CONTEXT AWARENESS USING ENVIRONMENTAL SOUND CUES AND COMMONSENSE KNOWLEDGE

Mostafa Al Masum Shaikh, Keikichi Hirose
Dept. of Information and Communication Engineering,University of Tokyo, 7-3-1 Hongo, Bunkyu Ku, Tokyo, Japan

Helmut Prendinger
Digital Contents and Media Sciences Research Div., National Institute of Informatics,Chiyoda Ku, Tokyo, Japan

Keywords: Acoustic Event Detection, Context Awareness, Activity Detection, Sound Cues, Auditory Scene Analysis, Commonsense Knowledge, Ambient Communication, Life-Logging.

Abstract: Detecting or inferring human activity (e.g., an outdoor activity) by analyzing sensor data is often inaccurate, insufficient, difficult, and expensive. Therefore, this paper explains an approach to infer human activity and location considering the environmental sound cues and commonsense knowledge of everyday objects usage. Our system uses mel-frequency cepstral coefficients (MFCC) and their derivatives as features, and continuous density hidden Markov models (HMM) as acoustic models. Our work differs from others in three key ways. First, we utilize both indoor and outdoor environmental sound cues which are annotated according to the objects pertaining to the sound samples to build the idea regarding sounds and the objects which produce that particular sound. Second, use of portable microphone instead of having a fixed setup of an array of microphones to capture environmental sound we can also infer outdoor environments like being on the road, in a train station, etc., which previous research was limited to perform. Thirdly, our model is easy to incorporate new set of activities for further needs by adding more appropriately annotated sound clips and re-training of the HMM based recognizer. A perceptual test is made to study the human accuracy in the task and to obtain a baseline for the assessment of the performance of the system. Though the direct comparison of the system's performance to human performance is somewhat worse but the preliminary results are encouraging with the accuracy rate for outdoor and indoor sound categories for activities being above 67% and 61% respectively.

1 INTRODUCTION

Although speech is the most informative acoustic event, other kind of sounds may also carry useful information regarding the surrounding environment. Many sources of information for sensing the environment as well as activity are available (Kam et al., 2005; Temko and Nadeu, 2005). In this paper our primary objective is, sound-based context awareness, where the decision is based merely on the available acoustic information at the surrounding environment of the user where the system detects particular sounds and an activity is inferred thereby at that particular time. Acoustic Event Detection (AED) is a recent sub-area of computational auditory scene analysis (Wang and Brown, 2006) 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. 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. 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 limited number of locations of our interest and the activities to recognize at a particular location.

Al Masum Shaikh M., Hirose K. and Prendinger H. (2009).
CONTEXT AWARENESS USING ENVIRONMENTAL SOUND CUES AND COMMONSENSE KNOWLEDGE.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 193-196
DOI: 10.5220/0002230901930196

2 ACOUSTIC MEASUREMENT

The sound data set, the annotation of the sound samples, and the features used to train the HMM based recognizer.

2.1 Date Set

A number of male and female subjects collected 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 (ICR-PS380RM) and signals were recorded as Stereo, 44.1 KHz, .wav formatted files. 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 second to 25 seconds.

Table 1: List of locations and activities of our interest.

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

2.2 Annotation

A list of 63 objects is used in the annotation to denote their pertinence in a given sound sample. 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 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". If an annotator found that there was an overlapping sounds of more than one objects in a selected portion, 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.

2.3 Feature

According to (Okuno et al., 2004; Eronen et al., 2003), it is concluded that MFCC might be the best transformation for non-speech environmental sound recognition. 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.

2.4 The HMM based Recognizer

We have chosen an approach based on HMM as the model has a proven track record for many sound classification applications (Okuno et al., 2004). We modelled 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. Each HMM state was composed of two Gaussian mixture components and trained in eight iterative cycles. The recognition 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 and each label may be repeated many times at any sequence.

3 SYSTEM DESCRIPTION

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.

3.1 System Architecture

According to the system's architecture given in Figure 1, sound signal is passed through a signal processing and each input sound sample is detected as a set of object labels by the 88 trained HMM classifiers. Based on the detected list of objects and commonsense knowledge regarding human activity, object interaction, along with temporal information (e.g., morning, noon etc.) are applied to infer both the activity and the surrounding environment of the user.

3.2 Notes on Commonsense Knowledge

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. Requiring humans to specify these probabilities is time consuming and difficult. Instead, the system has utilized a technique adopted from Semantic Orientation (SO) (Hatzivassiloglou and McKeown, 1997), 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. 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.

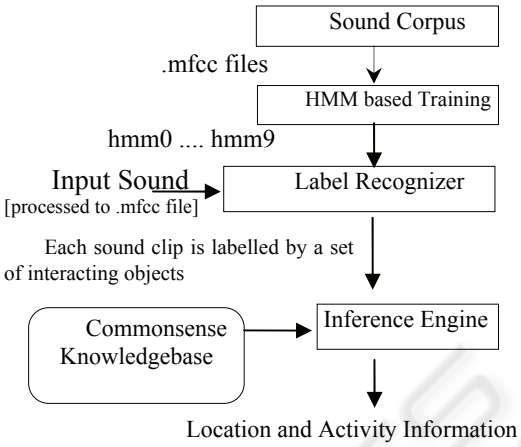

Figure 1: The Architecture of the System.

3.3 Inference Engine

For example, the system recorded six sound clips of 10 second long in two minutes to infer an activity and location. The HMM model 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 commonsense 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.

4 EXPERIMENT RESULT

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. 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. Recognition results for activity and location are presented in Figure 2 and 3 respectively.

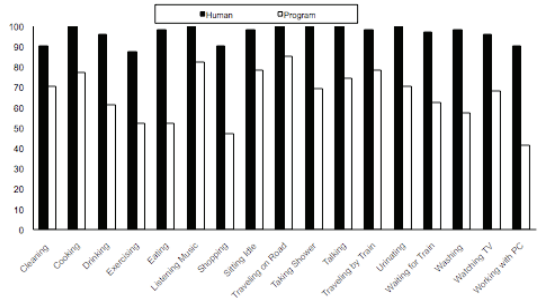

Figure 2: Comparisons of recognition rates for 17 activities of our interest with respect to human judges.

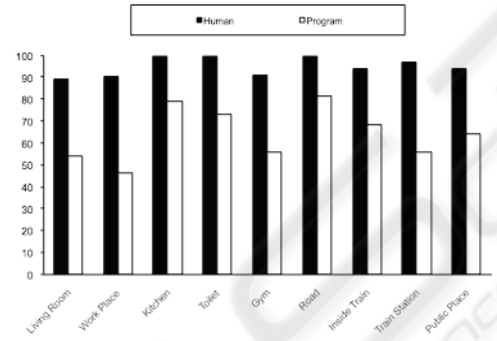

Figure 3: Comparisons of recognition rates for 9 locations of our interest with respect to human judges.

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 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. Preliminary results are encouraging with the accuracy rate for outdoor and indoor sound categories for activities being above 67% and 61% respectively. We 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 neediest elderly peoples residing alone within the cities of Tokyo evaluating its suitability as a benevolent behavior understanding system carried by them.

REFERENCES

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.

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.

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

Okuno, H.G., Ogata, T., Komatani, K., and Nakadai, K., 2004. Computational Auditory Scene Analysis and Its Application to Robot Audition. In *International Conference on Informatics Research for Development of Knowledge Society Infrastructure,* pp., 73–80

Eronen, A., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J., 2003. Audio-based Context Awareness-Acoustic Modeling and Perceptual Evaluation. In *ICASSP '03, Int'l Conference on Acoustics, Speech, and Signal Processing*, pp. 529-532

Hatzivassiloglou, V. and McKeown, K. R., 1997.

Predicting the Semantic Orientation of Adjectives. In 35th annual meeting on ACL, pp.174-181