Easy Living in the Virtual World: A Noble Approach to Integrate Real World Activities to Virtual Worlds

Mostafa Al Masum Shaikh[1], Helmut Prendinger[2], Keikichi Hirose[1], Ishizuka Mitsuru[1]
[1]*Department of Information and Communication Engineering, University of Tokyo, Japan*
[2]*Digital Contents and Media Sciences Research Division, National Institute of Informatics, Japan*
{almasum, hirose}@gavo.t.u-tokyo.ac.jp, helmut@nii.ac.jp, ishizuka@i.u-tokyo.ac.jp

Abstract

Virtual worlds like "Second Life" are popular graphical representations of real (and fictitious) places, which are inhabited by real people in the form of personal avatars. The existence of people in these worlds is either (1) as avatars manipulated by users (to make them walk, fly, chat, etc), or (2) as pre-scripted agents, called "bots", which are programmed to display some predefined behavior in the virtual world. Research that aims to bridge real life and these virtual worlds to simulate virtual living, while challenging and promising, is currently rare. Only very recently the mapping of real-world activities to virtual worlds has been attempted by processing multiple sensors data along with inference logic for real-world activities. Detecting or inferring human activity using such simple sensor data is often inaccurate and insufficient. Hence, this paper explains to infer human activity from environmental sound cues and common sense knowledge, which is an inexpensive alternative to other sensors (e.g., accelerometers). We discuss the challenges to implement such a system from the signal processing and agent based system point of view. 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 in virtual worlds.

1. Introduction and Motivation

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 suggest that someone is on the road, mob sound along with sound of cutleries support the inference that the person is in a restaurant and so on.

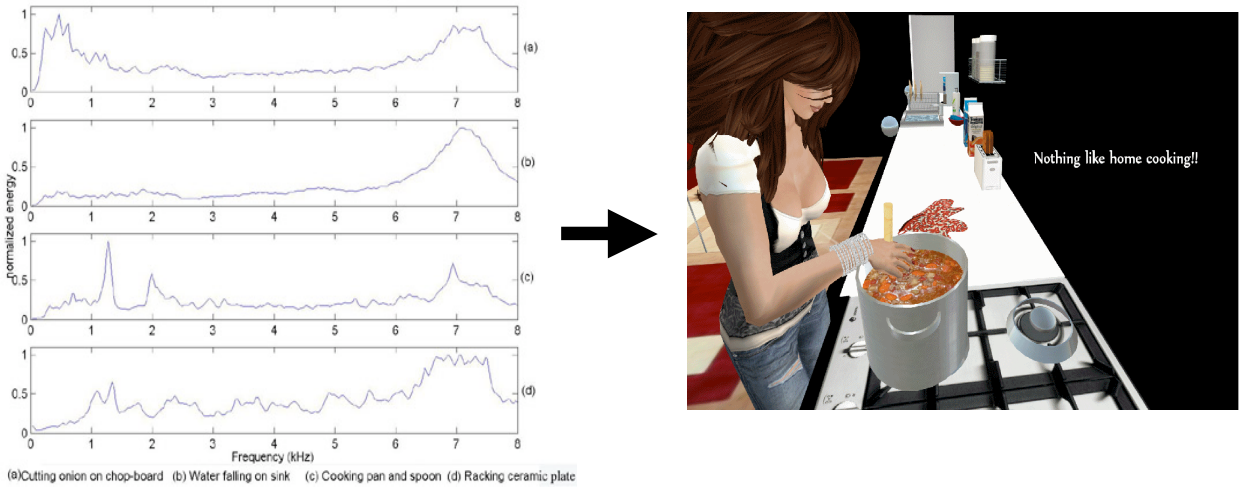
Many sources of information for sensing the environment as well as activity are available [1][2][3]. In this paper, we consider two objectives: first, sound-based context awareness, where the decision is based merely on the available acoustic information at the surrounding environment of the user and second, automatic virtual living, where the detected sound context infers an activity to be mapped with a virtual world activity. 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. Virtual living is a concept of living in a virtual world with a resident population of millions of real people from around the world. Each person is represented by an avatar that represents their chosen digital persona. A user will be able to walk, "teleport" or even fly to thousands of exciting 3D locations and can also use voice and text chat to communicate with other real people from around the world. In this manner, the animating behavior and life-likeness of the avatars as well as the representation of the real-world environments in the virtual worlds render the idea of "virtual living". Would there be a technology to synchronize a user's (e.g., an elderly people) real world with the user's virtual world, the concept of "virtual living" might be applied to monitor the user's well-being or abnormality by someone (e.g., relatives or care-givers) who cares about the user. Nourishing such a vision in mind we apply the concept of AED to perform automatic generation of life-log which is represented as activities in the virtual world.

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 virtual world activities 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. We envision that with the proliferation of computing power of hand held devices (HHD), availability of internet connectivity and improvements in communication

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.196

technologies such ambient communication to the virtual world will find a universal place in our daily lives and allow us to create a vivid and intelligent online social network. Let's consider the following scenario to clarify our motivation.

Scenario: Sami, Anny, Harry, and Silvia have become friends in a virtual world but in real life they live at different corners of the world. They often login to a virtual world and frequently update their status to let others know what they are doing just for fun. They use "Second Life" [5] to interact with each other in the virtual world. They are looking forward to use a HHD (e.g., iPhone) that can automate the process to update their real-world status. Let's assume that on the HHD they have installed our system that can capture and allow processing of environmental sounds at some time interval. The processed sound cues are used together with common sense knowledge to infer the present activity and automatically reflects the real-world activity on the virtual world. For example, while Silvia is cooking in real-world (i.e., the sound cues like cutting onion on chopping board, water falling on sink, cooking pan and spoon, arranging plates as indicated in Figure 1 are generate), her friends see her moving around the kitchen in the Second Life, as indicated at the right-side of Figure 1.



(a)Cutting onion on chop-board (b) Water falling on sink (c) Cooking pan and spoon (d) Racking ceramic plate

Figure 1. Cooking activity is inferred from the sound cues produced in kitchen and the activity is mapped to "cooking" activity to reflect that Silvia is also cooking in her virtual world while she is cooking in her real-world kitchen

The paper is organized as follows: Section 2 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 3. Section 4 explains the experimental setup, the results obtained by the system as well as user evaluations. Conclusions are presented in Section 5.

2. Background

A number of researchers have investigated to infer activities of daily living (ADL). In [6] authors have successfully used cameras and a bracelet to infer hand washing. The authors of [7] used radio-frequency-identification (RFID) tags functionally as contact switches to infer when users took medication. The system discussed in [8] used contact switches, temperature switches, and pressure sensors to infer meal preparation. Authors of [9] used cameras to infer meal preparation. In [10] 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 [11]. Author of [12] used 13 sensors to infer home energy use, focusing on the heating-use activity. Motion detectors to infer rough location were used in [13]. Several sensors like motion sensors, pressure pads, door latch sensors, and toilet flush sensors to infer behavior are reported in the system described in [14]. 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 [15] 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 [6][10][13][14]. 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.

A similar approach to automatic virtual living is automatic life-logging. The idea of a "life-log" or a personal digital archive is a notion that can be traced back at least 60 years [16]. Since then a variety of modern projects have spawned such as the *Remembrance Agent* [17], *the Familiar* [18][19], *myLifeBits* [20], *Memories for Life* [21] and *What Was I Thinking* [22]. In [23] 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. 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. But specifically speaking, a user cannot automatically mirror/reflect his current movements, activities or surrounding environment (e.g., park, shopping mall, etc.) in his real life to the virtual life of his avatar. Only very recently [24] the mapping of real-world activities to virtual worlds has been attempted by processing multiple sensors data along with inference logic for real-world activities. Detecting or inferring human activity using such simple sensor data is often inaccurate and insufficient. Moreover deploying a sophisticated ubiquitous sensor network at outdoor environment is often expensive and not feasible.

Our work differs from others in four key ways. First, we utilize environmental sounds cues to infer regarding the interactions with objects or environment instead of sensor or camera data. 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 Hidden Markov Model (HMM) based recognizer. Finally, the system can be used as a life-logging agent as well as a mean to seam someone's real-world with a virtual world.

3. The System

The goal of the system is to detect activities of daily living (e.g., laughing, talking, travelling, cooking, sleeping, etc.) and situational aspects of the person (e.g., inside a train, at a park, at home, at school, etc.) by processing environmental sounds. For example, 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'.

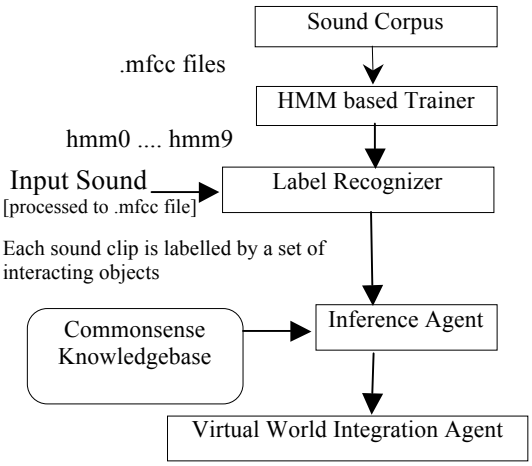


Figure 2. The System Architecture

3.1. System Architecture

Because of their ubiquity we plan to use hand held devices (e.g., portable computer or smart phone) to deploy this application that will capture environmental sound at some intervals to be processed. According to the system's architecture given in Figure 2, environmental sound signals are processed through signal processing and then each input sound sample is recognized as a set of object labels by HMM based label recognizer. The detected object list and commonsense knowledge regarding human activity, object interaction, along with temporal information (e.g., morning, noon etc.) are utilized by the inference agent to infer both activity and surrounding environment of the user. Recognized real-world activity and location are then mapped to the virtual world of the user by a scripting language. In the following sections the system's components are described with necessary examples.

3.2. 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 of 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, Tokyo University and apartments in Tokyo. 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 1 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 recorded simultaneously. 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. According to the location and activities of our interest mentioned in Table 1, we have collected 114 types of sounds. Each of the sound types has 15 samples of varying length from 10 to 25 seconds.

location	Activities
Living Room	Listening Music, Watching TV, Talking, Sitting Idle, Cleaning (vacuum-cleaning)
Work Place	Sitting idle, Working with PC, Drinking
Kitchen	Cleaning, Drinking, Eating, Cooking
Toilet	Washing, Urinating
Gym	Exercising
Train Station	Waiting for Train
Inside Train	Travelling by Train
Public Place	Shopping, Travelling on Road
On the Road	Traveling on Road

Table 1: List of locations and activities of our interest

3.3. HMM based Trainer

3.3.1. Sound Clip Annotation: We have listed 63 objects that are used in the annotation to denote their pertinence in a given sound sample. During the annotation an annotator opened a sample (.wav formatted) sound file by WaveSurfer, an open source tool for sound visualization and manipulation, setting the annotation configuration as "HTK Transcription". HTK is a speech recognition toolkit based on Hidden Markov Model. In our case, a sound sample usually contains different kind of sounds eventually produced by different kind of objects. An annotator selected a particular portion of the sample sound by listening that represented any of the 63 listed objects and thus that region of the signal is annotated by assigning a short name (as shown in Table 2) of a particular object which was producing or associated with that sound portion. For example if a sound portion is produced by a "plate" object, that portion is annotated as "plt".

Object	Tag	Object	Tag
ambulance	amb	cash register	reg
announcement	ann	CD player	cdp
basin	bsn	chopping board	chp
bicycle	ckl	computer	com
blender machine	bln	computer keyboard	kbd
boiling	boi	computer mouse	cms
bottle	btl	cracking	crk
bowl	bwl	drawer	dwr
bus	bus	dumbbell	dbl
car	car	electric train	trn

Table 2: Object names and their Tag used for labelling

If the annotator found that there was an overlapping sounds of more than one objects in a selected portion, in that case the following was considered.

- Initially it was tried if the selection can be shortened to represent the sound portion associated with a single object as mentioned above.
- If the selected audio portion cannot be represented by one object due to auditory distinctness of more than one sounds produced by simultaneous interaction of more than one objects, an annotator was allowed to denote maximum of two objects to annotate such complex sound. For example, if a selected audio portion was found representing both "human coughing", "music", and "tv program" sound, in this case an annotator tagged this portion of the sound as either "ppl_mus" or "ppl_tel" or "ppl_tel" or "mus_tel" according to the prominence of the sounds of the pertaining objects.

3.3.2. Training Features: It is obvious that simple frequency characterization would not be robust enough to produce good classification results. To find representative

features, previous study [25] 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 [26,27]. These findings provide the essential motivation for us to use MFCC in extracting features for environmental sound classification.

The input signal is first pre-emphasized with the FIR filter 1, -0.97z-1. MFCC analysis is performed in 25 ms windowed frames advanced every 10 ms. For each signal frame, the following coefficients are extracted as a feature vector:

- The 12 first MFCC coefficients [c1,…, c12]
- The "null" MFCC coefficient c0, which is proportional to the total energy in the frame
- 13 "Delta coefficients", estimating the first order derivative of [c0, c1,…, c12]
- 13 "Acceleration coefficients", estimating the second order derivative of [c0, c1,…, c12]

Altogether, a 39 coefficient vector is extracted from each signal frame window.

3.4. Label Recognizer

Training is performed in the training set that consists of the recordings and their associated class (i.e., object) labels. Typically, the HMM parameters are iteratively optimized with the Baum-Welch algorithm that finds a local maximum of the maximum likelihood (ML) objective function.

We modeled each sound using a left-to-right 88-state (63 for simple object tag + 25 for complex object tag) continuous-density HMM without state skipping using HTK Toolkit [28]. 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. For classification, continuous HMM recognition is used. The grammar (denoted partially) used is as follows:
(<alr|amb|ann|bsn|ckl|bln|boi|btl|bwl|bus|car|reg|cdp|…|flu|fwt|sus|srb|…|shr|sng|snk|tap|wnd…|ppl_tv|…|mus_ppl|ppl_tv| … |wtr_ppl>), which means that there is no predefined sequence for all the activities and each label may be repeated many times at any sequence.

3.5. Commonsense Knowledgebase

Once we get the list of objects involved in recognized sound samples, we must define the object involvement probabilities with respect to the activities of our interest. 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 has utilized a technique adopted from Semantic Orientation (SO) [29,30] employing NEAR search operator of AltaVista's web search result.

List of objects, $O = \{O_1, O_2, \ldots O_K\}$ (K=63)
List of locations, $L = \{L_1, L_2, \ldots L_M\}$ (M=9)
List of activities, $A = \{A_1, A_2, \ldots A_N\}$ (N=17)
Each location is represented by a set of English synonym words. $WL_i = \{W_1, W_2, \ldots, W_P\}$. For example, L_1= "kitchen" and it is represented by, $W_{kitchen}$ = {"kitchen", "cookhouse", "canteen", "cuisine"}
$SA(O_i|L_j)$ = Semantic Associative value representing the object O_i to be associated with location L_j
$SA(O_i|A_j)$ = Semantic Associative value representing the object O_i to be associated with activity A_j
The formulae to get the *SA* values are,

$$SA(O_i \mid L_j) = \log_2\left(\frac{\prod_{W \in WL_j} hits(O_i \text{ NEAR } W)}{\prod_{W \in WL_j} \log_2(hits(W))}\right) \quad (1)$$

$$SA(O_i \mid A_j) = \log_2\left(\frac{hits(O_i \text{ NEAR } A_j)}{\log_2(hits(A_j))}\right) \quad (2)$$

The obtained values support the concept that if an activity name and location 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 these obtained values as commonsense knowledgebase to assign a semantic associative value to the object pertaining to a sound sample as a model of relatedness in human activity. Thus, for example, if the system detects that the sound samples represent frying, saucepan, water sink, water, and chopping board from consecutive input samples the commonsense knowledge usually infers a cooking activity located in kitchen.

3.6. Inference Agent

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 equal length (i.e., ten seconds) that serves as the input to the classifier to process three sound samples 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 each 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 Semantic Associative (SA) value of the activities and locations stored in the commonsense knowledgebase.

As an example, for a two-minute interval the system gets six sound clips. These six sound clips are considered to infer an activity at that moment. Each sound clip is processed by HMM based recognizer that performs

continuous recognition of desired labels. For example, let's assume that the system receives the following six sound clips and the HMM recognized the following objects.

- clip 1→ {knife, chopping board, people}
- clip 2→{knife, spoon, chopping board, people}
- clip 3→{water sink, water, wind, male voice}
- clip 4→{spoon, frying, people, wind}
- clip 5→{frying, saucepan, spatula, people}
- clip 6→{frying, saucepan, spoon, water}

The unique list of objects obtained from the clips, U= {chopping board, frying, knife, male voice, people, saucepan, spatula, spoon, water sink, water, wind}. This list of objects is dealt with common sense knowledge by obtaining a normalized SA values for each Activity and Location. In the above example the objects yield a maximum SA value of having a relationship with "cooking" activity in "kitchen" location and the near candidates are "eating", "drinking tea/coffee" as activities.

3.7. Virtual World Integration Agent

The agent receives location and activity information from the inference agent and on the basic of the input this agent invokes necessary pre-defined scripts. The scripts usually contains the necessary commands to move the person's avatar to a specific location in the virtual world (i.e., Second Life) and do a sequence of animations while interacting with the virtual world objects that eventually represents the person's real-world activity in that virtual world. At present this agent integrates to the virtual world with the kitchen related activities only and in future we plan to extend the mapping with other activities and locations. Screenshots of the mapped activities in Second life is given in Figure 3.

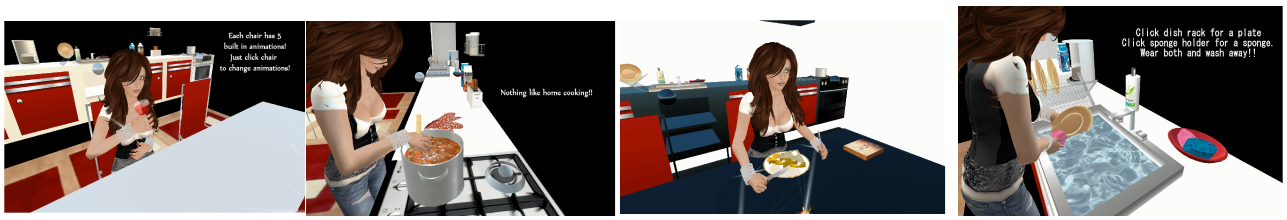


Figure 3. Kitchen activities like "Drinking", "Cooking", Eating" and "Cleaning" are represented in Second Life

4. Test and Evaluation

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, Travelling by Train, Shopping, Travelling 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. 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 clip which is considered to infer activity and location of the user at that time, 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 63 objects 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 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 input 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 least 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 4 and 5 respectively.

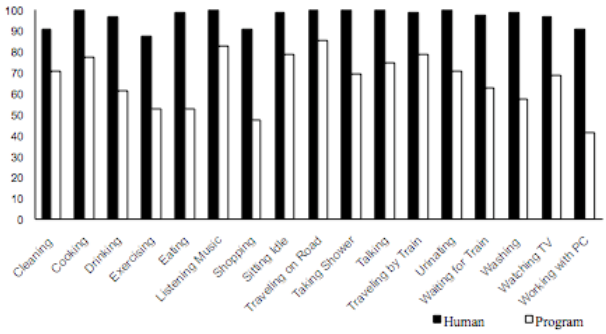


Figure 4. Comparisons of recognition rates for 17 activities of our interest with respect to human judges

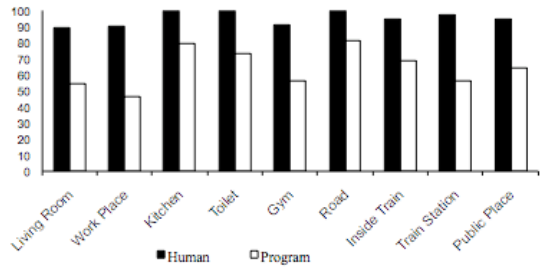


Figure 5. Comparisons of recognition rates for 9 locations of our interest with respect to human judges

The recognition accuracy for activity and location is encouraging with most being above than 66% and 64% respectively. From Figure 4 and 5, 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 'wind' type object recognition.

5. Conclusion

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 detects 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 loved ones of elderly people. In future we plan to integrate different sensors (e.g., pressure and/or proximity sensors) into the system and conduct experiments to acquire better understanding of human activities. We envision that the enhanced system will be tested on the neediest elderly peoples residing alone within the cities of Tokyo to monitor their living in an unobtrusive manner. Enabling a user to represent real world activities to a virtual world and thereby continue the concept of "virtual living" is surely a source of excitement for young generation but it can come across potential usages like virtual shopping mall for product or service advertisement, collaborative learning, easy monitoring for elderly people for the caregivers and so on.

References

[1] Kam, A. H., Zhang, J., Liu, N., and Shue, L., 2005. Bathroom Activity Monitoring Based on Sound. In *PERVASIVE'05, 3rd International Conf. on Pervasive Computing.* Germany, LNCS 3468/2005, pp. 47-61.

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

[3] Temko, A., Nadeu, C., 2005. Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering. In *ICASSP'05,* pp. 505-508.

[4] Wang, D., and Brown, G., 2006. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE

[5] Linden Research Inc. SecondLife. http://secondlife.com/

[6] Mihailidis, A., Fernie, G., and Barbenel, J.C., 2001. 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.

[7] Wan, D., 1999. Magic Medicine Cabinet: A Situated Portal for Consumer Healthcare. In *HUC'99, 1st Int'l Symp. Handheld and Ubiquitous Computing*, LNCS 1707, Springer-Verlag, pp. 352–355

[8] Barger T., Alwan, M., Kell, S., Turner, B., Wood, S., and Naidu, A., 2002. Objective Remote Assessment of Activities of Daily Living: Analysis of Meal Preparation Patterns. *Medical Automation Research Center, Univ. of Virginia Health System*.

[9] Tran, Q., Truong, K., and Mynatt, E., 2001.Cook's Collage: Recovering from Interruptions. *Demo at Ubi-Comp'01, 3rd Int'l Conf. Ubiquitous Computing.*

[10] Glascock A., and Kutzik, D., 2000. Behavioral Telemedicine: A New Approach to the Continuous Nonintrusive Monitoring of Activities of Daily Living, *Telemedicine Journal*, Vol. 6, No. 1, pp. 33–44.

[11] Korhonen, I., Paavilainen, P., and Särelä, A., 2003. Application of Ubiquitous Computing Technologies for Support of Independent Living of the Elderly in Real Life Settings. In *UbiHealth'03, 2nd Int'l Workshop Ubiquitous Computing for Pervasive Healthcare Applications*

[12] Mozer, M., 1998. The Neural Network House: An Environment That Adapts to Its Inhabitants. In *AAAI Spring Symposium, Intelligent Environments, tech. report SS-98-02*, AAAI Press, pp. 110–114.

[13] Campo E., and Chan, M., 2002. Detecting Abnormal Behavior by Real-Time Monitoring of Patients. In *AAAI Workshop Automation as Caregiver*, AAAI Press, pp. 8–12

[14] Guralnik V., and Haigh, K., 2002. Learning Models of Human Behaviour with Sequential Patterns. In *AAAI Workshop Automation as Caregiver*, AAAI Press

[15] MIT house_n Project, http://architecture.mit.edu/house_n

[16] V. Bush, "As we may think", Atlantic Monthly, 1945

[17] B. Rhodes and T. Starner, "Remembrance Agent: A Continuously Running Automated Information Retrieval System," Proc. 1st Int'l Conf. Practical App. of Intelligent Agents and Multi-Agent Technology, 1996, pp. 487-495.

[18] B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," Proc. IEEE Int'l Conf. Acoustics, Speech, and SignalProcessing, IEEE CS Press, vol. 6, 1999, pp. 3037-3040

[19] B. Clarkson, K. Mase, and A. Pentland, "The Familiar: A Living Diary and Companion," Proc. ACM Conf. Computer–Human Interaction, ACM Press, pp. 271-272.

[20] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "MyLifeBits: Fulfilling the Memex Vision," Proc. ACM Multimedia, ACM Press, 2002, pp. 235-238.

[21] A. Fitzgibbon and E. Reiter, "'Memories for Life': Managing Information over a Human Lifetime," UK Computing Research Committee Grand Challenge proposal, 2003.

[22] S. Vemuri and W. Bender, "Next-Generation Personal Memory Aids," BT Technology J., vol. 22, no. 4, 2004.

[23] M. Blum, A. Pentland, G. Troster, et al., "InSense: Internet-Based Life Logging", IEEE Multimedia vol. 13, Issue 4, pp.40-48, 2006

[24] Mirco, M., Emiliano, M., Nicholas D. L., Shane B. E., Tanzeem, C., Andrew T. C., 2008. The Second Life of a Sensor: Integrating Real-world Experience in Virtual Worlds using Mobile Phones. In Proceedings of the Fifth Workshop on Embedded Networked Sensors (Charlottesville, Virginia, June 2-3, 2008). HotEmNets 2008

[25] M. Cowling, Non-Speech Environmental Sound Recognition System for Autonomous Surveillance, Ph.D. Thesis, Griffith University, Gold Coast Campus (2004)

[26] H.G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational Auditory Scene Analysis and Its Application to Robot Audition," International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS), (2004), 73–80

[27] 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. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 5, (2003), pp. 529-532

[28] S. Young, The HTK Book, User Manual, Cambridge University Engineering Department, 1995

[29] Hatzivassiloglou, V. and McKeown, K. R., 1997. Predicting the Semantic Orientation of Adjectives. In 35th annual meeting on ACL, pp.174-181

[30] Grefenstette, G., Qu, Y. Evans, D., and Shanahan, J., 2004. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In Computing Attitude and Affect in Text: Theory and Applications, eds. J. Shanahan, Y. Qu, and J. Wiebe, 93-107. The Information Retrieval Series Vol. 20, Netherlands: Springer Verlag