J Ambient Intell Human Comput
DOI 10.1007/s12652-012-0122-2

ORIGINAL RESEARCH

Wearable sensor-based human activity recognition from environmental background sounds

Yi Zhan · Tadahiro Kuroda

Received: 20 October 2011 / Accepted: 26 March 2012

Abstract Understanding individual's activities, social interaction, and group dynamics of a certain society is one of fundamental problems that the social and community intelligence (SCI) research faces. Environmental background sound is a rich information source for identifying individual and social behaviors. Therefore, many power-aware wearable devices with sound recognition function are widely used to trace and understand human activities. The design of these sound recognition algorithms has two major challenges: limited computation resources and a strict power consumption requirement. In this paper, a new method for recognizing environmental background sounds with a power-aware wearable sensor is presented. By employing a novel low calculation one-dimensional (1-D) Haar-like sound feature with hidden Markov model (HMM) classification, this method can achieve high recognition accuracy while still meeting the wearable sensor's power requirement. Our experimental results indicate an average recognition accuracy of 96.9 % has been achieved when testing with 22 typical environmental sounds related to personal and social activities. It outperforms other commonly used sound recognition algorithms in terms of both accuracy and power consumption. This is very helpful and promising for future integration with other sensor(s) to provide more trustworthy activity recognition results for the SCI system.

Y. Zhan (✉) · T. Kuroda
Department of Electronics and Electrical Engineering, Keio University, Hiyoshi 3-14-1, Kouhoku-ku, Yokohama-shi, Kanagawa-ken 223-8522, Japan
e-mail: yizhan@kuroda.elec.keio.ac.jp

T. Kuroda
e-mail: kuroda@elec.keio.ac.jp

Keywords Social and community intelligence · Digital footprint · WSNs · Sound recognition · Haar-like feature · HMM

1 Introduction

The past decade witnessed rapid development in basic Internet, communications theories and in some newly emerging technologies, such as wireless sensor networks (WSNs) (Culler et al. 2004), wearable sensing and computation (Bonfiglio and Rossi 2011); these technologies are gradually entering an applicable stage where they can be used for various purposes. The technological advancement has recently led to the emergence of a brand-new research area: social and community intelligence (SCI) (Zhang et al. 2011). With the SCI technology, individual's behavior patterns, social interactions and community dynamics inside a certain society can be explored, collected, analyzed, and well managed. In addition, applications of SCI technology will be helpful to enrich our life contents and improve our society's efficiency.

"Community detection and social behavior analysis" and "socially-aware computing" are two major topics of SCI research (Zhang et al. 2011; Pentland 2005). Reliable detection as well as comprehension of individual activities and person-to-person interactions in a certain society is a fundamental problem encountered in SCI research. Information pertaining to personal and social activities can be detected and traced by the so-called "digital footprint" left behind by people while interacting with cyber-physical spaces (Zhang et al. 2011; Guo et al. 2011b). With various sensing capabilities sensors embedded into the mobile phones, wearable devices and WSNs technology's involvement, people's daily information can be digitalized

Published online: 26 May 2012

and perceived. This will facilitate the understanding of "digital footprints" information of individual and social interactions inside an organization (Pentland 2005; Choudhury 2004; Laibowitz et al. 2006; Yano et al. 2009; Yano et al. 2008; Guo et al. 2011a). Among these sensing media, acoustic sound is a rich information source for identifying the individual and social behaviors.

In this study, a sound sensor embedded in the wearable sensor node (Yano et al. 2009; Nishimura et al. 2008) shown in Fig. 1, is utilized to recognize environmental background sounds happening around people. These sounds contain useful information to understand what activities an individual does. They can also act as a social "bridge" among people. By recognizing these sounds continuously for a whole day, the people's daily activities log can be established accordingly. This log indicates personal and social interactive information. Many SCI applications can be created based on the log information. For example, it is very helpful to establish household medical systems such as remote monitoring and diagnosis for patients, and individual's daily physical and health monitoring at home. Log information can also assist in understanding social interactions in a particular group or society; for example, the working status of employees and their efficiency in offices or working places (Yano et al. 2009). A good example for a group dynamics application is to determine a common favorite individual in the group. The utilized wearable device "UberBadge" mounted on each participant of the group contents sound sensor (Pentland 2005; Laibowitz et al. 2006).

Energy efficiency plays an important role for mobile and wearable devices in the SCI system (Bonfiglio and Rossi 2011). In order to reveal individual activities and social interactions, most front-end sensing units are mobile and portable, for example mobile phones, PDAs and wearable devices. In addition, power supply for these devices is an energy limited battery, unlike a DSP and FPGA board fitted with a power adaptor. Conventional sound recognition and acoustic signal processing algorithms that can be executed on the DSP or FPGA (Dong et al. 2007; Veitch et al. 2011) platforms may not perform well on our wearable sensor node. Therefore, a major challenge for this research is development of a new sound recognition algorithm for achieving high accuracy with low calculation cost to meet the energy requirement.

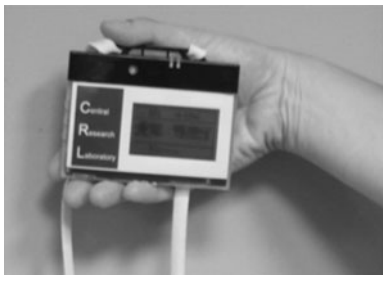

Fig. 1 Our power-aware wearable sensor node consisting of embedded sound, acceleration, IR sensor and other sensors with a size of an ID card (3.86 inch × 2.87 inch × 0.35 inch)

Some environmental sound recognition researches have been reported previously (Chen et al. 2005; Goldhor 1993; Ma et al. 2006; Chu et al. 2009; Cowling and Sitte 2003; Peltonen et al. 2002; Dong et al. 2007; Bharatula et al. 2005). At the feature extraction stage, conventional state-of-the-art Mel-frequency cepstrum coefficients (MFCCs) filtering is used to extract the sound feature and obtain good recognition accuracy (Chen et al. 2005; Goldhor 1993; Ma et al. 2006; Dong et al. 2007). However, computationally expensive FFT is calculated before entering a bank of Mel-scale filters in the extraction flow. This increases the calculation complexity of sound feature extraction. At the classification stage, performance of the Gaussian mixture model (GMM), support vector machine (SVM), Linde–Buzo–Gray algorithm (LBG), *k*-means, and hidden Markov Model (HMM) classifiers has been studied and compared in work (Cowling and Sitte 2003). Through the work, we have learned that the HMM (Ma et al. 2006; Rabiner 1989) classifier can achieve high recognition accuracy with an acceptable increment of calculation cost compared with other classifiers.

In this paper, a novel Haar + HMM algorithm is proposed to recognize the environmental background sounds. Haar-like filtering is a commonly used feature extraction method for 2-D image processing fields. This method was first used in 2-D face detection and yielded good performance (Viola and Jones 2004); it was also applied to speech and non-speech detection (Nishimura and Kuroda 2008b). In order to utilize its low cost and high efficiency aspects, 1-D Haar-like filtering is newly employed for environmental sound recognition. The integral signal (*IS*) method (Nishimura and Kuroda 2008a) can further decrease the calculation cost considerably during the Haar-like filtering without compromising accuracy. Furthermore, the HMM classifier can achieve comparatively high recognition accuracy at the classification stage. With the above mentioned advantages, our Haar + HMM algorithm is very effective and can be used for environmental background sound recognition on the power-aware wearable sensor node.

The rest of this paper is organized as follows. Relevant previous work is discussed in "Sect. 2". In "Sect. 3", our proposed Haar + HMM algorithm is introduced in detail. Evaluation benchmarks for our proposed sound recognition algorithms are presented in "Sect. 4". "Section 5"

introduces a detailed experimental process. In "Sect. 6", with the introduced sound recognition algorithm and experimental data, system results and discussions are presented. Finally, the conclusion and future work are given in "Sect. 7".

2 Review of related work

In this section, we discuss three questions. First, the reasons why the sound is used as a detection medium to recognize people's daily activities are studied. Second, the researches related to sound recognition in general and for human activities are reviewed. Finally, researches about activity recognition utilizing wearable devices are reviewed.

Accurately knowing an individual and understanding the person-to-person's activities inside a society is a premise for the SCI system to fulfill its functions. Many detection media are used to recognize human activities, the most commonly used are acceleration (Bao and Intille 2004; Yin et al. 2008; Krause et al. 2005), video (Rota and Thonnat 2000), infrared ray (IR), and sound (Chen et al. 2005; Bharatula et al. 2005; Pentland 2005; Laibowitz et al. 2006; Yano et al. 2009). In research (Bao and Intille 2004), five two-axis accelerometers were attached on the tester's joints to recognize 20 human daily activities, and this was done successfully with 84 % accuracy. Work (Yin et al. 2008) also used the acceleration sensors to detect people's abnormal activities caused by Parkinson or Alzheimer's diseases. Based on their reports, we can conclude that the acceleration is mainly applied for detection of an individual's activities. It is rarely employed for detection of social activities. Video is also widely used to detect people's individual and social activities (Rota and Thonnat 2000). Because of security and privacy concerns, employing images as an activity detection medium is inconvenient or not allowed in some unobtrusive locations, such as in a hospital or a restroom. In addition, image signal processing is more computationally complex than acoustic signal processing. Sound has unique advantages in terms of detection accuracy, algorithm complexity, and operational convenience. Therefore, it is an ideal detecting medium to be utilized for the personal and social activity recognition.

Recently, in study (Chu et al. 2009), a new matching pursuit (MP) algorithm was introduced to decompose sound's time–frequency feature. In each step, the best decomposed matching atom from a redundant dictionary (such as Gabor dictionary) is searched. The sound can be presented by a linear combination with those atoms. A drawback of the MP algorithm is that the calculation cost for the searching enlarges significantly as the number of the atoms in the dictionary increases. In work (Dong et al. 2007), a complicated MFCC-based sound feature with HMM classification is implemented on the Ezairo 5900 SoC system. It is used to classify environmental sounds for a hearing aid application. A 24-bit specific DSP IP core is employed to process acoustic environmental sounds. It is difficult for our power-aware wearable sensor to execute these complex algorithms. In work (Chen et al. 2005), seven bathroom activities are recognized by detecting sounds, such as shower and brush tooth sounds. They are sampled by a microphone and are subsequently recognized by utilizing the MFCC + HMM algorithm on a PC. An average recognition accuracy of 83.5 % has been achieved. The difference between our research and Chen's work is that the recognition of Chen's work is processed off-line on a PC. In our case, processing must be done by using the limited power available in the wearable sensor node.

To accomplish the activity recognition on a power-aware wearable device, lightweight signal processing is necessary. In work (Krause et al. 2005), the following five activities can be discriminated using a wrist-worn eWatch accelerometer platform: walking, running, sitting, standing, and ascending/descending stairs. The detection accuracy evidently decreases with a reduction in the accelerometer's sampling rate. An optimized sampling scheme facilitates realization of a tradeoff—increase in the deployment lifetime of the eWatch without significant deterioration in accuracy. In work (Bharatula et al. 2005), how to trade off the power consumption and accuracy of a sound-based context recognition system is reported. Free combinations of nine time-domain features (such as mean and variance) and five frequency-domain features (such as bandwidth and frequency centroid) constitute sound feature sets. Different recognition results are obtained using different classifiers. A target sound feature set and classifier is decided by the tradeoff between accuracy and power consumption. However, exploring the ideal sound feature set and classifier is an empirical and complicated process. Hence, compared with this method, our proposed Haar-like sound feature with HMM classification is more effective.

3 Sound recognition implementation by utilizing the Haar + HMM algorithm

The proposed sound recognition flow is shown in Fig. 2. It follows two sequential steps: generation of off-line sound templates and on-line sound classification. Features of the template sound can be extracted by low computational Haar-like filtering. After training them off-line, the sound template is completed and stored in memory in advance. When the input test sound comes, its feature can be extracted on-line by applying the same filtering method. Following this, the recognition result is finally achieved by comparison with the prepared templates using the HMM classifiers (Rabiner 1989; Rabiner and Juang 1993).

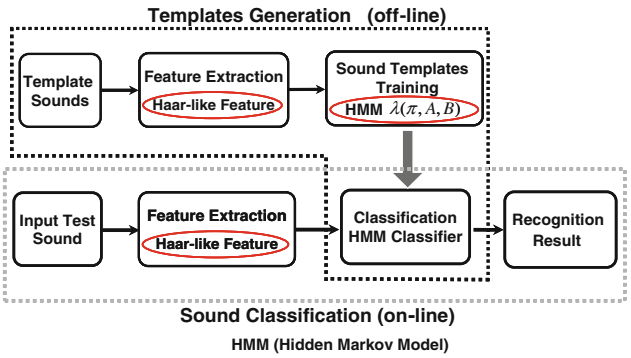

Fig. 2 Sound recognition flow

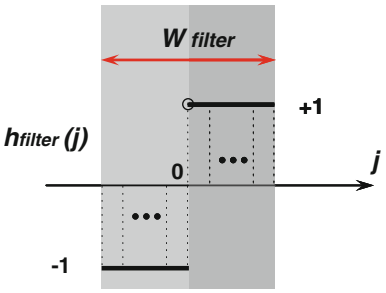

Fig. 3 One-dimensional (1-D) Haar-like filter $h_{filter}(j)$

3.1 Haar-like sound feature extraction

3.1.1 1-D Haar-like filtering

Inspired by the low cost and efficient feature extraction of Haar-like filtering used in 2-D face detection (Viola and Jones 2004), this novel filtering method is also applied to a 1-D signal, for example speech/non-speech detection, acceleration processing and recognition (Nishimura and Kuroda 2008b; Hanai et al. 2009).

A basic Haar-like filter $h_{filter}(j)$ is denoted by Eq. (1) and shown in Fig. 3.

$$h_{filter}(j) = \begin{cases} -1 & -W_{filter}/2 < j \le 0 \\ +1 & 0 < j \le W_{filter}/2 \end{cases}, \tag{1}$$

where, W_{filter} is the width of the Haar-like filter $h_{filter}(j)$.

In comparison with the MFCC's Mel-scale filter, Haar-like filter is simple and has a low calculation cost. Its filter width W_{filter} and shift width W_{shift} between neighbor filters, as shown in Fig. 4, are adjustable. These simple controllable parameters can be designed and applied for the feature extraction of environment sound in our research.

One frame length's sound signal (256 sampling points) processed by Haar-like filtering is shown in Fig. 4. The Haar-like feature x_m is calculated by the sum of the absolute outputs of Haar-like filtered signals:

$$x_m = \sum_{n=0}^{N-1} \left| \sum_{k=1}^{W_{filter}} h_m(k) \times s(nW_{shift} + k) \right|, \tag{2a}$$

$$= \sum_{n=0}^{N-1} |oneFilterValue(n)|, \tag{2b}$$

where $s(t)$ is the input sound signal and $h_m(k)$ denotes a Haar-like filter whose length can have a different value. W_{shift} is the shift width between neighbor filters. The filters number N in one frame is calculated as

$$N = (W_{frame} - W_{filter})/W_{shift} + 1. \tag{3}$$

Parameter W_{shift} is adjustable as α change [α is defined in Eq. (4)]. A longer W_{shift} (larger α) helps to reduce the N value and decrease the calculation of each frame's sound data accordingly. The variation of α also affects the final recognition result. When $\alpha = 0$, W_{shift} is set to 1.

$$\alpha = W_{shift}/W_{filter} \tag{4}$$

3.1.2 Integral signal (IS)

From Eq. (1) and Fig. 3, it follows that the coefficients of the Haar-like filter are -1 when $j \le 0$, and then change to $+1$ when $j > 0$. Thus, after the sound signal $s(t)$ passes a W_{filter} width Haar-like filter, the final filtering result is the absolute value of the difference between the sum of the sampling sound's $(-W_{filter}/2, 0]$ and $(0, W_{filter}/2]$ two-parts data. Based on this and borrowing from the integral image concept introduced in work (Viola and Jones 2004), a novel concept

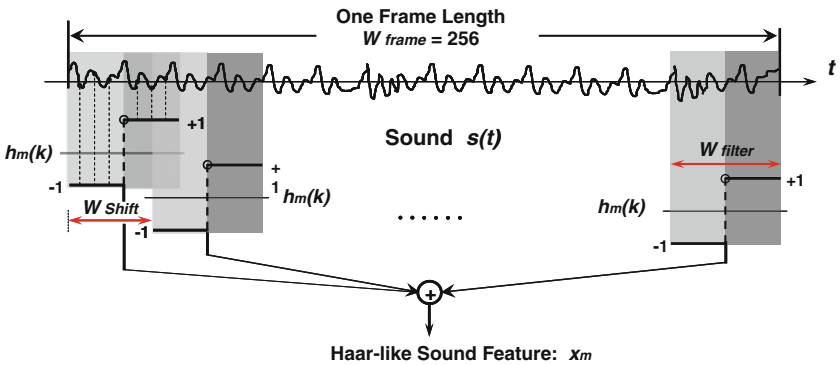

Fig. 4 One-dimensional (1-D) Haar-like filtering for one frame's sound signal

called Integral Signal (Nishimura and Kuroda 2008a) is newly utilized in this work. The IS of each sound frame has been calculated and stored in memory as a preprocessed intermediate signal for later use. It is defined as follows:

$$IS(n) = \sum_{t \le n} s(t) \tag{5}$$

Therefore, the filtered sound signal calculation can be denoted as

$$oneFilterValue = IS(t + W_{filter}) - 2 \times IS(t + W_{filter}/2) + IS(t) \tag{6}$$

In Eq. (2a), W_{filter} multiplication and $W_{filter} - 1$ addition calculations are need in order to obtain the filtering result of each frame sound. However, with the proposed *IS* method in Eq. (6), the calculations are reduced to one multiplication and two addition calculations. Therefore, it is obvious that the computational complexity of x_m in Eq. (2b) decreases. At the same time, the accuracy does not deteriorate.

3.1.3 Haar-like sound feature

A Haar-like filters group $h_v = \{h_{v1}, h_{v2}, \ldots, h_{vi}, \ldots, h_{vn}\}$ $(1 \le i \le n)$ chosen from M filters groups' pool is utilized to extract the feature of sound $s_v(t)$. $1 \le v \le p$, p is the number of all detected sounds. h_{vi} is an 1-D Haar-like filter which is as previous "Sect. 3.1.1" defined. Value n is the feature dimension of each sound frame.

Two parameters that decide the pool size M are defined as *H*aarWidMax (Maximum Haar filter Width) and *H*aarFilNum (Haar Filters Number). M's value is decided by combination expression below:

$$M = \begin{pmatrix} HaarFilNum \\ HaarWidMax/2 \end{pmatrix}. \tag{7}$$

For each frame of sound $s_v(t)$, its Haar-like feature X_v is formed by passing the Haar-like filters group $h_v = \{h_{v1}, h_{v2}, \ldots, h_{vi}, \ldots, h_{vn}\}$. Therefore, the sound feature X_v can be calculated by utilizing the *IS* method and is denoted as

$$X_v = \{x_{v1}, x_{v2}, \ldots, x_{vi}, \ldots, x_{vn}\}, \tag{8}$$

where $1 \le i \le n$, n = *H*aarFilNum is the feature dimension of each sound frame, and x_{vi} is as the previously introduced Haar-like feature x_m.

Sound feature plays an important role in achieving the expected final recognition results. With the simple Haar-like filters group and applying the *IS* method for the calculation, the extraction process to form the Haar-like sound feature can be completed with an extremely low computational cost. The achieved Haar-like sound features are simple and effective. These are very helpful in efficiently speeding up the feature extraction process and reducing the calculation cost significantly to meet the energy requirement.

3.2 Off-line training for the Haar-like filters group

Haar-like filters group h_v decides the feature X_v of the individual sound $s_v(t)$. The detailed training process to select the filters group h_v is described in work (Nishimura and Kuroda 2008b). The group's selection result is based on the training error. It is evaluated by matching feature vectors extracted from training data against the clustering model. Minimum error yielding of the filters group is selected.

Two assumptions are established in the training stage:

1. Once the value of the *H*aarFilNum has been decided, the dimension of all p sounds' feature is the same. That is similar to how X_v in Eq. (8) defines (n = *H*aarFilNum).
2. Once an h_v for the test sound $s_v(t)$ has been chosen, the left $p-1$ sound's filters group should be chosen from the remaining $M-1$ candidate filters groups' pool. This can guarantee that the different sound $s_v(t)$ adopts the different filters group h_v.

The two introduced parameters *H*aarWidMax and *H*aarFilNum in Eq. (7) decide the training complexity and searching scale during the h_v's selection stage. The size of the searching pool M is shown in Table 1 with combinations of these two parameters' variation. For example, when *H*aarWidMax = 18 and *H*aarFilNum = 5, feature $\{x_1, x_2, x_3, x_4, x_5\}$ of each sound is according to one Haar-like filters group among $M = 126$ filter groups pool.

During h_v's training, the LBG clustering model (Linde et al. 1980) is employed to develop new cluster centers in work (Nishimura and Kuroda 2008b). In this research, k-means cluster (Duda et al. 2001; Wiki_k-means 2012) is applied instead. This is because the k-means cluster is more controllable than the LBG cluster. It means that the number of clustering centers in LBG is split with a power of 2,

Table 1 Training Haar-like filters pool size M with relation to the "*H*aarWidMax" and "*H*aarFilNum"

*H*aarWidMax \ *H*aarFilNum.	2	3	4	5	6
20 [20 18 16 14 12 10 8 6 4 2]	45	120	210	252	210
18 [18 16 14 12 10 8 6 4 2]	36	84	126	126	84
16 [16 14 12 10 8 6 4 2]	28	56	70	56	28
14 [14 12 10 8 6 4 2]	21	35	35	21	7
12 [12 10 8 6 4 2]	15	20	16	6	1
10 [10 8 6 4 2]	10	10	5	1	
8 [8 6 4 2]	6	4	1		
6 [6 4 2]	3	1			
4 [4 2]	1				
2 [2]					

In this table, the *bottom gray cells* are impossible cases. The *middle white cells* are non-executable cases because the M value is less than our target 22 testing sounds. The *top gray cells* are our experimental cases

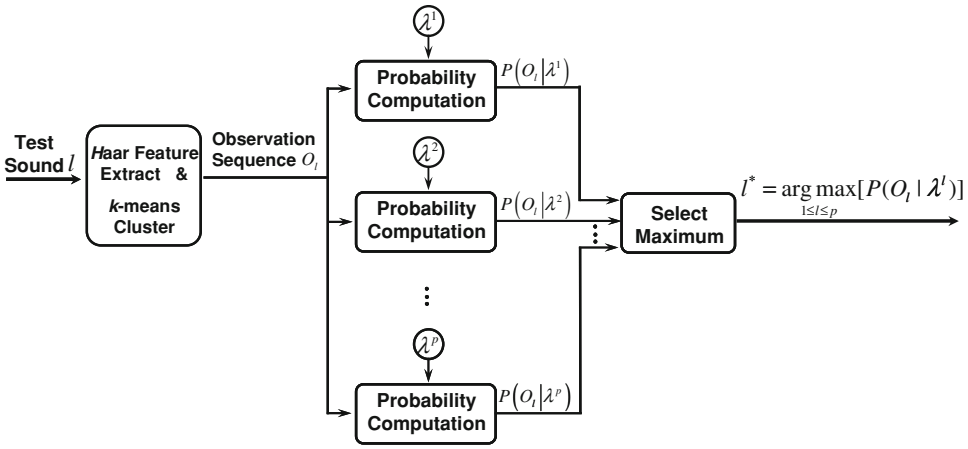

Fig. 5 Block diagram of a test sound's HMM classification

whereas it can adopt a value less than that of the LBG in *k*-means clustering. Moreover, in the following HMM classification stage, the number of observation states in the HMM model is equal to that of the *k*-means clusters. This clustering method change is of benefit to reduce the size of HMM's observation sequence, and further decreases the HMM classifier's calculation cost.

3.3 HMM classification

As shown in Fig. 2 to classify different environmental sounds, the appropriate off-line trained HMM classifier $\lambda^{v}(\pi, A, B)$ for individual sound $s_v(t)$ is necessary. After obtaining the updated centroids of sound $s_v(t)$ by *k*-means clustering, an observation O_q is formed by mapping the training sound vector q into a centroid index. Namely, the training vector is assigned to the index of the nearest centroid. Therefore, an observation sequence of sound $s_v(t)$ can be denoted as $O_v = \{O_1, O_2, \ldots, O_q, \ldots, O_T\}_v$. With the composed training sound's O_v and initial HMM parameter $\lambda^{v}(\pi, A, B)_0$, the Baum–Welch algorithm is applied to refine the model $\lambda^{v}(\pi, A, B)$ until it converges less than ε in the HMM classifier's training stage (Welch 2003; Rabiner and Juang 1993; Rabiner 1989).

The block diagram of an on-line test sound HMM classifier is shown in Fig. 5. In a real recognition stage, the extracted Haar-like feature of the unknown test sound l is quantized and establishes an observation sequence O_l. After computing the probability of all template sounds' $P(O_l|\lambda^l)$ $(1 \le l \le p)$ that employs the Viterbi algorithm (Rabiner and Juang 1993; Gold and Morgan 2000), the result with the highest likelihood among all the templates is recognized as the most similar to the test sound.

$$l^* = \arg\max_{1 \le l \le p}[P(O_l|\lambda^l)] \tag{9}$$

After analyzing Eq. (9), we can find that the calculation cost is on the order of $p \times N^2 \times T$ for each sound. The cost is proportional to the number of all detected sounds p, the square of the number of state N, and the number of observations in sequence T in the HMM model (Rabiner 1989; Rabiner and Juang 1993).

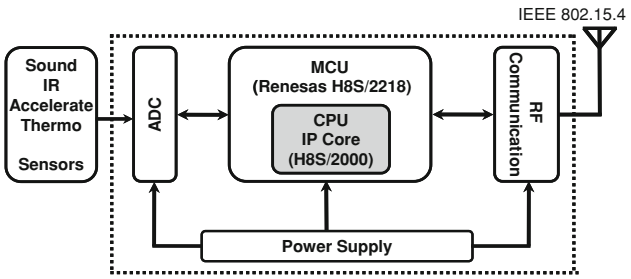

Fig. 6 Schematic diagram of the wearable sensor node

4 Benchmark values to evaluate our proposed sound recognition algorithms

As mentioned in "Sect. 1", a wearable device can help in accurately understanding individual and social interactions. These devices mostly have limited battery power. The following three blocks inside the wearable device mainly consume the limited energy (Yamashita et al. 2006; Bonfiglio and Rossi 2011; Bharatula et al. 2005): analog to digital converter (ADC), communication block, and MCU microprocessor, as depicted in Fig. 6.

Among them, the ADC and the communication blocks consume most energy; the energy remaining for the MCU is limited (Doherty et al. 2001; Yamashita et al. 2006; Bonfiglio and Rossi 2011). It has also been proven that locally processing the sampling data consumes less energy than transmitting them to the upper servers to process (Lynch and Loh 2006; Bharatula et al. 2004; Bonfiglio and Rossi 2011). Thus, the MCU should utilize the limited energy to complete signal processing inside the sensor node locally. That means the applied algorithm should be operated within the node energy

budget. At the same time, the final recognition accuracy should be guaranteed to a reasonable degree.

In aspect of how much the accuracy a sound recognition system could achieve, it has been reported in some researches (Chu et al. 2009; Ma et al. 2006; Eronen et al. 2006). If the recognition targets are environmental sounds, the listening test experiments performed in above researches indicate that people's hearing can achieve approximate 82 % accuracy. This conclusion provides a benchmark for deciding the accuracy level of our environmental sound recognition research.

Another aspect is to evaluate whether the applied recognition algorithm(s) can achieve the sound recognition by using the limited power assigned for the MCU inside our wearable sensor. The MCU is a Renesas Technology's H8S/2218 chip (Renesas_H8S_2218 2011; Yamashita et al. 2006; Nishimura et al. 2008). It is a microprocessor with a 0.35 μm process, 16-bit architecture, 65 basic instructions, 6 mA working current, and 3.0–3.6 V working voltage. Inside this MCU chip, there is an embedded low power H8S/2000 CPU core in which our proposed sound recognition algorithm is executed. The CPU core works at 20 MHz (50 ns per cycle), 1.8 V input voltage with an average 4 mA working current. The main parameters of the MCU and the CPU core are summarized in Table 2. From the specification (Renesas_H8S_2218 2011), we can calculate that for one-cycle commands, such as addition and subtraction operations, it consumes $4\ \text{mA} \times 1.8\ \text{V} \times 50\ \text{ns} = 0.36\ \text{nJ}$ energy. For four-cycle commands, such as multiplication operation, it consumes $4\ \text{mA} \times 1.8\ \text{V} \times 4 \times 50\ \text{ns} = 1.44\ \text{nJ}$ energy.

We aim that the sound module in the sensor node could continuously work for 24 h ($3{,}600 \times 24 = 86{,}400$ s), and the CPU core can finish the sound recognition algorithm within each 1 s sampling. Therefore, the recognition results can help capture a person's activities for a whole day. The algorithm is executed by individual addition and multiplication operations in the CPU.

- $1.8\ \text{V} \times 10\ \text{mAh} = 18\ \text{mWh} = 64.7\ \text{J}$ ($1\ \text{J} = 2.78 \times 10^{-4}\ \text{Wh}$) energy in CPU for calculation.
- $64.7\ \text{J}/86{,}400\ \text{s} = 7.5 \times 10^{5}\ \text{nJ/s} = 0.75$ mil. nJ/s energy assigned for execution of the sound recognition algorithm

Therefore, based on our hardware platform, a minimum 82 % sound recognition accuracy and maximum 0.75 million nJ/s power consumption for computation are decided. These two values are used as benchmarks to evaluate the performance of the sound recognition algorithms. They are indicated as dashed-lines in Fig. 11 for performance comparison. If the performance marks of the applied algorithms drop into the top left region of the figure, it can be concluded that the algorithms are suitable for our sound recognition application.

5 Experimental process

5.1 Test target sounds

Many personal and social activities happen in our daily life. We can understand these activities by recognizing their background sound. 22 experimental sounds in our research are listed below. They are sampled in a real environment, and not in a noise-isolated space. Among them, the background sounds of social activities are sampled in noisy environments.

- Background sounds of personal activities

1. Vacuum cleaner (house cleaning)
2. Washing machine (wash clothing)
3. Water sound from tap —Household clean
4. Brush teeth
5. Shaving (shave beard)
6. Taking shower (bath)
7. Hair dryer (dry hair)
8. Urination (man)
9. Flush toilet (use water closet) —Sanitary
10. Chewing cake (eat)
11. Drinking (drink something)
12. Oven-timer (toast some food) —Dietetic
13. Walk inside room
14. Walk (walk in street)
15. Run
16. Moving train (travelling in a train)
17. Rain hits an umbrella (in the rain) —Outdoor acts
18. Telephone ringing (phone call)

Table 2 Main technical parameters of the H8S/2218 MCU and embedded H8S/2000 CPU Core

	Voltage (V)	Current (mA)	Energy (mAh)
MCU (H8S/2218)	3.0~3.6V	6mA	150mAh
CPU core (H8S/2000)	1.8V	4mA	10mAh*

* 10 mAh is the energy assigned for the sound processing module in CPU

- Background sounds of social activities

19. Supermarket (shopping)
20. Discussion/talking in lab (discuss/talk with others)
21. Restaurant (outside dining)
22. Front square of a subway entrance (meet friends)

5.2 Experimental data collection and data sets

The sampling mode of our wearable sensor node introduced in Fig. 1 has been wirelessly configured in advance. During data collection, it operates at the setting configuration. The node is hung in front of the tester's chest or set within the environment depending on the test activity. For example, it can be placed on the bathroom's countertop when the tester takes a shower. Under normal circumstances, it is hung in front of the chest. The sampling rate of these 22 sounds is 16 kHz with 16-bit resolution. The data are stored in the sensor node's on-board memory, and used for the sounds' templates training and test inputs.

Each of the above mentioned 22 type of sounds is recorded more than three times on different days. Among many recordings of each sound, one recording is randomly picked as the testing input, and these different 22 testing input sounds compose the testing data set. At the same time, the remaining records of each sound are collected together as the templates training set. The durations of the testing set vary from 14.9 to 256.8 s (indicated on the 2nd column of Table 4). For the templates training set, their durations range from 16 to 277 s and total length is 1,788 s.

5.3 Performance evaluation

During the recognition process, each unit length of the detected sound is 1 s. It means that the algorithms for our sound recognition should finish within each one second as discussed in "Sect. 4". Each sound frame contains 256 sampling points with a 50 % overlap.

The recognition accuracy rate AR is defined as:

$$AR = \frac{C_u}{A_u} \times 100\ \% \tag{10}$$

where C_u stands for the number of correctly recognized units (1 s period), A_u stands for the number of all input sound units (1 s period).

Another evaluation factor of the performance of our sound-context recognition system is the calculation cost. It can be determined by the amount of multiplication and addition calculations within the whole algorithm flow.

6 Experimental results and discussion

As analyzed in "Sect. 4", the sound recognition algorithm executed on the wearable sensor requires that the recognition accuracy should be improved while satisfying the sensor node's computational power budget. After conducting experiments and analyzing their results in this section, we can find that our proposed Haar + HMM algorithm for environmental sound recognition can successfully satisfy these requirements.

6.1 Parameters tuning and recognition accuracy

Figure 7 indicates how the parameters *H*aarFilNum and *H*aarWidMax affect the average accuracy of the sound recognition system. Among all these cases, when *H*aarFilNum = 5, *H*aarWidMax = 18, $\alpha = 0$ ($W_{shift} = 1$), number of HMM states = 7, number of HMM observe symbol = 15, and $\varepsilon = 0.01$, the average accuracy of the 22 sounds can reach highest at 98.2 %. Even with *H*aarFilNum = 2 (other parameters are identical), it can yield accuracy of more than 94.0 %. These results greatly outperform the required minimum accuracy of 82 % decided in "Sect. 4", and also prove that our proposed Haar + HMM environmental sound recognition algorithm with the proposed training method is effective.

Besides $\alpha = 0$, the recognition results of typical $\alpha = W_{shift}/W_{filter} = 0.5$ and $\alpha = W_{shift}/W_{filter} = 1$ are also illustrated in Fig. 8. Except for the value of α, the parameters are set as in the previous experiment with a maximum accuracy 98.2 %. From this figure, we can observe that the accuracy of all cases surpasses the required minimum accuracy of 82 %. The variation of α does not significantly affect the accuracy of our proposed sound recognition system. The accuracy range is from a minimum 93.7 % to a maximum 98.2 %. Different combinational values of the *H*aarFilNum and α introduce only 4.5 % variation. For the maximum accuracy which happens at *H*aarFilNum = 5, the variation of accuracy is only 1.3 %. So the influence of the value of α on accuracy is not significant if the appropriate *H*aarFilNum is chosen.

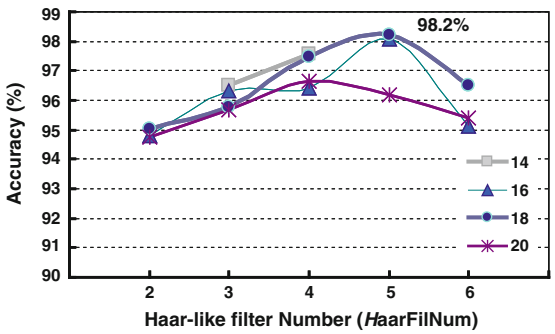

Fig. 7 Average accuracy in function of the parameters: *H*aarFilNum and *H*aarWidMax ($\alpha = 0$)

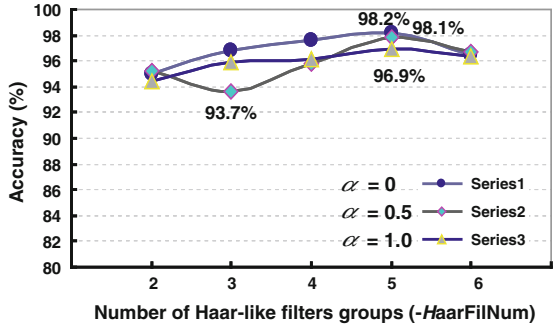

Fig. 8 Average accuracy in function of the parameter: *H*aarFilNum and α

6.2 Different sound features' performance comparison

Different sound features yield different performances. With the same HMM classifier utilized in "Sect. 6.1", the accuracy and calculation cost of the MFCC (Davis and Mermelstein 1980) and three Haar-like features ($\alpha = 0$, 0.5, 1.0, *H*aarFilNum = 5, *H*aarWidMax = 18) are compared. The process of the MFCC feature extraction is complex which contents FFT, logarithm, discrete cosine transform (DCT) and many multiplication computations. On the other hand, the Haar-like feature only requires a small number of addition and multiplication as Eq. (6) denotes. The experimental results shown in Fig. 9 and Table 3 prove that our proposed Haar + HMM outperforms MFCC + HMM in terms of both accuracy and calculation cost. The most aggressive case with $\alpha = 1.0$ can obtain 96.9 % accuracy by employing only 8.3 % of MFCC's multiplication and 8.2 % of MFCC's addition calculations.

Parameter α is an important and effective variable that affects system's accuracy and calculation cost. From Figs. 8, 9 and Table 3, it is evident that the average recognition accuracy drops by 1.3 % when the value of α changes from 0 to 1. However, this trivial 1.3 % decrease in accuracy helps to considerably reduce the calculation cost. The multiplication calculation can be reduced by 72.2 % and the addition calculation by 79.8 % compared with the referenced $\alpha = 0$ case. It is because the filters number N in Eq. (3) deceases with increasing α and further reduces the calculation cost in sound's feature extraction stage dramatically. Meanwhile, the increase of α slightly deteriorates the final recognition accuracy. We believe this limited accuracy decrease is because most of the environmental sounds are quasi-stationary.

6.3 Performance comparison of different classifiers

With the same $\alpha = 1.0$ Haar-like feature configuration used in "Sect. 6.2", the performance of the HMM classifier is investigated with the referenced *k*-means and LBG classifiers. The comparison results are shown in Fig. 10 and Tables 4, 5. The clusters number of the HMM and the

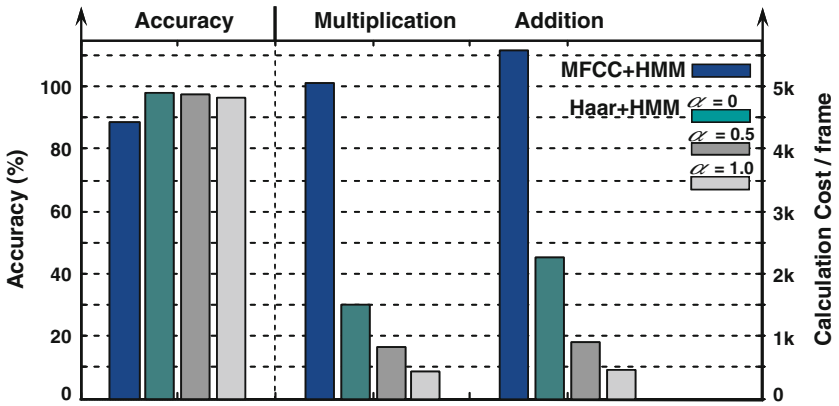

Fig. 9 Performance comparison of proposed Haar-like and traditional MFCC sound features with same HMM classifier—average accuracy and multiplication/addition calculation cost (256 samples/frame)

Table 3 Different sound feature—MFCC and Haar-like feature ($\alpha = 0$, 0.5, 1.0). Performance comparison (per frame = 256 samples)

Feature + Classifier	Average Accuracy	Multiplication (per frame)	Addition (per frame)
MFCC + HMM (Mel-filters=22)	88.7%	5,050	5,580
Haar + HMM ($\alpha = 0$)	98.2%	1,510 (MFCC's 29.9%)	2,255 (MFCC's 40.4%)
Haar + HMM ($\alpha = 0.5$)	98.1%	0,834 (MFCC's 16.5%)	0,903 (MFCC's 16.2%)
Haar + HMM ($\alpha = 1.0$)	96.9%	0,420 (MFCC's 8.3%) (α =0's 27.8%)	0,456 (MFCC' 8.2%) (α =0's 20.2%)

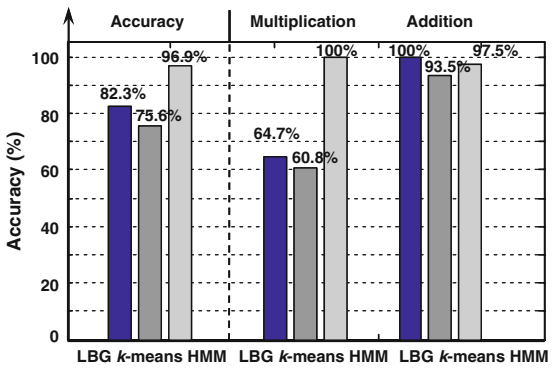

Fig. 10 Performance comparison of LBG, *k*-means, and HMM classifiers with same Haar-like sound feature (Haar-like feature's $\alpha = 1.0$)

k-means classifiers are 15. The LBG's cluster is set to $16 = 2^4$ which is close to the *k*-means and HMM's 15 clusters for comparison. It can be seen that the Haar + HMM algorithm achieves the highest average accuracy of 96.9 % among these three cases.

During the classification, the HMM classifier needs more computation than the *k*-means classifier does. As in "Sect. 3.3" mentioned, the Viterbi algorithm determines the final recognition performance from the on-line observation sequence O_l in the HMM classification. The algorithm is additionally employed to estimate the likelihood of O_l sequence that is calculated from the *k*-means cluster's centroids developed during the off-line training stage. Moreover, the Viterbi algorithm itself employs many multiplications as Eq. (9) indicated. These obviously lead to an increase of multiplication calculation compared with *k*-means cluster in Fig. 10.

6.4 Performance comparison of whole system

Performance comparison of the recognition algorithms of different environmental sounds is illustrated in Table 5 and

Table 4 Recognition accuracy confusion matrix of 22 different tested sounds with Haar + HMM algorithm ($\alpha = 1.0$); accuracy comparison with other Haar + HMM two cases ($\alpha = 0/0.5$), Haar + *k*-means and Haar + LBG

	Sound Length	Twenty-two Test Sounds' Recognition Confuse-Matrix																							Haar+ (α=1.0) HMM	Haar+ (α=0.5) HMM	Haar+ (α=0) HMM	Haar+ (α=1.0) k-means	Haar+ (α=1.0) LBG
	Len. (S)	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	Cu*	Au*	Acc.	Acc.	Acc.	Acc.	Acc.
A1	48.3	47	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	48	**97.9%**	100%	100%	100%	79.2%
A2	170.5	0	169	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	169	170	**99.4%**	100%	100%	70.0%	65.3%
A3	141.4	0	0	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	141	141	**100%**	93.6%	99.3%	99.3%	99.3%
A4	26.1	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	26	**100%**	100%	96.2%	76.9%	65.4%
A5	71.3	1	0	0	1	55	5	0	0	0	0	0	5	0	0	0	0	4	0	0	0	0	0	55	71	**77.5%**	97.2%	95.8%	12.7%	56.3%
A6	105.7	3	0	0	0	0	97	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	97	105	**92.4%**	98.1%	98.1%	30.5%	96.2%
A7	25.0	0	0	0	0	0	0	20	0	1	0	0	0	0	1	0	3	0	0	0	0	0	0	20	25	**80.0%**	100%	92.0%	100%	100%
A8	14.9	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	14	**100%**	100%	100%	35.7%	50.0%
A9	17.5	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	17	17	**100%**	100%	100%	100%	100%
A10	66.9	0	0	0	0	0	0	0	0	0	66	0	0	0	0	0	0	0	0	0	0	0	0	66	66	**100%**	98.5%	95.5%	100%	92.4%
A11	20.4	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	20	20	**100%**	95.0%	100%	75.0%	95.0%
A12	51.5	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	1	50	51	**98.0%**	98.0%	100%	100%	100%
A13	35.6	0	0	0	0	0	0	0	1	0	0	0	0	34	0	0	0	0	0	0	0	0	0	34	35	**97.1%**	100%	97.1%	45.7%	91.4%
A14	21.9	0	0	0	0	0	0	0	1	0	0	0	0	0	20	0	0	0	0	0	0	0	0	20	21	**95.2%**	100%	100%	100%	100%
A15	36.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	36	36	**100%**	94.4%	100%	86.1%	86.1%
A16	40.2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	39	40	**97.5%**	100%	97.5%	97.5%	90.0%
A17	142.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0	0	2	140	142	**98.6%**	95.8%	97.2%	100%	99.3%
A18	21.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	21	21	**100%**	95.2%	95.2%	100%	100%
A19	97.3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	0	0	96	97	**99.0%**	97.9%	99.0%	52.6%	61.9%
A20	256.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	255	0	0	255	256	**99.6%**	100%	99.16%	94.1%	75.0%
A21	152.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	152	152	**100%**	99.3%	100%	24.3%	74.3%
A22	117.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	116	116	117	**99.2%**	95.7%	98.3%	61.5%	32.5%
		Average Accuracy																								**96.9%**	98.1%	98.2%	**75.6 %**	**82.3%**

Cu*: Correctly recognized ***u***nits **Au*: *All*** input sound ***u***nits

******* : for space limitation, above Confuse-Matrix is for Haar+HMM (α =1.0). α =0 / 0.5 two cases, Haar+k-mean, Haar+LBG just has accuracy.

A1: Vacuum cleaner (house cleaning)
A2: Washing machine (wash clothing)
A3: Water sound from tap (wash something)
A4: Brush teeth
A5: Shaving (shave beard)
A6: Take shower
A7: Hair dryer (Dry hair)
A8: Urination (man)
A9: Flush toilet (use water closet)
A10: Chewing cake (eat)
A11: Drinking (drink something)
A12: Oven-timer (toast some food)
A13: Walk inside room
A14: Walk (walk in street)
A15: Run
A16: Moving train (travelling in a train)
A17: Rain hits an umbrella (in the rain)
A18: Telephone ring (phone call)
A19: Shopping in supermarket
A20: Discuss in lab (discuss with others)
A21: Restaurant (dining)
A22: Station

Table 5 Comprehensive performance comparison of four different sound recognition algorithms: MFCC + HMM, Haar + LBG, Haar + *k*-means, and Haar + HMM (1 s unit = 124 frames in each 1 s sound unit, Haar-like feature's $\alpha = 1.0$)

Feature + Classifier	Average Accuracy	Feature (F) (mil.)		Classifier (C) (mil.)		Total (F+C) (mil.)		Energy * (mil. nJ)
		Mul.	Add	Mul.	Add.	Mul.	Add	
MFCC + HMM (Mel-filters=22, training centroids=15)	88.7%	0.625	0.692	0.938	0.888	1.563	1.580	2.920
Haar + LBG (*H*aarFilNum=5, LBG codebooks=16)	82.3%	0.052	0.056	0.198	0.358	0.250	0.414	0.509
Haar + *k*-means (*H*aarFilNum=5, training centroids=15)	75.6%	0.052	0.056	0.186	0.334	0.238	0.390	0.483
Haar + HMM (*H*aarFilNum=5, training centroids=15)	**96.9%**	**0.052**	**0.056**	**0.306**	**0.349**	**0.358**	**0.405**	**0.661**

* The whole energy = 1.44 nJ × Mul. + 0.36 nJ × Add (mil. nJ) based on Sect. 4's discussion

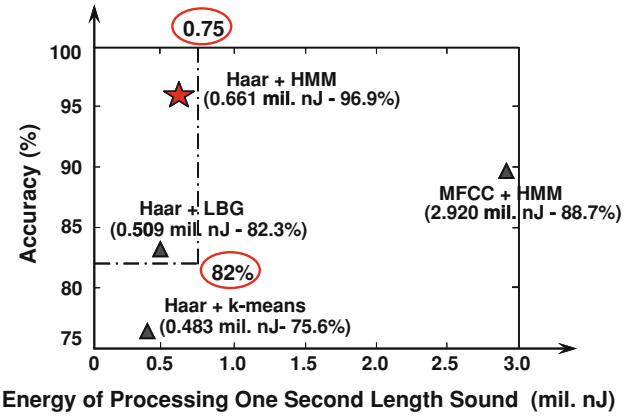

Fig. 11 Performance comparison of MFCC + HMM, Haar + LBG, Haar + *k*-means, and Haar + HMM (Haar-like feature's $\alpha = 1.0$)

Fig. 11. Results of four algorithms—MFCC + HMM, Haar + LBG, Haar + *k*-mean, and Haar + HMM are compared. Among them, the average accuracy of the three algorithms: MFCC + HMM, Haar + LBG, Haar + HMM outperforms the 82 % benchmark decided in "Sect. 4". The highest accuracy is achieved by the Haar + HMM algorithm.

In Fig. 11, we also find that the sound feature extracted by the Haar-like filtering needs less calculation energy than the MFCC filtering. Three algorithms with the Haar-like feature are executable based upon the wearable sensor's energy budget. However, the MFCC sound feature with the HMM classifier is so complicated that it goes beyond the 0.75 mil. nJ/s calculation energy benchmark. Compared with the Haar + *k*-mean method, the Haar + HMM algorithm's calculation energy increases $0.661-0.483 = 0.178$ mil. nJ/s. However, the accuracy obviously increases by a further $96.9-75.6\ \% = 21.3\ \%$ due to the effective HMM classification.

Within the top left region confined by the two benchmarks, the Haar + HMM algorithm achieves better comprehensive performance. It consumes a little more energy $0.661-0.509 = 0.152$ mil. nJ/s compared to the Haar + LBG spends. However, it can achieve a much higher 96.9 % accuracy than the Haar + LBG's 82.3 %. When the requirement of the calculation energy becomes stricter, Haar + LBG can be a candidate solution.

7 Conclusion and future work

Environmental background sound is a rich information source for identifying individual and social behaviors. In this study, a power-aware wearable sensor is utilized to recognize the environmental sounds happening in the background of these activities. A novel low calculation with high recognition accuracy Haar + HMM algorithm is utilized to realize this function.

Based on the wearable sensor power budget and listening test results, the target recognition accuracy and power consumption benchmarks to evaluate the applied sound algorithms have been decided. By utilizing the Haar + HMM algorithm, an average accuracy of 96.9 % of 22 typical personal and social related environmental sounds has been achieved. This proves that our proposed algorithm performs well in the application of the sound-context recognition. At the same time, it still satisfies the wearable sensor's power requirement. Experimental comparison also indicates that our method outperforms other commonly used sound recognition algorithms with respect to the accuracy and power consumption. This method is promising and applicable for future systems combined with other sensor(s), such as accelerometers, to achieve higher accuracy rate and more reliable human activity recognition results for the SCI system.

There are some interesting tasks to be conducted in the future. One is to implement our applicable algorithm and evaluate it upon the real power-aware wearable sensor node. Another, as the usual environmental sound

recognition researches, the test sound templates of our research have been trained and registered. However, the input can be a new non-registered sound in real applications. To solve this problem, some methods in the similar speech and face recognition researches can be considered. Moreover, the sound–context recognition of more complex social activities is also a direction of our future work.

Acknowledgments The authors want to sincerely thank Dr. Yano K., Senior Chief Researcher of Central Research Laboratory at Hitachi Ltd. for providing us an opportunity to take part in this research. We want to express our sincere acknowledges to Mr. Ohkubo N. and Mr. Wakisaka Y. for developing the wearable sensor node well used in our experiments. We would also thank Dr. Daribo Ismael, Mr. Jun Nishimura, and Mr. Hao Zhang for their helpful discussion and comments during this research. Finally, we gratefully acknowledge the anonymous reviewers. Their valuable comments and suggestions are very helpful to improve the presentation of this paper and our future work.

References

Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data. In: Pervasive 2004, LNCS 3001, pp 1–7

Bharatula NB, Ossevoort S, Stager M, Troster G (2004) Towards wearable autonomous microsystems. In: PERVASIVE 2004, LNCS 3001, pp 225–237

Bharatula NB, Stager M, Lukowics P, Troster G (2005) Empirical study of design choices in multi-sensor context recognition systems. In: The 2nd international forum on applied wearable computing (IFAWC'05), pp 79–93

Bonfiglio A, Rossi DR (2011) Wearable monitoring systems. Springer, NY

Chen J, Zhang JA, Kam H, Shue L (2005) Bathroom activity monitoring based on sound. In: PERVASIVE 2005, LNCS 3468, pp 47–61

Choudhury T (2004) Sensing and modeling human networks. PhD Dissertation, MIT. http://hd.media.mit.edu

Chu S, Narayanan S, Kuo C-CJ (2009) Environmental sound recognition with time-frequency audio features. IEEE Trans Audio Speech Lang Process 17:1142–1158

Cowling M, Sitte R (2003) Comparison of techniques for environmental sound recognition. Pattern recognit lett 24:2895–2907

Culler D, Estrin D, Srivastava M (2004) Overview of sensor networks. Computer 37:41–49

Davis SB, Mermelstein P (1980) Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. IEEE Trans Speech Audio Process 28:357–366

Doherty L, Warneke BA, Boser BE, Pister KSJ (2001) Energy and performance considerations for smart dust. Int J Parallel Distrib Syst Netw 4:121–133

Dong R, Hermann D, Cornu E, Chau E (2007) Low-power implementation of an HMM-based sound environment classification algorithm for hearing aid application. In: Proceedings of EUSIPCO 2007

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, NY

Eronen AJ, Peltonen VT et al (2006) Audio-based context recognition. IEEE Trans Audio Speech Lang Process 14:321–329

Gold B, Morgan N (2000) Speech and audio signal processing. Wiley, NY

Goldhor RS (1993) Recognition of environmental sounds. In: IEEE ICASSP. pp 149–152

Guo B, Zhang D, Imai M (2011a) Toward a cooperative programming framework for context-aware applications. J Pers Ubiquitous Comput 15:221–233

Guo B, Zhang D, Wang Z (2011b) Living with internet of things: the emergence of embedded intelligence. In: IEEE international conference on cyber, physical and social computing (CPSCom), pp 297–304

Hanai Y, Nishimura J, Kuroda T (2009) Haar-like filtering for human activity recognition using 3D accelerometer. In: IEEE 13th digital signal processing workshop and 5th IEEE signal processing education workshop, pp 675–678

Krause A et al (2005) Trading off prediction accuracy and power consumption for context-aware wearable computing. In: Proceeding of the 9th IEEE international symposium on wearable computers (ISWC'05), pp 20–26

Laibowitz M, Gips J, Aylward R, Pentland A, Paradiso J (2006) A sensor network for social dynamic. In: IEEE IPSN'06, pp 483–491

Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun 28:84–95

Lynch JP, Loh KJ (2006) A summary review of wireless sensors and sensor networks for structural health monitoring. Shock Vib Dig 38:91–128

Ma L, Milner B, Smith D (2006) Acoustic environment classification. ACM Trans Speech Lang Process 3:1–22

Nishimura J, Kuroda T (2008a) Haar-like filtering based speech detection using Integral Signal for sensornet. In: International conference on sensing technology, pp 52–56

Nishimura J, Kuroda T (2008b) Low cost speech detection using Haar-like filtering for sensornet. In: 9th international conference on signal processing, pp 2608–2611

Nishimura J, Sato N, Kuroda T (2008) Speech "siglet" detection for business microscope. In: IEEE international conference on pervasive computing and communications (PerCom08), pp 147–152

Peltonen V, Tuomi J, Klapuri A, Huopaniemi J, Sorsa T (2002) Computational auditory scene recognition. In: IEEE ICASSP, pp 1941–1944

Pentland A (2005) Socially aware computation and communication. IEEE Comput 38:33–40

Rabiner LR (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc IEEE 77:257–286

Rabiner LR, Juang BH (1993) Fundamentals of speech recognition. Prentice-Hall, Englewood Cliff

Renesas_H8S_2218 (2011) Renesas H8S_2218 MCU technical details. http://www.renesas.com/fmwk.jsp?cnt=h8s2218_h8s2212_root.jsp&fp=/products/mpumcu/h8s_family/h8s2200_series/h8s2218_h8s2212_group/. Accessed Dec 2011

Rota N, Thonnat M (2000) Activity recognition from video sequences using declarative models. In: Proceedings of the 14th European conference on artificial intelligence, pp 673–680

Veitch R, Aubert LM, Woods R, Fischaber S (2011) FPGA implementation of a pipelined Gaussian calculation for HMM-based large vocabulary speech recognition. Int J Reconfi Comput. doi:10.1155/2011/697080

Viola P, Jones M (2004) Rapid object detection using a boosted cascade of simple features. In: Computer society conference on computer vision and pattern recognition, pp 511–518

Welch LR (2003) Hidden Markov models and the Baum–Welch algorithm. IEEE Inf Theory Soc Newslett 53:9–13

Wiki_*k*-means (2012) Wikipedia introduction of the *k*-means algorithm. http://en.wikipedia.org/wiki/K-means_clustering. Accessed Jan 2012

Yamashita S, Shimura T, Aiki K, Ara K, Ogata Y, Shimokawa I, Tanaka T, Kuriyama H, Shimada K, Yano K (2006) A 15 × 15 mm, 1 μA, reliable sensor-net module: enabling application-specific nodes. In: The fifth international conference on information processing in sensor networks (IPSN 2006), pp 383–390

Yano K, Sato N, Wakisaka Y, Tsuji S, Ohkubo N, Hayakawa M, Moriwaki N (2008) Life thermoscope: integrated microelectronics for visualizing hidden life rhythm. In: IEEE ISSCC digests of technical papers, pp 136–137

Yano K, Ara K, Moriwaki N, Kuriyama H (2009) Measurement of human behavior: creating a society for discovering opportunities. Hitachi Rev 58:139–144

Yin J, Yang Q, Pan JJ (2008) Sensor-based abnormal human-activity detection. IEEE Trans Knowl Data Eng 20:1082–1090

Zhang D, Guo B, Yu Z (2011) The emergence of social and community intelligence. IEEE Comput 44:21–28