking tea/coffee". From the ontology of daily life the system finds that it is likely to have "cooking" activity at the concerned time (e.g., Table III). Therefore, the activity inference engine considers this event as a legitimate event and logs it by either animating this event on a virtual world or generating a text message. This schematic solution is depicted in Figure 4.

5 DEVELOPMENT CHALLENGES

In order to develop this system we have the following challenges and concerns:

5.1 Environmental sound corpus and features

For simplicity our collected sound corpus is limited to sound cues related to certain genres like, cooking, bathroom activity, human physiological action, outdoor, home, etc. Since the recognition rate is the only performance measure of such system, it is hard to judge how suitable the selected feature is. To solve this problem, it is essential to analyze the feature itself, and measure how good the feature itself is. It is concluded that MFCC may be the best transformation for non-speech environmental sound recognition [29]. This finding provides the essential motivation for us to use MFCC in extracting features for environmental sound classification.

5.2 Computing power of iPhone or HHD

The computing potentials of hand help devices (HHDs) are increasing in a rapid manner with respect to memory and incorporation of full-fledged operating system but they are still inferior to personal computer comparing to the speed of data processing. Running of the core signal-processing task requires high-end computing and therefore the processing should be done by means of external

devices connected to the HHDs through the Bluetooth or wireless data transmission. Therefore in this case a light process to capture environmental sounds after some intervals will be running on the HHDs to be transmitted wirelessly to a server for activity detection.

5.3 Privacy and Social Issues

To manage their exposure, users should be able to disconnect from the virtual worlds at any time or be able to interrupt the sound capturing of their activities. The actual representation of users in the virtual world may be disclosed according to pre-configured user policies or users list. Additional privacy issues are related to the collection of environmental data by means of people carrying devices to different places and data collected about them. We consider important privacy and security challenges related to people-centric sensing [5][6][7].

5.4 Scalability of the solution

Our default approach is to run activity recognition algorithms on the mobile HHDs to decrease communication costs and also to reduce the computational burden on the server. However, we recognize that signal processing based classification algorithms may be too computationally intensive for present HHDs and propose to run the classifier on a back-end server in this case. This may be particularly appropriate during the training phase of a classification model.

6 EXPERIMENTAL RESULTS

The purpose is to test the performance of the system in recognizing the major activities of our interest. The system was trained and tested to recognize the following 17 activities: Listening Music, Watching TV, Talking, Sitting Idle, Cleaning, Sitting idle, Working with PC, Drinking, Eating, Cooking, Washing, Urinating, Exercising, Waiting for Train, Traveling by Train, Shopping, Traveling on Road. As explained earlier, the sound samples recording for each activity was carried out separately. For example, for Listening Music, each subject played a piece of music of his/her choice, with this repeated a number of times for the same individual. The other subjects followed the same protocol and the entire process was repeated for developing the sound corpus for each activity being tested. It is also noted that we have various kinds of sounds (i.e., different sound-clip types) that are grouped together to represent one particular activity. The training data set was formed utilizing a 'leave-one-out' strategy. That is, all the samples would be used for their corresponding models' training except those included in the signal under testing. Hence, each time the models were trained respectively to ensure that the samples in the testing signal were not included in the training data set.

Since each sound clip resolves to a set of objects pertaining to the recognized sound class which is then considered to infer activity and location related to that sound class, we developed perceptual testing methodology to evaluate the system's performance on continuous sound streams of various sound events to infer location and activity. 420 test signals were created, each of which contained a mixture of three sound clips of respective 114 sound types. Since these 420 test signals are the representative sound clues for the 40 sound classes to infer 17 activities, we grouped these 420 test signals into 17 groups according to their expected affinity to a particular activity and location. Ten human (i.e., five male, five female) judges were engaged to listen to the test signals and judge an input signal to infer the activity from the given list of 17 activities (i.e., forced choice judgment) as well as the possible location of that activity from the list of given nine locations of our choice. Each judge was given all the 17 groups of signals to listen and assess. The number of test signals in a group varied from 3 to 6 and each test signal was the result of three concatenated sound clips of same sound type. Therefore a judge had to listen each test signal to infer the location and activity that the given signal seemed most likely to be associated with. In the same way the signals were given to the system to process. For the system the entire group of signals was given at a time to output one location and activity for each group. Since human judges judged each signal individually, in order to compare the result with the system, a generalization on the human assessment was done. The generalization was done in the following manner. A group of signals had at lest more than 3 signals and each of the signals was assigned a location and activity label by the judges. Thus a group of signals obtained a list of locations and activities. We counted the frequencies of location and activity labels for each group assigned by each judge and took the maximum of the respective labels to finally assign the two types of labels (i.e., activity and location) for the group of signals. For each type of label, if more than one labels obtained equal frequency the random choice of the labels are considered. Thus we considered the judges' labels and system's inference with respect to the expected labels for the 17 groups of signals. Recognition results for activity and location are presented in Figure 5 and 6 respectively. The recognition accuracy for activity and location is encouraging with most being above than 66% and 64% respectively. From Figure 5 and 6, we notice that humans are skillful in recognizing the activity and location from sounds (i.e., for humans' the average recognition accuracy of activity and location is 96% and 95% respectively). It is also evident that the system receives the highest accuracy (i.e., 85% and 81% respectively) to detect "traveling on road" activity and "road" location respectively, which is a

great achievement and pioneer effort in this research that no previous research attempted to infer outdoor activities with sound cues. The correct classification of sounds related to activity "working with pc" and location "work place" were found to be very challenging due to the sounds' shortness in duration and weakness in strength, hence the increased frequency for them to be wrongly classified as 'silence' type sound class.

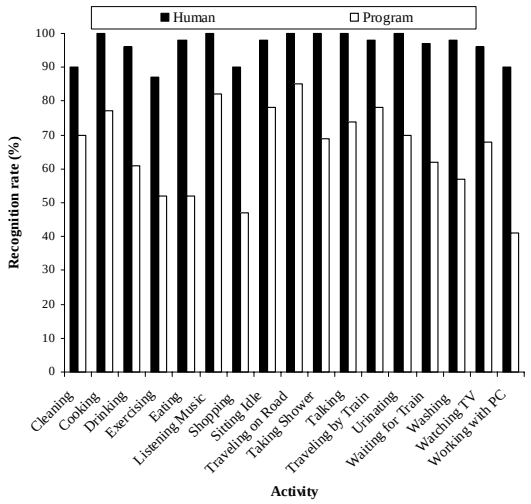

Figure 5: Comparisons of recognition rates for 17 activities of our interest with respect to human judges

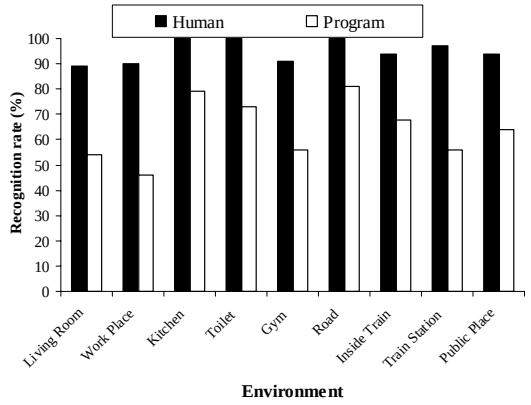

Figure 6: Comparisons of recognition rates for 9 locations of our interest with respect to human judges

7 DISCUSSION AND CONCLUSIONS

In this paper, we described a novel acoustic indoor and outdoor activities monitoring system that automatically detects and classifies 17 major activities usually occur at daily life. Carefully designed HMM parameters using MFCC features are used for accurate and robust sound based activity and location classification with the help of commonsense knowledgebase. Experiments to validate the utility of the system were performed firstly in a constrained setting as a proof-of-concept and in future we plan to perform actual trials involving peoples in the normal course of their daily lives to carry the device running our application that listens to the environment and automatically logs the daily event based on the mentioned approach. Preliminary results are encouraging with the accuracy rate for outdoor and indoor sound categories for activities being above 67% and 61% respectively. We sincerely believe that the system contributes towards increased understanding of personal behavioral problems that significantly is a concern to caregivers or loving ones of elderly people. Besides further improving the recognition accuracy, we plan to enhance the capability of the system to identify different types of human vocalization, which provides useful information pertaining to the mental wellbeing of the subject. We also believe that integrating sensors into the system will also enable acquire better understanding of human activities. The enhanced system will be shortly tested in a full-blown trial on the most needy elderly peoples residing alone within the cities of Tokyo evaluating its suitability as a benevolent behavior understanding system carried by them.

This type of application may have another potential use. Software like Google Earth and Microsoft Virtual Earth map the real world with great accuracy in terms of geographical locations and local features. We believe that the next step is to enable a user to represent real world activities in this kind of virtual world whereby personal avatars will be able to reflect what the real persons are doing in the real world by inferring activities from sound cues. This type of application is a source of fun for young generation while it has lots of potential regarding virtual shopping mall in e-commerce context, easy monitoring for elderly people for the caregivers etc.

ACKNOWLEDGMENT

The authors wish to thank the judges and the peoples who participated in the system's evaluation and collected the environmental sounds to build the sound corpus. This work was supported in part by a grant from Japan Society for the Promotion of Science (JSPS).

8 REFERENCES

[1] A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom Activity Monitoring Based on Sound Jianfeng Chen1 Contact Information," Int. Conf. on PERVASIVE, LNCS 3468/2005, pp. 47-61, 2005.

[2] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, D. HahnelInferring, "Activities from Interactions with Objects," IEEE Pervasive Computing, Vol. 3,No. 4, pp. 50-57, 2004.

[3] A. Temko, C. Nadeu, "Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering," Proc. of IEEE ICASSP'05, pp. 505-508, 2005.
[4] D. Wang, G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley-IEEE Press, 2006.
[5] A. Mihailidis, G. Fernie, and J.C. Barbenel, "The Use of Artificial Intelligence in the Design of an Intelligent Cognitive Orthosis for People with Dementia," Assistive Technology, Vol. 13, No. 1, pp. 23–39, 2001.
[6] D. Wan, "Magic Medicine Cabinet: A Situated Portal for Consumer Healthcare," Int. Sym. On Handheld and Ubiquitous Computing (HUC'99), LNCS Vol. 1707/1999, pp. 352–355, 1999.
[7] T. Barger et al., "Objective Remote Assessment of Activities of Daily Living: Analysis of Meal Preparation Patterns," Presentation at Medical Automation Research Center, Univ. of Virginia Health System, 2002.
[8] Q. Tran, K. Truong, and E. Mynatt, "Cook's Collage: Recovering from Interruptions," Proc. of 3rd Int. Conf. on Ubiquitous Computing (Ubi-Comp), 2001.
[9] A. Glascock and D. Kutzik, "Behavioral Telemedicine: A New Approach to the Continuous Non-intrusive Monitoring of Activities of Daily Living," Telemedicine Journal, Vol. 6, No. 1, pp. 33–44, 2000.
[10] I. Korhonen, P. Paavilainen, and A. Särelä, "Application of Ubiquitous Computing Technologies for Support of Independent Living of the Elderly in Real Life Settings," Proc. of 2nd Int. Workshop Ubiquitous Computing for Pervasive Healthcare Applications (UbiHealth), 2003.
[11] M. Mozer, "The Neural Network House: An Environment That Adapts to Its Inhabitants," Proc. of AAAI Spring Sym. On Intelligent Environments, Tech. AAAI Press, pp. 110–114, 1998.
[12] E. Campo and M. Chan, "Detecting Abnormal Behavior by Real-Time Monitoring of Patients," Proc. of AAAI Workshop Automation as Caregiver, AAAI Press, pp. 8–12, 2002.
[13] V. Guralnik and K. Haigh, "Learning Models of Human Behaviour with Sequential Patterns," Proc. of AAAI Workshop Automation as Caregiver, AAAI Press, pp. 24–30, 2002.
[14] house_n Project: *http://architecture.mit.edu/house_n*
[15] V. Bush, "As we may think", Atlantic Monthly, 1945
[16] B. Rhodes and T. Starner, "Remembrance Agent: A Continuously Running Automated Information Retrieval System," Proc. of Int. Conf. on Practical App. of Intelligent Agents and Multi-Agent Technology, pp. 487-495, 1996.
[17] B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," Proc. of IEEE ICASSP'99, pp. 3037-3040, 1999.
[18] B. Clarkson, K. Mase, and A. Pentland, "The Familiar: A Living Diary and Companion," Proc. of ACM Conf. on Computer–Human Interaction, ACM Press, pp. 271-272, 2001.
[19] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "MyLifeBits: Fulfilling the Memex Vision," Proc. of ACM Multimedia, pp. 235-238, 2002.
[20] A. Fitzgibbon and E. Reiter, "'Memories for Life': Managing Information over a Human Lifetime," UK Computing Research Committee Grand Challenge Proposal, 2003.
[21] S. Vemuri and W. Bender, "Next-Generation Personal Memory Aids," BT Technology Journal, Vol. 22, No. 4, pp. 125-138, 2004.
[22] M. Blum, A. Pentland, G. Troster, et al., "InSense: Internet-Based Life Logging", IEEE Multimedia, Vol. 13, Issue 4, pp.40-48, 2006.
[23] C. Dickie, R. Vertegaal , D. Fono , C. Sohn , D. Chen , D. Cheng , J. S. Shell , and O. Aoudeh, "Augmenting and sharing memory with eyeBlog," Proc. of ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, 2004.
[24] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, "Passive Capture and Ensuing Issues for a Personal Lifetime Store," Proc. of Continuous Archival and Retrieval of Personal Experiences, 2004.
[25] M. Raento, A. Oulasvirta, R. Petit, H. Toivonen, "ContextPhone: A Prototyping Platform for Context-aware Mobile Applications", IEEE Pervasive Computing, Vol. 4, Issue 2, pp. 51–59, 2005.
[26] K. Panu, M. Jani, K. Juha, K. Heikki, M. Esko-Juhani, "ContextPhone: Managing Context Information in Mobile Devices", IEEE Pervasive Computing, Vol. 2, Issue 3, pp. 42–51, 2003.
[27] K. Kim, M. Jung, S. Cho, "KeyGraph-based chance discovery for mobile contents management system Source", Int. Journal of Knowledge-based and Intelligent Engineering Systems, Vol. 11, Issue 5, pp.313-320, 2007.
[28] M. Cowling, "Non-Speech Environmental Sound Classification System for Autonomous Surveillance", Ph.D. Thesis, Griffith University, Gold Coast Campus, School of Information Technology, 2004.
[29] H.G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational Auditory Scene Analysis and Its Application to Robot Audition," Proc. of Int. Conf. on Informatics Research for Development of Knowledge Society Infrastructure (ICKS), pp. 73–80, 2004.
[30] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based Context Awareness-Acoustic Modeling and Perceptual Evaluation," Proc. of IEEE ICASSP'03, Vol. 5, pp. 529-532, 2003.
[31] L. R. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*, PTR Prentice-Hall Inc, New Jersey, 1993
[32] S. Young, *The HTK Book*, User Manual, Cambridge University Engineering Department, 1995
[33] H. Liu, and P. Singh, "ConceptNet: A Practical Commonsense Reasoning Toolkit," BT Technology Journal, Vol. 22, Issue 4, pp. 211-226, 2004.