Low-Complex Environmental Sound Recognition Algorithms for Power-Aware Wireless Sensor Networks

September 2012

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Engineering

Keio University

Graduate School of Science and Technology
School of Integrated Design Engineering

Zhan, Yi

Abstract

The past decade witnessed rapid development in the basic Internet, communications theories and in some newly emerging technologies, such as wireless sensor networks (WSNs), wearable sensing and computation. With the rapid development of these technologies, understanding individual's activities, social interaction, and group dynamics of a certain society becomes possible and plays an important role for creation a ubiquitous information society around us. This will inevitably enrich our life's content and improve our society's efficiency.

Environmental background sound is a rich information source for identifying individual and social behaviors. Therefore, many power-aware wearable devices in the WSNs system with sound recognition function are widely used to trace and understand human activities. Design of these sound recognition algorithms has two major challenges: limited computation resources and a strict power consumption requirement. These motivate us to develop a new method for recognizing environmental background sounds upon our power-aware wearable sensor node. Therefore, we address to develop a new and low-complex sound recognition algorithm which can achieve high recognition accuracy while still meeting the wearable sensor's power requirement in the dissertation.

In Chapter 1, the motivation and challenge of this study are introduced. Related work is also surveyed.

In Chapter 2, hardware architecture of the power-aware wearable senor node for detection and software-level sound recognition flow are introduced. Upon this resource limited platform, the assumptions and special constrains of this research are discussed. Basic approaches to tradeoff the system's accuracy and power consumption problem are proposed.

In Chapter 3, the experimental setup and process are presented. Comprehensively considering the accuracy and power consumption as the proposed sound recognition algorithms' performance evaluation criteria is also discussed.

In Chapter 4, sound feature extraction Mel-frequency cepstral coefficients (MFCC) and vector quantization (VQ) classification Linde-Buzo-Gray (LBG) algorithm is applied for

recognizing the environmental background sounds. Applying this algorithm to 20 typical daily activity sounds, average recognition accuracy of 93.8% can be achieved. In this algorithm, how the three parameters (i.e., Mel filters number, frame-to-frame overlap and LBG codebook cluster number) affect the system's calculation burden and accuracy is also investigated. Based on the performance evaluation method in Chapter 3, the comprehensive performance of proposed MFCC+LBG algorithm is evaluated.

In Chapter 5, a new low-complex sound feature extraction Haar-like filtering with hidden Markov model (HMM) classification algorithm is proposed and applied to recognize the environmental sounds. Average recognition accuracy 96.3% of 20 typical daily activity sounds by the proposed algorithm can be achieved, which outperforms normal personal hearing capacity 82% accuracy. At the same time, it also satisfies the amount of calculation cost decided by the wearable sensor node's energy resource. Through experimental comparison, the proposed method outperforms other normally utilized sound recognition algorithms as the recognition accuracy and calculation cost two evaluation parameters concerned.

In Chapter 6, summary of this study is concluded. Overview of the future work is also mentioned.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my advisor Professor Tadahiro Kuroda for his kind support, encouragement and patience during my Ph.D study. His passion, dedication, and dexterity towards the research and work are a good example, which activates me to keep on going forward in and out of school.

I would also like to deliver my appreciation to Prof. Yoshimistu Aoki, Prof. Nobuhiko Nakano, and Prof. Hideo Saito for their time, valuable comments, helpful discussion on my thesis and effort to serving on my thesis committee.

I wish to thank the Japanese Government (Monbukagakusho: MEXT) for providing me a precious opportunity and generous support to pursue my graduate study at Keio University. I would also like to show my sincere acknowledgement to Prof. Tadahiro Kuroda and Prof. Zhihua Wang of Tsinghua University for their strong recommendations for this opportunity.

I would like to appreciate my team mates - Shun Miura and Jun Nishimura for giving me a lot of help during the study. My gratitude should also deliver to Dr. Kazuo Yano and Dr. Nobuo Sato of Central Research Laboratory, Hitachi, Ltd. for their helpful discussion, encouragement, and generous support at the initial stage of this research.

I am also deeply thankful to all the lab members. Experience of studying these years with them in Kuroda Laboratory is a wealth for me, and this precious memory is to bear in my mind forever. Especial thanks should be given to Noriyuki Miura, Yasumoto Tomita, Takayuki Shibasaki, Shun Miura, Mari Inoue, Yuxiang Yuan, Vishal Kulkarni, Yanfei Chen, Hitoshi Kikuchi, Yoichi Yoshida, Xiaolei Zhu, Andrzej Radecki, Tsutomu Takeya, Mitsuko Saito, Yasuhiro Take, Takayuki Abe, Wataru Mizuhara for their kind help and pleasant companion with me in and outside school. I would like to thank Tsunaaki Shidei, Chika Wada, Ritsuko Mukai, and Chika Kijima for various experimental and academic issues. Staying with the special research professors Won-Joo Yun, Lan Nan, Hayun Chung is also a pleasant and thankful period of time.

Friends gave me tremendous help, constructive suggestions and encouragement during the doctoral study. Studying and living with them widen my knowledge scope and enrich my life,

I would like to thank them and cherish this precious friendship. I also owe a special appreciation to Ms. Lijiang Niu for her kind encouragement.

I would like to thank Prof. Keqian Zhang, Prof. Lian Gong, and Prof. Zhibin Pan for teaching me basic knowledge, giving me cordial help and encouragement during my growth and study. I am also blessed with a family full of love, dedication, and trust. I would like to express the gratitude to my beloved parents and sister. This work could not have been completed without their everlasting and dedicated cultivation, trust and love.

Keio University, Yokohama, Japan
August 8th, 2012
Yi Zhan

Contents

Abstract ..... **I**
Acknowledgement ..... **III**
Contents ..... **V**
List of Figures ..... **VIII**
List of Tables ..... **X**

Chapter 1 Introduction **1**
1.1 Research Motivation 2
1.2 Wireless Sensor Networks and Front-End Wearable Sensors 3
1.2.1 Introduction of Wireless Sensor Networks 3
1.2.2 An Application Example of the WSNs System 4
1.2.3 Front-End Wearable Sensor Node in the WSNs System 6
1.2.4 Unique Constrains and Challenges 9
1.3 Environment Background Sound Detection for Activity Recognition 10
1.3.1 Low-Level Activity Recognition 10
1.3.2 Why Applies Sound as the Detection Media? 12
1.3.3 Application Domains 14
1.4 Related Work 15
1.4.1 Environmental Background Sound Recognition 15
1.4.2 Audio-Context Recognition on Hardware Platforms 16
1.4.3 Tradeoffs of the Sound-Context Recognition on Wearable Platform 18
1.5 Research Objects and Contributions 20
1.6 Thesis Organization 22

Chapter 2 Our System Study **24**
2.1 Our Hardware and Software System 25
2.1.1 Hardware Platform and Specifications of Our Wearable Sensor 25
2.1.1.1 Hardware Schematic Diagram 25
2.1.1.2 Why MCU? DSP, FPGA, and MCU Comparison 27
2.1.1.3 Hardware Specification 28
2.1.2 Recognition Flow in Software Aspect 29
2.2 Assumptions and Constraints of This Research 30
2.2.1 Placement of the Wearable Sensor 30

2.2.2 Dominant and Single-Content Sounds 31
2.2.3 Local Processing 32
2.2.4 Length of Processing Unit: One Second 34
2.3 Basic Approaches and Principles of Our Solution 35
2.4 Chapter Summary 36

Chapter 3 Experiment Setup and System's Performance Evaluation 37
3.1 Experimental Setup 38
3.1.1 Test Environmental Sounds 38
3.1.2 Experimental Data Collection and Data Sets 39
3.1.3 Recognition Flow 40
3.2 Evaluation Approach: System's Accuracy and Power Consumption 42
3.2.1 Recognition Accuracy 42
3.2.2 Power Consideration and Evaluation 44
3.2.2.1 Algorithm's Evaluation in Power Consumption Aspect 44
3.2.2.2 Power Consideration in Previous Sound Recognition Algorithms .. 47
3.2.3 Our Evaluation Approach 48
3.3 Chapter Summary 51

Chapter 4 Mel-Scale Feature with LBG Classification for Environmental Sound Recognition 52
4.1 Introduction and Related Work 53
4.2 Sound Recognition Algorithm's Flow 55
4.2.1 Why Mel-Scale? 56
4.2.2 Feature Extraction – MFCC Flow 58
4.2.3 Why MFCC Can Have Less Overlap? 61
4.2.4 Classification – LBG Algorithm 62
4.3 Experimental Process and Consideration of Some Parameters 64
4.3.1 Experimental Setup and Details 64
4.3.2 Recognition Flow 64
4.3.3 Consideration of Some Parameter Values 66
4.4 Experimental Results and Discussion 67
4.4.1 Recognition Accuracy (Mel-filter Number, Frame Overlap) 68
4.4.2 Calculation Cost (Mel-filter Number, Frame Overlap) 69
4.4.3 Experimental Results 71
4.4.4 Performance Comparison and Whole System's Evaluation 73

4.5 Chapter Summary 76

Chapter 5 Low-Complex Haar-Like Feature with HMM Classification for Environmental Sound Recognition 77

5.1 Introduction and Related Work 78
5.2 Implementation of Sound Recognition by the Haar+HMM Algorithm 82
5.2.1 Why Employ Haar-like Sound Feature with HMM Classification? 82
5.2.2 Haar-like Sound Feature Extraction 84
5.2.3 Off-Line Training for the Haar-like Filters Group 88
5.2.4 HMM Classification 89
5.3 Experimental Process and Consideration of Some Parameters 91
5.3.1 Experimental Setup and Details 91
5.3.2 Recognition Flow 92
5.4 Experimental Results and Discussion 93
5.4.1 Parameters Tuning and Recognition Accuracy Rate 93
5.4.2 Comparison of Different Sound Features' Performance 95
5.4.3 Performance Comparison of Different Classifiers 97
5.4.4 Performance Comparison of Whole System 99
5.5 Chapter Summary 101

Chapter 6 Conclusions 102

6.1 Conclusions 103
6.2 Scope of Future Work 105

Bibliography 107
List of Abbreviations 120

List of Figures

Figure 1.1 An Example of Habitat Monitoring – "Great Duck Island" Project by Employing WSNs Technology. 5

Figure 1.2 MIT Media Lab's "Sociometer" Which Can Detect the Carrier's Physical Information and Notify Wearer's Location and Proximity. 7

Figure 1.3 Our Wearable Sensor Node Embedded Sound, Acceleration, IR Sensor in Size of Worker's ID Card (3.86 inch × 2.87 inch × 0.35 inch). 8

Figure 1.4 Main Components of a Wearable Sensor Node. 9

Figure 1.5 Flowchart of This Dissertation. 23

Figure 2.1 Wearable Sensor Recharging on a Charging Pad (a) and Inner Hardware Prototype (b). 25

Figure 2.2 Schematic Diagram of the Front-End Wearable Sensor Node. 26

Figure 2.3 Sound Recognition Flow. 30

Figure 2.4 Energy Assignment to the Three Main Blocks inside the Front-End Sensor Node – Rene. (Ref. [72_*L. Doherty*]) 33

Figure 3.1 Sound Matching and Recognition Flow. 41

Figure 3.2 Our Sound Recognition Performance Evaluation Approach – Average Accuracy and Power Benchmarks. 51

Figure 4.1 Sound Recognition Flow with the MFCC+LBG Algorithm. 56

Figure 4.2 Relationship of the Frequency *f* and Mel Frequency *Mel (f)*. 57

Figure 4.3 MFCC Algorithm Flow. 58

Figure 4.4 Mel Domain Diagram of Two Sounds - Train Start and Train Running (1.5-second length). 60

Figure 4.5 An Experimental Waveform That Explains Less Overlap in Sound Process Is Available. 61

Figure 4.6 Average Recognition Accuracy as a Function of the Template Length and LBG Codebook Cluster Number *k*. 67
Figure 4.7 Multiplication and Addition Calculation Cost as a Function of the LBG Codebook Cluster Number *k*. 67
Figure 4.8 Accuracy Rate in Function of Mel-filter Number and Frame Overlap. 69
Figure 4.9 Calculation Cost of Multiplication and Addition in Function of the Mel-filter Number and Frame Overlap. 70
Figure 4.10 Comparison of Optimized B, Referenced A and C's Accuracy and Calculation Cost. 72
Figure 4.11 Accuracy Comparison of the MFCC+LBG, MFCC+DTW, and MFCC+GMM Algorithms. 73
Figure 4.12 Performance Comparisons of MFCC+LBG, MFCC+DTW, and MFCC+GMM Algorithms. 75
Figure 5.1 Sound Recognition Flow with the Haar+HMM Algorithm. 82
Figure 5.2 One-Dimension (1-D) Haar-like Filter $h_{filter}(j)$ 84
Figure 5.3 One-Dimension (1-D) Haar-like Filtering for One Frame's Sound Signal. 85
Figure 5.4 Block Diagram of a Test Sound's HMM Classification. 90
Figure 5.5 Average Accuracy in Function of the Parameters: HaarFilNum and HaarWidMax. 93
Figure 5.6 Average Accuracy in Function of the Parameter: *H*aarFilNum and α. 94
Figure 5.7 Performance Comparison of Proposed Haar-like and Traditional MFCC Sound Features with Same HMM Classifier – Average Accuracy and Multiply / Addition Calculation Cost (256 samples/frame). 96
Figure 5.8 Performance Comparison of LBG, *k*-means and HMM Classifiers with Same Haar-like Sound Feature (Haar-like Feature's α=1.0). 97
Figure 5.9 Performance Comparison of MFCC+HMM, Haar+LBG, Haar+*k*-means, and Haar+HMM (Haar-like Feature's α=1.0). (* Ref. [93_*J. Nishimura*_ICSP'2008]) 100

List of Tables

Table 1-1: Approximate Data Rate of Different Sensor Modalities. 13

Table 2-1: DSP, FPGA, and MCU Technical Parameters' Comparison. 28

Table 3-1: Power Consumption Evaluation Methods in Different Design Stages. 47

Table 3-2: Main Electronic Parameters of the H8S/2218 MCU and Embedded H8S/2000 CPU Core. 49

Table 4-1: Performance Comparison of Optimized Case B, Reference Cases A and C with the Same MFCC+LBG Algorithm. 71

Table 4-2: Twenty Sounds Recognition Accuracy Confusion Matrix of Optimized Case B, Reference Cases A and C. 72

Table 4-3: Performance Comparison of the MFCC+DTW, MFCC+GMM, and MFCC+LBG Algorithms..... 74

Table 5-1: Comparison of Nishimura's Studies (Ref. [82, 93, 104]) and This Work. 81

Table 5-2: Training Haar-like Filters Pool Size with Relation to the Two Parameters - "*H*aarWidMax" and "*H*aarFilNum"..... 89

Table 5-3: Different Sound Feature – MFCC and Haar-like Feature (α=0, 0.5, 1.0) Performance Comparison (per Frame =256 Samples)..... 96

Table 5-4: Recognition Accuracy Confusion Matrix of 20 Different Tested Sounds with Haar+HMM Algorithm (α=1.0); Accuracy Comparison with Other Haar+HMM Two Cases (α=0/0.5), Haar+*k*-means and Haar+LBG. 98

Table 5-5: Comprehensive Performance Comparison of Four Different Sound Recognition Algorithms - MFCC+HMM, Haar+LBG, Haar+*k*-means, Haar+HMM (1 Second / *u*nit =124 Frames in Each Second Sound *u*nit, Haar-like Feature's α=1.0)..... 100

Chapter 1 Introduction

Firstly, the motivation of this research "Low-Complex Environmental Sound Recognition Algorithms for Power-Aware Wireless Sensor Networks (WSNs)" is introduced. A well-known bird habitat monitoring system in the "Great Duck Island" project is taken as an example to briefly introduce the WSNs system. At the same time, unique constrains in the WSNs system and research challenges, especially its front-end wearable sensors are introduced. The importance of activity recognition and reason to employ sound as a detection media are presented. Related researches of environmental sound recognition and its implementation on a hardware platform are also surveyed. Finally, the research targets and our contributions are concluded, outline of this dissertation is also delivered.

1.1 Research Motivation

Wireless sensor networks (WSNs) [1, 2, 3] becomes an active research area these years, its research results are gradually being applied in various fields and plays an important role for creation a ubiquitous information society around us. Its application ranges from initial battlefield surveillance to industrial fields, such as industrial monitoring, inventory tracking, and so on; and also to personal applications, such as household health care and elder-people caregiver systems [29, 91], etc.

To help realization these functions, employing the sound sensor embedded in some wearable device and recognizing personal daily activities are meaningful and challenging work. Background environmental sounds contain a lot of useful information to tell what activities people are doing. Through recognition these sounds continuously for a day, the people's daily activities log can be established. This log contains abundant information of individual self and between others. With the WSNs involvement, it is very helpful to

establish household medical systems like long distance diagnose for patients, physical and health monitoring for people in normal daily life, etc. The 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. However, these sound recognition algorithms executed on the power-aware WSNs platform are difficult because the power assigned for the signal processing block is very limited. Therefore, a so-called "smart sensing" which processes raw data and makes decision *locally* is absolutely required (Refer to Section 2.2.3). This demands the sound-context recognition algorithm to achieve high accuracy with low calculation cost to satisfy the energy requirement.

1.2 Wireless Sensor Networks and Front-End Wearable Sensors

1.2.1 Introduction of Wireless Sensor Networks

With the development of the micro-fabrication and integration, such as sensors and actuators manufactured using advanced micro-electromechanical system or MEMS technology, the transistors integrated in a IC chips has been doubled every two years based on Moore's Law and improved the computational performance by 70% every year. These advantages provide more low-cost and high performance front-end sensors which can sense fields and forces in our physical world.

Another, with the development of wireless communication, system software, hardware technologies that supports networks , and with sensor itself improvement in this decade make our scope of probing and understanding the outside physical world to an extent which human being has not even had before. Under this background, a new technology - wireless sensor

networks (WSNs) [1-6] arouses the researchers' great interest in both industry and academic fields.

These front-end sensors in the WSNs system mainly include sensing, data processing, and communication components. They can be self-organized and self-adjusted to build up a network and complete more complex functions than individual sensor does. Potential applications are described in references [2, 3] and specified as follow:

- Environmental and habitat monitoring (e.g., traffic, habitat, security monitoring) [3, 7]
- Industrial sensing and diagnostics (e.g., factory, inventory tracking)
- Infrastructure monitoring and protection [8, 9] (e.g. structural health monitoring)
- Battlefield awareness (e.g. multi-target tracking)
- Context-aware computing [10, 12-17] (e.g. intelligent home, responsive environment)
- Body Sensor Networks (BSN) [6].

1.2.2 An Application Example of the WSNs System

A well-known research for the "environmental and habitat monitoring" – the "Great Duck Island" monitoring system [103] is taken as an example to explicate the WSNs system. In year 2002, this project was initiated near the coast of "Great Duck Island" in Maine, USA by a combined research group from the University of California Berkeley and College of the Atlantic. The research target is to long-distance monitor the habitats of the local bird - Leach's Storm Petrel without personal interference by employing many locally embedded sensors and WSNs system.

The whole system is as Fig 1.1 shown, various type of sensor nodes called “motes” (marked as ① and ②) are embedded in and outside nest. They can measure the environmental temperature, light, infrared, relative humidity, and barometric pressure around the nest. The birds’ living environmental information is sampled, collected, and processed real-time inside the motes locally. Monitoring results and environmental information are transmitted by the mote’s transmitter to nearby Gateway (③) and to the faraway base station (④). Finally, the observation data is sent to remote lab in California through the satellite and internet (⑤).

In this way, the scientists who are not locally can learn the bird’s habitat information quickly. With the WSNs system involvement, disturbance from the researches was minimized compared with the traditional on-site study. Even now, this system is still working and we can learn the local environmental and habitat information from the internet [103].

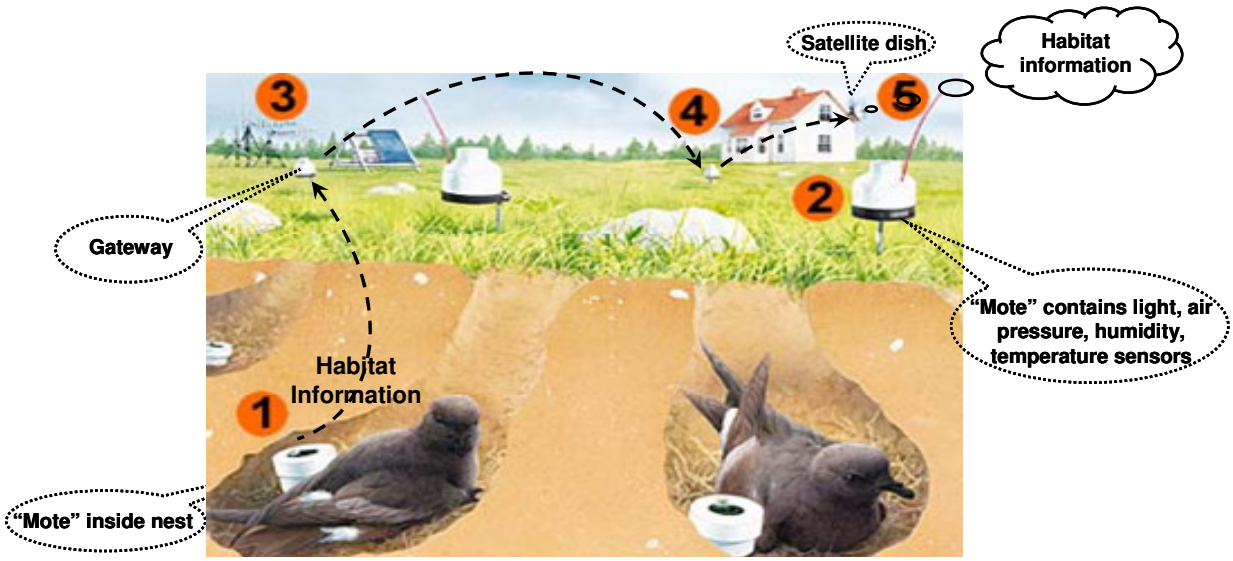


Figure 1.1 An Example of Habitat Monitoring – “Great Duck Island” Project by Employing WSNs Technology.

1.2.3 Front-End Wearable Sensor Node in the WSNs System

Previous Section 1.2.2 simply introduces what major components compose a WSNs system and how they work cooperatively to complete an environmental monitoring work. Similarly, multi-functional sensors can be integrated into a wearable device and applied to people. These wearable devices are easily and comfortable attached to human body. From them, carrier's activities, behaviors, and person-to-person's relationship information can be detected.

Supposed every member inside a society wearing these multi-functional sensors and with the WSNs involvement, detecting and understanding the individual's activities, person-to-person interaction of the society are available. This is of benefit to fulfill the "Community Detection and Social Behavior Analysis" and "Socially-Aware Computing" [10-19, 26, 27, 77, 78, 99] functions in the near future. Among them, the MIT Media Lab. and the Hitchai Ltd. research groups had developed their own front-end wearable sensor nodes.

It is reported that active pattern recognition of face-to-face interactions within a workplace can radically improve the function of the organization [20]. In order to improve it by detection the face-to-face interaction, researchers of the MIT Media Lab. developed some wearable sensors, such as "MIThril", "Uber-Badge" [11, 13], and "Sociometer" [12], etc. By using the "Sociometer" as Fig. 1.2 shown, ambient audio, acceleration information of the wearers can be sampled, processed and detected. The infrared ray (IR) sensor inside can inform the wearer's location and proximity. Therefore, individual and whole community's physical information can be collected, analyzed and conveyed, such as in an office, school or company. This community "networks" is helpful to understand their collaboration, team

formation, knowledge management, and dynamic communication conditions inside it. All these provide a powerful tool to understand and organize a dynamic human organization.

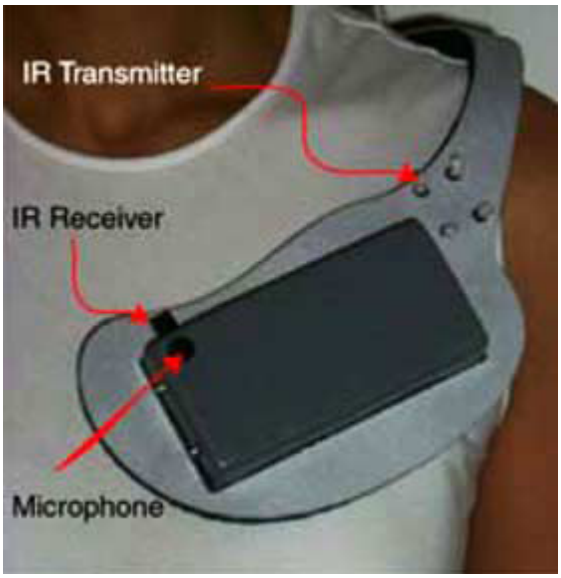


Figure 1.2 MIT Media Lab's "Sociometer" Which Can Detect the Carrier's Physical Information and Notify Wearer's Location and Proximity.

In order to perform the "context-aware computing" function for realization the ubiquitous society, Hitachi researchers have also developed their low-power wearable sensor nodes - "Life microscope" [15, 16, 18], "Business microscope" [15, 16, 19] and "Life Thermoscope" [17]. These designs also figure out a prosperous version of "knowledge-creating" and "opportunities-discovering" society in the near future by using these wearable sensors with the WSNs technology [15]. People's daily household and working information can be collected, analyzed, and well managed. These systems will inevitably enrich our life content and improve our society's efficiency.

A low-power wearable sensor - "Business microscope" [15, 16] is as Fig.1.3 shown. It is designed for understanding individuals and their interactive relations with others inside an

organization. The system uses an ID-tag-shape wearable sensor node that transmits and receives infrared light to detect face-to-face interaction between people. It can track individuals' movement using embedded accelerometers. At the same time, it can also detect and understand voice and ambient sound acoustic information inside a community by the integrated sound sensor. In this way, activities of all members within the organization can be sampled, collected, analyzed and illustrated. For example, this technology's application is great benefit to the employees' self-study and growth, the company's management and efficiency improvement [15, 16, 17, 21, 22, 23, 24, 63]. Besides inside a company, this low-power wearable sensor node can also be utilized at home. It is helpful to implement "household monitoring and assistance" and "household health monitoring and diagnose" functions.

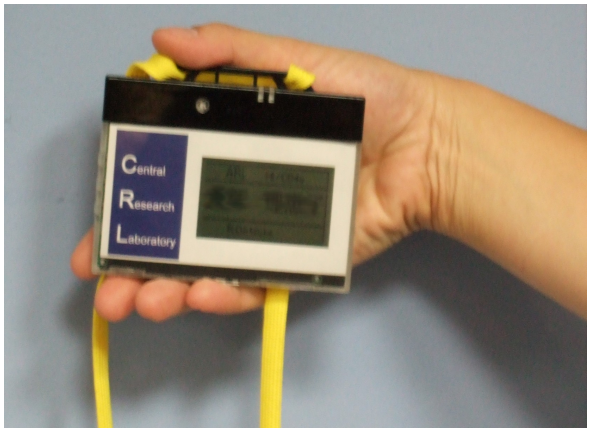


Figure 1.3 Our Wearable Sensor Node Embedded Sound, Acceleration, IR Sensor in Size of Worker's ID Card (3.86 inch × 2.87 inch × 0.35 inch).

From introduction of these front-end wearable sensors, we learn that their architecture is quite similar. They mainly consist of six major components as depicted in Fig. 1.4.

- MCU processor: the brain of the wearable sensor node.

- Wireless Communication: wireless communication between sensor nodes.
- Memory: external storage for sensor reading or program.
- Sensor Interface: interface with sensors and other devices.
- Power Supply: power provides for the sensor node.
- Operating System: software for managing the networks and resources.

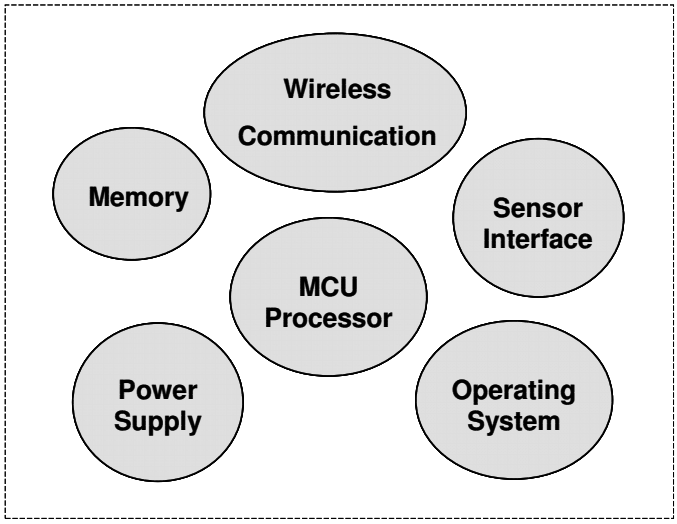


Figure 1.4 Main Components of a Wearable Sensor Node.

1.2.4 Unique Constrains and Challenges

Three main constrains lead to special research challenges during designing the WSNs system and applications [1, 2, 6].

- Limited support for networking: each node in the WSNs system acts as a router and as an application host. The network is peer-to-peer, mesh topology, and dynamic, mobile and unreliable connectivity. How to manage the networks effectively is a challenge work and have some discussion in [2, 4, 5, 9].

- Limited hardware: front-end sensor of the WSNs system has limited energy supply (on-board battery), memory, and communication capacity. Therefore, the methodology of signal processing, data storage and communication bandwidth of the WSNs system is different from the normal network system with a constant power supply.
- Limited support for software development: energy is limited in the WSNs system. For this reason, algorithms and network protocols need to maximize the system's lifetime, address robustness and fault tolerance, and self-configure.

Different constrains lead to different research problems in the WSNs system. In this research, we just focus on the first issue - energy supply and hardware resource assigned to our wearable sensor in Fig. 1.3 are limited (technical details are presented in the Chapter 2). This requires the applied algorithms upon the sensor platform must be operated within the very limited energy budget. Followed this energy constrains, the final performance must achieve to a reasonable and practical accuracy becomes meaningful. This is the difference from normal sound recognition research that mainly focuses on the recognition accuracy.

1.3 Environment Background Sound Detection for Activity Recognition

1.3.1 Low-Level Activity Recognition

Activity recognition [28] aims to recognize the actions and goals of one or more agents. It can be detected from a series of observations on the agents' actions and the environmental context conditions. It is a part of research field of pattern recognition.

There are three levels of activity recognition. At low-level activity recognition, relative information is collected by the sensors and processed locally or transmitted to higher level. At intermediate-level, statistical inference concerns about how to recognize individuals' activities from the inferred location and environmental conditions from the low-level. At the high-level, major concern is to discover the overall goal of an agent from the detected activity sequences through a mixture of logic and statistical reasoning. In this research, we focus on the low-level recognition by employing various kinds of sensors.

The low-level human activities can be recognized from two kinds of information. One is from the agent's body information. It can be sampled and collected by accelerometers [18, 30, 31, 32, 33], thermo-sensor [17], infrared ray (IR) sensor [12, 21, 23], etc. Another is from the context-aware sensing, for example, acoustic environmental sound [10, 13, 15, 25, 35, 36, 37, 38, 77, 78] and image processing [34]. These two ways are sometimes combined together with different functional sensors embedded into the wearable devices. Employing these easy-carrying movable devices is helpful to achieve better performance of tracking, monitoring and recognizing human daily activities.

Daily activity recognition has many important applications and plays an important role on improvement of personal and social life qualities. Health care is one of applications, such as nursing home for the elders, assisting the sick and disabled, fitness monitoring, etc. [6, 29]. Traditionally, people's physical and health information is acquired through self-reporting based on diaries or questionnaires from the doctors. This method is time-consuming and unreliable, especially for the elderly and subjects with memory impairment. Another method of acquiring this information is through clinical observation, but it requires expert's involvement and may not accurately reflect the patient's behaviors under the normal household environment. With the current advances in sensor and wireless technology, it is

now possible to provide ubiquitous monitoring of the subjects under their natural physiological status. The activity detection results can also be used to a person's behavior, intention, goal and social connection analyses. If these detection results are utilized in a company or an organization, every member's working status can be understood and illustrated. This can be a beneficial feedback for them. Their working efficiency can be improved and "healthcare" of the organization can also be realized [15, 21].

1.3.2 Why Applies Sound as the Detection Media?

Many detection media are used to recognize human activities, the most commonly used are acceleration [18, 30, 31, 32, 33], video [34], IR [12, 21, 23], and sound [10, 13, 15, 25, 35, 36, 37, 38, 77, 78], etc.

In Bao's work [30], five two-axis accelerometers were attached on the tester's joints and successfully recognized 20 human daily activities and achieve 84% accuracy. Work [31, 32] also used the acceleration sensor to detect people's abnormal activities which lead by Parkinson's or Alzheimer's disease. Through their reports, it can be conclude that the acceleration is mainly applied to detect individual activities. It is rarely employed to person-to-person social activity detection. Video is also widely used to detect people's individual and social activities [34]. Limitation of taking image as an activity detection media is at some unobtrusive situations, such as in hospital and toilet. In addition, image signal processing is more calculational complex than the acoustic signal processing.

In our research, the sound is chosen as the activities detecting media. Because compared with other detecting media, it possesses some unique advantages:

- *Wide detection scopes* - It can be person-to-person social activity detection, not like accelerometer is limited to the carrier's individual detection. IR detecting content is not as rich as sound sensing is. Sometimes, image detection quality is effected by the environmental illumination and setting position.
- *Convenient and comfortable* - Not like accelerometer and IR sensor which must be attached to the carrier, the sound sensor can be embedded into the background environment. This avoids inconvenience and discomfort of carrying the sensors.
- *Privacy protection* **-** Some of the people's activities are very personal, such as using toilet or taking bath. Applying sound as the activities detecting media is more suitable under these unobtrusive situations.
- *Appropriate calculation complexity* - For human being, sound is the second most important source of information after vision. However, image processing is more complex than sound processing. Image processing needs much more data rate and sampling rate than sound processing does as Table 1-1 shown. Therefore, sound processing is appropriate for the power-aware WSNs system.
- *Low cost* - From the Table 1-1, we notice that the number of the sound sensor is less than the accelerometers which makes the cost of sound system cheaper. Normally, sound sensor is cheaper than video sensor.

Table 1-1: Approximate Data Rate of Different Sensor Modalities.

Sensor Type	Sensors Num.	Sampling Rate	Resolution	Data Rate
Accelerometers	10	100 Hz	8bit	1k B/s
Sound	1	8k Hz	8bit	10k B/s
Video*	1	4608k Hz	8bit	4608k B/s

* Assuming 15 frames per second (fps) and VGA solution (640 x 480) x15=4608k B/s

As above described, sound is an ideal sensing media for the human activities recognition upon our wearable sensor platform. This is very helpful and promising for future integration with other sensor(s) to enhance daily activity recognition.

1.3.3 Application Domains

With the WSNs involvement, embedded acoustic sound and other functional sensors wearable devices can realize many applications. They can recognize the environmental background sounds happening around the people. These sounds contain a lot of useful information to understand what activity a person is doing. They also act as a social interactive "bridge" between people. Many applications can be built up based on the sound-context detecting results [11, 12, 15, 99].

Household Monitoring and Assistance is one of the main application fields. People's daily dietetic and sanitary information is hidden within the daily activities log. This is very helpful to understand the people's daily physical and health condition, and provides assistance to establish household medical systems, such as long distance diagnose for patients and elder-people health monitoring [29]. For example, we can deduce a person is having a food through chewing and drinking sounds. Toilet flush and urination sounds can indicate how often a person uses toilet, it is one of useful hints for doctor to diagnose the person's urinary system is abnormal or not. This household monitoring and assistance application is our main concerning. Therefore, in our experiments of later chapters, the detecting sounds are mainly targeted to the household events. The specific target test sounds are introduced in Section 3.1.1.

"Healthcare" of an Organization can be benefit form sound-context based human activities detection technology. Staffs of the organization's individual and social interaction, working status and efficiency can be illustrated from their daily activities log and other indicative methods [15, 16, 17, 18, 21]. Clearly understanding this information helps the staffs to realize their deficiencies during the working period and make improvement accordingly. This feedback-loop system inevitably improves the organization's efficiency and makes it more productive and healthier.

Social Aware and Communication is also an application domain for our sound-context detection. Acoustic voices and sounds are a rich information source for identifying the social behaviors and interactions. A good example for the group dynamics application is to find common favorite individual in the group. The utilized wearable device "UberBadge" mounted on each participants of the group employs sound sensor [11, 13]. A measuring interaction between people wearable sensor platform "Sociometer" shown in Fig. 1.2 is implemented with embedment of IR, acceleration and sound sensors [12].

1.4 Related Work

1.4.1 Environmental Background Sound Recognition

Some researches have been directed to recognize the environmental sounds happening around us [40-49, 80]. Most of them are algorithm level study and do not concern hardware implementation.

At the feature extraction stage, presenting the spectral envelope characteristic of a sound signal, linear prediction cepstral coefficients (LPCC) [46, 47] is a typical sound feature.

However, it can be clearly concluded that Mel-frequency cepstral coefficients (MFCC) outperforms the LPCC algorithm in normal sound recognition from previous work [43, 44]. Conventional state-of-art MFCC filtering is used to extract the sound feature and obtains good recognition accuracy [35, 41, 44, 45, 48]. However, computational expensive FFT is calculated before entering a bank of Mel-scale filters in the feature extraction flow. This increases the calculation complexity of sound feature extraction. Recently, in Chu's work [43], a new matching pursuit (MP) algorithm is 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 linearly combination with those atoms. Problem of the MP algorithm is that calculation cost for the searching enlarges dramatically with the number of the atoms in the dictionary increase.

At the classification stage, Cowling's work has a comprehensive comparison of most conventionally used classifiers [40]. Performance of the k-nearest neighbor (kNN), Gaussian mixture model (GMM), dynamic time wrapping (DTW), support vector machine (SVM), Linde-Buzo-Gray algorithm (LBG), *k*-means, and hidden Markov model (HMM) classifiers have been studied and compared.

1.4.2 Audio-Context Recognition on Hardware Platforms

Some researches about sound-context recognition based on DSP, FPGA, and MCU hardware platforms have been reported [38, 50-59]. Besides the system's recognition accuracy, these researches also consider how to implement the acoustic recognition algorithms on the hardware system.

A DSP system in Dong's work [57] is applied to execute sound environmental recognition for hearing aid application. A traditional sound feature extraction - Mel-frequency cepstral coefficient (MFCC) with hidden Markov model (HMM) classifier are implemented upon the DSP system. In this work, the complicated MFCC-based sound feature with HMM classification is implemented on the Ezairo 5900 SoC system. A 24-bit specific DSP IP core is employed to process the acoustic environmental sounds. For our power-aware wearable sensor, to execute these complex algorithms is difficult.

An interesting system upon a combined DSP and MCU hardware platforms to realize acoustic scene analysis is carried out by a research group of Arizona State University [52, 53, 54]. The system can process the acoustic signal which is sampled by the front-end sensor. Sampling data are transmitted through a RS232 serial link to the attached DSP board to process. These pre-processed acoustic features are wirelessly transmitted to the base station, and some functions are implemented inside the station. This system can fulfill speech/non-speech, gender (male or female) recognitions and other functions. In fact, these achieved various functions are at the cost of abundant power supply (power adaptor) in DSP board and base station. Moreover, all the classification is executed inside the base station, not inside the front-end sensor locally. This proposed system is not suitable for our wearable application.

MIT media center group basically completes social dynamics detection [10, 11, 12, 13]. Their wearable front-end sensor nodes - "UbER-Badge" [13] and "Sociometer" [12] have been developed. Each member carrying these wearable sensors inside a social organization can build up a social dynamic sensor networks. Social connections and interactions between them can be detected and understood. Both "Uber-badge" and "Sociometer" employ acceleration, microphone, IR sensors to complete the functions. The "Uber-badge" has four

AAA batteries with 100mA average current and continuously works for 15 hours. The microphone samples with 8-bit resolution and 8 kHz sampling rate. Only simple background sound's average and difference of the amplitude values are calculated to indicate the carrier's dynamic, such as during lunch, dinner, and buffet break these social dynamic moments. However, this work just achieves a rough function of understanding acoustic context around the people. As to comprehensively understand the detail acoustic context, their proposed sound feature is simple and does not work.

Researcher of the Waseda University employs some sensors called "Cookie" and "Muffin" [55, 56] to build up their sensor networks. This front-end wearable system can detect person's daily activity from 11 genres sensors (microphone, RFID, pulse, 3-D acceleration, etc.). These sampling data are transmitted to a host mobile terminal (such as cellular phone or a PDA) that provides enough power to analyze and detect the carrier's background context. The recognition software and hardware infrastructures have been introduced. Multi-sensors fusion and hierarchical context refinement methods help to complete context awareness function. However, as the system's power consumption and how long the system maintains, the authors do not provide enough explanation and research effort. In fact, power consumption is one of important factors during the algorithm's implementation on power-aware front-end wearable sensor node.

1.4.3 Tradeoffs of the Sound-Context Recognition on Wearable Platform

The design of a wearable computing system needs to consider various factors: low power, high performance, easily wearable, efficient communication channels, etc. In our research, we focus on the sound-context recognition algorithm's study. It must achieve acceptable

recognition accuracy and satisfy the power budget assignment of our wearable sensor [16, 18].

Including the algorithms introduced in Section 1.4.1, most reported environment sound recognition researches don't need to consider the hardware factor. In Chen's work [35], seven bathroom activities are recognized by detecting sounds happening in it, such as shower and brush tooth sounds, etc. The sounds are sampled by a microphone set on site and recognized by utilizing the MFCC+HMM algorithm on a PC afterwards. Similar cases also happen in acceleration-context activities study. Yin's Work [31] uses the acceleration sensor to detect people's abnormal activities which lead by Parkinson's or Alzheimer's disease. Even though the experiment raw data were sampled by some wearable sensors and transmitted to the computer, the recognition is not completed inside the sensor nodes locally, whereas inside the computer.

Comprehensively trading off the recognition system's performance and power consumption research was firstly reported by the ETH research group in 2004 [36, 38]. Their researches are most close to our research. They employed wearable accelerometer and microphone embedded sensor - ETH PadNET to detect carrier's activities happening in a wood shop [36, 38, 50, 51]. Recorded sounds are sampled at 48 kHz and down-sampled to 2 kHz, frame based FFT feature extraction is executed and linear discriminant analysis (LDA) is applied to decrease the feature's dimensions. Twenty-one sounds happened in wood shop, such as from filing, sawing, drilling, and hammering, etc. can be detected with combination of the carrier's acceleration information. In Stager and Bharatula's work [36, 38], how to trade off the accuracy and power consumption of a sound-based context recognition system upon a wearable platform is reported. Free combinations of nine time-domain features (mean, variance, etc.) and five frequency-domain features (bandwidth, frequency centroid, etc.)

constitute sound feature sets. With different classifiers, different recognition results are yielded. A target sound feature set and classifier is decided by the accuracy and power consumption's tradeoff. However, to explore this ideal sound feature set and classifier needs an empirical and complicated training process.

Power efficiency plays a crucial role for those wearable devices in WSNs system [61]. In our work, a sound sensor embedded in the power-aware wearable sensor node is utilized to recognize the environmental background sounds. Power supply for the wearable sensor is energy limited battery, not like DSP and FPGA board with adaptor power. Conventional sound recognition and acoustic signal processing algorithms which can be executed on the DSP or FPGA [57, 62] platforms may not perform well on our wearable sensor. Therefore, how to develop a new sound recognition algorithm to achieve high accuracy with low calculation cost to satisfy the energy requirement is the challenge of this research.

1.5 Research Objects and Contributions

Environmental background sound is a good context indicator for human activities, and contains rich information for identifying individual and social behaviors. Therefore, many front-end wearable devices in the WSNs system with sound recognition function are widely used to trace and understand human activities. Because those front-end sensor nodes are low-powered and the WSNs system has limited resource, these limitations decide our unique research objects:

1: the sound-context detection function in front-end wearable sensor node should work *continuously* for 24 hours for a whole day observation.

2: the sound-based context recognition algorithms should be *local* processing. This can save energy than wireless transmitting the raw data to upper server to process.

3: the local processing decides the sound recognition algorithms must be of low computational complex. Therefore, our developed algorithms should achieve high recognition accuracy while still be with low calculation cost to satisfy our wearable sensor's power requirement. This is the difference from the normal sound recognition researches of which mainly focus on the recognition accuracy.

In order to complete the above mentioned research objects, we make efforts and achieve these goals in this research.

Our power-aware front-end wearable sensor node inside the WSNs system shown in Fig. 1.3 has been thoroughly studied. Upon this resource limited platform, the assumptions and special constrains of this research are analyzed and discussed. Especially the local environmental background sound detection is the most crucial problem which must be solved.

After understanding the system's limited recourse provided for our sound recognition algorithms, both the final detection accuracy and its power consumption are considered as the evaluation approach to those candidate algorithms. The target values of these two factors have been discussed and decided.

Two of our proposed sound recognition algorithms – MFCC+LBG and Haar+HMM are studied. Their recognition accuracy and approximate power consumption for execution these algorithms upon the wearable sensor node are also evaluated.

1.6 Thesis Organization

This dissertation is divided into six chapters, and its flowchart is shown in Fig. 1. 5.

In chapter 2, the power-aware wearable sensor's hardware platform and software-level sound recognition flow are introduced. Assumptions and constrains of this research are also presented and discussed. Based on the introduced resource limited sensor node platform, basic solution approaches to satisfy both the recognition accuracy and energy budget requirements are proposed.

In chapter 3, experiment details including target detected 20 sounds, training and detected experimental data sets and recognition flow are introduced. Two evaluation benchmarks - the accuracy expected to achieve and computational power budget for the applied sound recognition algorithms are discussed and decided.

In chapter 4, sound feature extraction Mel-frequency cepstral coefficients (MFCC) and vector quantization (VQ) classification Linde-Buzo-Gray algorithm (LBG) algorithm is applied for the sound-based context recognition. How three parameters (i.e., Mel filters number, frame-to-frame overlap and LBG codebook cluster number) of the algorithm affect the system's calculation burden and accuracy is investigated. Based on the performance evaluation method in Chapter 3, the comprehensive performance of proposed MFCC+LBG algorithm is evaluated.

In chapter 5, an extreme low calculation sound feature extraction Haar-like filtering with hidden Markov model (HMM) classification algorithm is newly proposed and applied to recognize the environmental sounds. Through experimental comparison, the proposed method outperforms other normally utilized sound recognition algorithms as the recognition accuracy and calculation cost two evaluation parameters concerned. Average recognition

accuracy 96.3% of 20 typical daily activity sounds can be achieved. At the same time, it also satisfies the amount of calculation cost decided by the wearable sensor node's energy resource.

In chapter 6, we conclude the dissertation and also discuss potential directions for future work.

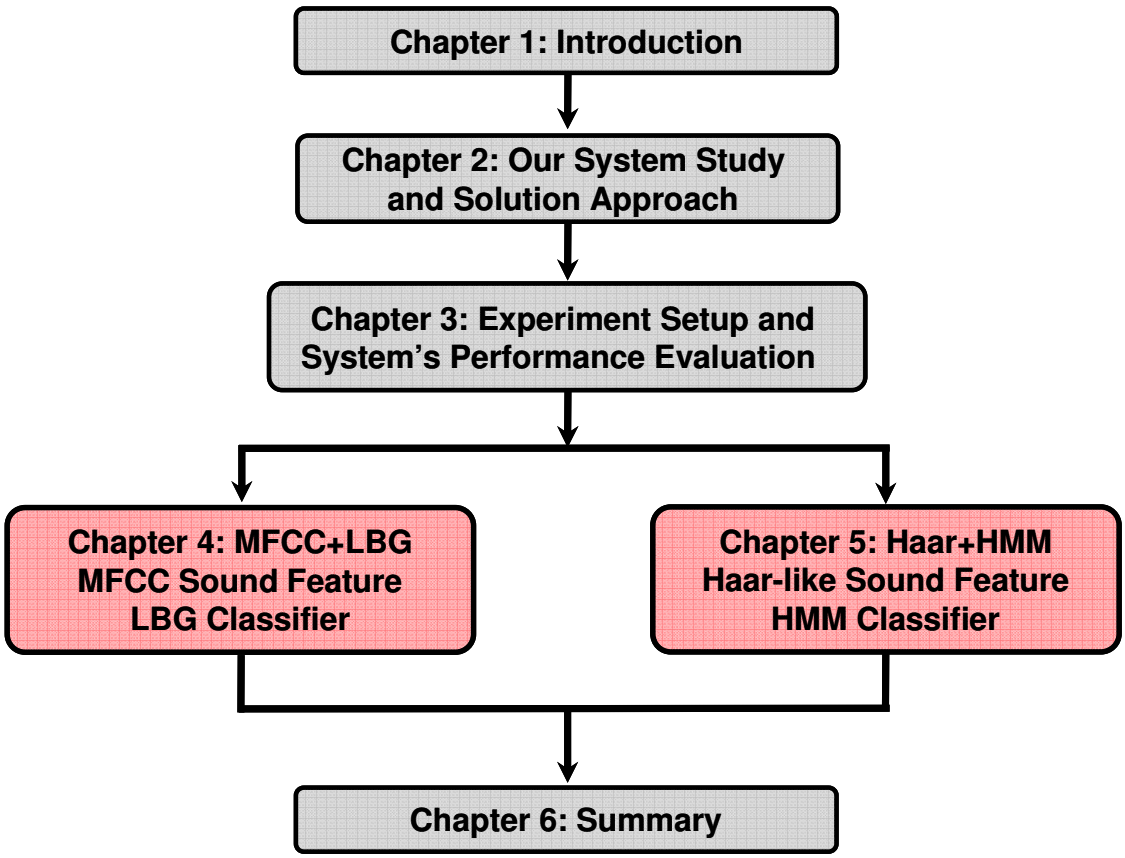


Figure 1.5 Flowchart of This Dissertation.

Chapter 2 Our System Study

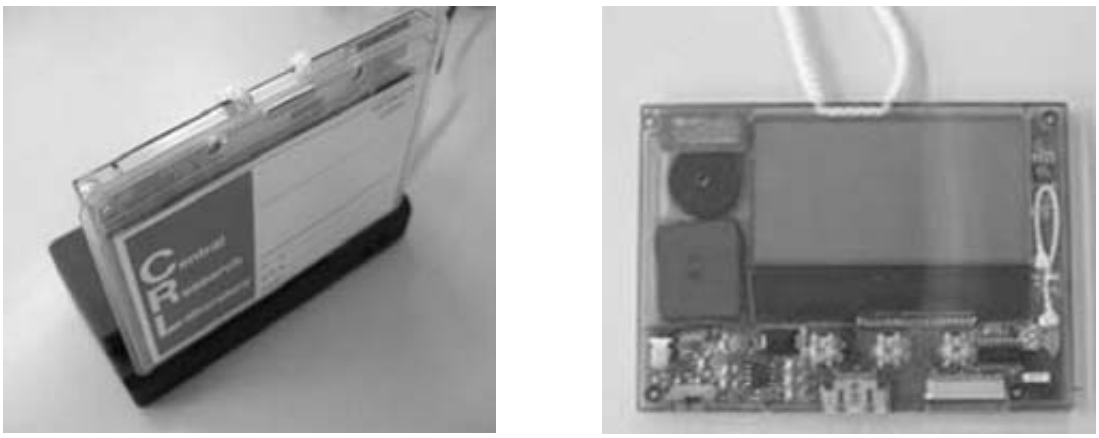
In order to achieve our sound-based activity recognition upon the power-aware wearable senor node, we must have a clear understanding of the wearable system. Therefore, in the first part of this chapter, the system's hardware-level architecture and software-level sound recognition flow of this research are introduced. Next, based on the introduced resource limited sensor node platform, important assumptions and constrains for this research are presented and discussed. Finally, aiming at achieving certain high recognition accuracy with limit assigned power, our basic approaches are proposed.

2.1 Our Hardware and Software System

2.1.1 Hardware Platform and Specifications of Our Wearable Sensor

2.1.1.1 Hardware Schematic Diagram

The wearable sensor used in our search is provided by the Hitachi's Central Research Laboratory.



(a) **(b)**

Figure 2.1 Wearable Sensor Recharging on a Charging Pad (a) and Inner Hardware Prototype (b).

Recharging status and inner hardware structure outlooks of the wearable sensor are indicated in Fig. 2.1. From the inner hardware outlook indicated in Fig. 2.1 (b), we notice that the sensor node mainly includes:

- Various types of sensors (acceleration, sound, temperature, etc.)
- Analog to digital converter (ADC)
- RF communication module (IEEE 802.15.4 wireless communication protocol)
- Micro control unit (MCU) processor which contains CPU for calculation.
- Li-ion battery power supply

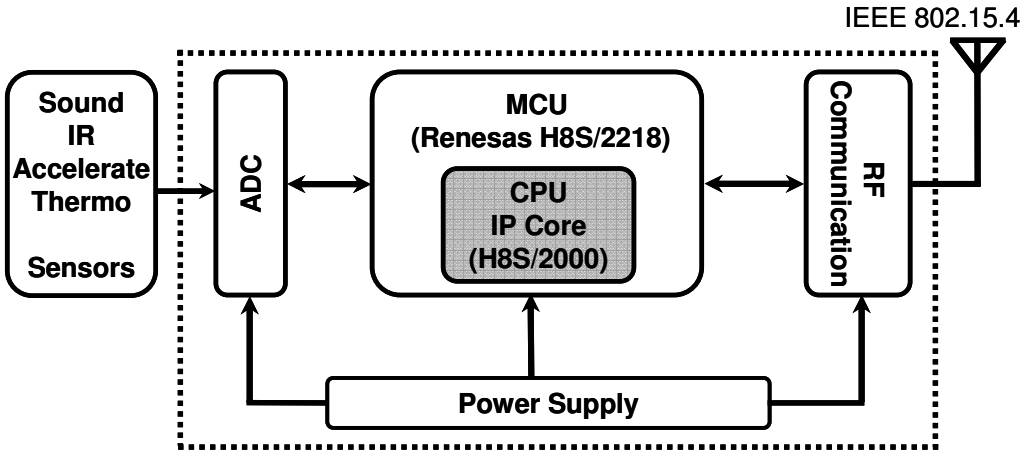


Figure 2.2 Schematic Diagram of the Front-End Wearable Sensor Node.

The schematic diagram of the front-end wearable sensor is illustrated in Fig. 2.2. Three blocks mainly consume the sensor's limited energy: ADC block, communication block, and MCU microprocessor [16, 38, 61, 72].

ADC block can sample and convert a continuous physical environmental analog signal into discrete digital signals for later processing. Optimized low power MCU processor [64] processes the converted signals, and can complete some control functions based on processed

results. The technical details of our MCU processor are introduced and discussed in Section 2.1.1.2, Section 2.1.1.3 and Section 3.2.3. RF and communication block are in charge of exchanging information between other sensors and with upper gateway nodes. Because this sensor node is applied to the power-aware WSNs system, this limitation decides the communication employs low-rate IEEE 802.1.4 and ZigBee protocols [6, 16]. The sensor node is with Li-ion rechargeable battery powered because of its superior discharge characteristics at high current as well as high energy density [16].

Inside the MCU processor, there is an embedded low-power CPU core in which our proposed sound recognition algorithm is executed. The algorithm is executed by individual addition and multiplication operations in the CPU [65, 66, 67]. In this work, we focus on sensor nodes which perform the whole environmental sound recognition process locally – from signal acquisition to classification. Thus, the challenge is to develop a sound-based context recognition algorithm upon the power-aware wearable sensor node. Executed algorithm in the MCU should guarantee certain recognition accuracy, and on the other hand satisfy the energy requirement.

2.1.1.2 Why MCU? DSP, FPGA, and MCU Comparison

To utilize which kind of processor as the processing and control unit decides the system's performance and cost. Comparison of candidate processors - DSP, FPGA, and MCU is listed in Table 2-1. Because our sensor node is wearable and battery power supply for the application is limited, therefore, the low-power MCU can be an appropriate choice. The MCU's energy requirement is less than that of DSP and FPGA. Another attractive advantage is that the price of MCU is cheaper compared with the DSP and FPGA. Therefore, in our

research, a middle class MCU in the Renesas H8S series had been decided and used inside our wearable sensor [16, 64].

Table 2-1: DSP, FPGA, and MCU Technical Parameters' Comparison.

	Data (bit)	ROM/RAM (KBit)	Working Freq. (MHz)	Supply Voltage (V)	Average Current (mA)	Energy (mAh)	Power Source	Price (piece)
Ultra-Low-Power TI_DSP(TMS320 C54X)[a] (Audio Processing) [d] (0.18 μm process)	16	4~256(ROM) 5~640(RAM)	66~300MHz 600MIPSmax	3.3V I/O 1.8V Core	64mA	650mAh [d]	Power Adaptor ---------- Battery	5~75 USD
Xinlinx_FPGA (Spartan-6)[b] (40 *nm* process)	1~64 (customizable)	Bigger than DSP & MCU (customizable)	50M~500MHz (configurable)	3.3V I/O	Application Dependent (X0 mA)	Application dependent	Power Adaptor	20~200 USD
Renesas_MCU (H8S/2218)[c] (0.35 μm process)	16	128/12	4~24MHz	3.0~3.6V	6mA	150mAh	Battery	7~10 USD
CPU core inside the 2218 MCU (H8S/2000)	16	X	20MHz (50ns/Cycle)	1.8V	4mA	10mAh*	X	X

a: TI_TMS320C54x data sheet (www.ti.com) [Ref. 68]

b: Xinlinx_FPGA_Spartan-6 data sheet (www.xinlinx.com) [Ref. 69]

c: Renesas_MCU_H8S/2218 data sheet [Ref. 64]

d: Refer to book "The application of programmable DSPs in mobile communications" [Ref. 70_*A. Gatherer*]

***:** 10mAh is the energy assigned for the sound processing module in H8S/2000 CPU.

X -- none

2.1.1.3 Hardware Specification

Our wearable sensor node (Fig. 1.3 and Fig. 2.1) contains many types of sensors and modules. They are illuminometer, thermometer, microphone, accelerometer, infrared rays (IR) transceiver, RF communication module and MCU processor [16, 19, 21].

- Approximate 150mAh battery energy is the sensor node's total energy budget.

- Among the 150mAh energy, 10mAh is assigned to the microphone's acoustic signal processing. (Battery resource inside our sensor node is limited. From Table 2-1, it is obvious that the 150mAh energy is much less than a typical mobile acoustic codec mode's energy consumption on a DSP platform as item "d" indicates.)
- The MCU is Renesas Technology's H8S/2218 chip [16, 19, 64]. It is a microprocessor with 35μm process, 16-bit architecture, 65 basic instructions, 6mA working current, and 3.0~3.6V 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 20MHz (50ns per cycle), 1.8V input voltage with 4mA average working current.
- Size of ROM = 128 KB, size of RAM = 12 KB.

We prefer the sound module in the sensor node could continuously work for 24 hours (3,600×24=86,400 seconds), and CPU core can finish the sound recognition algorithm within *each* second sampling. Therefore, the recognition results can help to capture a person's activities for a whole day with one second unit recording.

2.1.2 Recognition Flow in Software Aspect

Based upon the hardware platform shown in Fig. 1.3, Fig. 2.1 and Fig. 2.2, sound sensor inside the front-end wearable node can sample environmental sounds happening around the people. As the Section 2.2.3 analyses, the sound recognition algorithm should be locally processed. The local recognition algorithm flow is as Fig. 2.3 indicated.

It includes two steps sequentially: off-line sound templates generation and on-line sound classification.

With training data set, features of the template sound can be extracted. 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 feature extraction method. Following this, the recognition result is finally achieved by comparison with the prepared templates by using certain classification method.

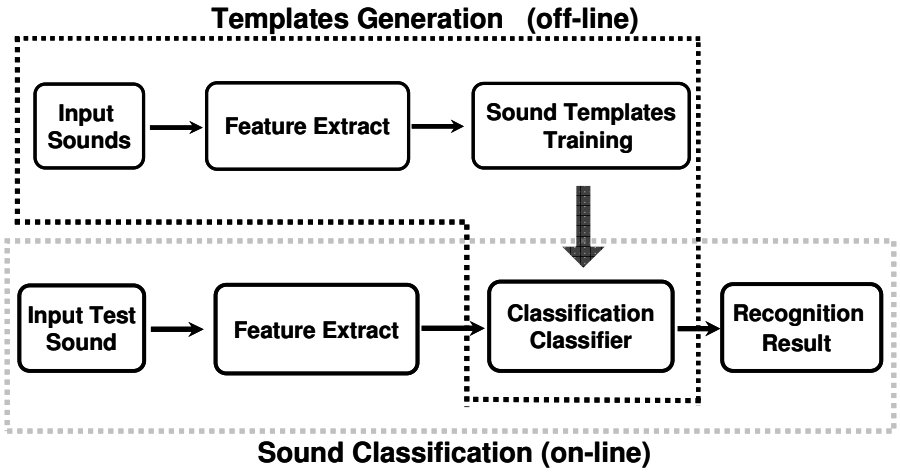


Figure 2.3 Sound Recognition Flow.

2.2 Assumptions and Constraints of This Research

2.2.1 Placement of the Wearable Sensor

Various genres of sensor are employed to recognize human activities. Placement of the sensor depends on specific research [12, 30, 31, 33, 36, 71], and mounted place on body has an obvious effect on final results. E.g., the accelerometers are often mounted on some body

joints when executing some activity detections [30, 31, 33, 71]. Only at these joints, the detected unique action's characteristics feature can be embodied and well collected.

As utilizing acoustic sound to complete the human activities detection, Stager's work [36] focuses on some activities happens in a small working mill and kitchen. The microphone – sound sensor is attached on participator's arm. Choudhury's work in MIT [12] employs a "Sociometer" wearable sensor to understand the connections between members inside a social group. The "Sociometer" contains IR, microphone, and acceleration three kinds of sensors. It is attached on the participator's chest. With the wearable sensor node shown in Fig. 1.3, our case is similar to Choudhury's work. The node is hung on the tester's chest as an ID card or set under test environment (such as testing shower action). In this way, the sound sensor can sample the environmental sounds information to a maximum extent.

2.2.2 Dominant and Single-Content Sounds

Our test target sounds are required "dominant" and "single-content". There are two extreme situations which are not included in our research scope.

One is the real target recognized environmental sound is dominated by noise or other sounds, it is undetectable. For example, a vacuum is operated by mom while her kid is brushing teeth. From the activity detection angle, the kid can't brush and operate vacuum at the same time. Among the kid's daily activities, possibility to encounter his mom to use vacuum while brushing teeth is very low. Therefore, our test target environmental background sounds are often caused by a single action, and they are defined dominant and detectable. By "dominant", it means that the test sound is the loudest and dominant sound received by our system. Our target 20 test sounds are listed in Section 3.1.1 specifically. They

are recorded in real environment, not in a noise-isolated room. It means our sampling sounds data comprises both real test sound and background noise.

Another extreme case is the test sound is composed of multiple components sounds. For example, a TV program is broadcasted when a person is on shaver. It is happens at home sometime. To detect shaving activity, we must recognize the shaving sound overwhelmed by the TV sound. However, extracting the target sound from a mixture sound is another research topic. It is beyond this research scope and not our research focus.

2.2.3 Local Processing

As previous Section 2.1.1 introduced, three blocks inside the front-end sensor mainly consume the limited assigned energy from the WSNs system. They are ADC, communication, and MCU microprocessor three blocks. Among them, ADC and the communication blocks consume most energy; the remaining for the MCU is limited [16, 38, 61, 72].

A "Traffic Tracking Scenario" employed the WSNs technology in reference [72] illustrates proportion of each block within the whole energy budget. In this project, Rene sensor node's ADC sampling (10-bit) energy is 375pJ/sample, communication energy (short-range) is 800pJ/byte, and computation is 1pJ/instruction. Three blocks work together for one second with 30 Hz sampling rate and communicate 30-bytes message to the network. The required energy in each block is as below described, and their proportion is as Fig. 2.4 shown.

- 30 Hz sampling × 375pJ/sample = 11.3 nJ — ADC sampling (31.1%)
- 30 bytes/message × 800pJ/byte = 24nJ — Communication (66.1%)
- 1000 instructions/sample × 1pJ/instruction = 1nJ — Computation (2.8%)

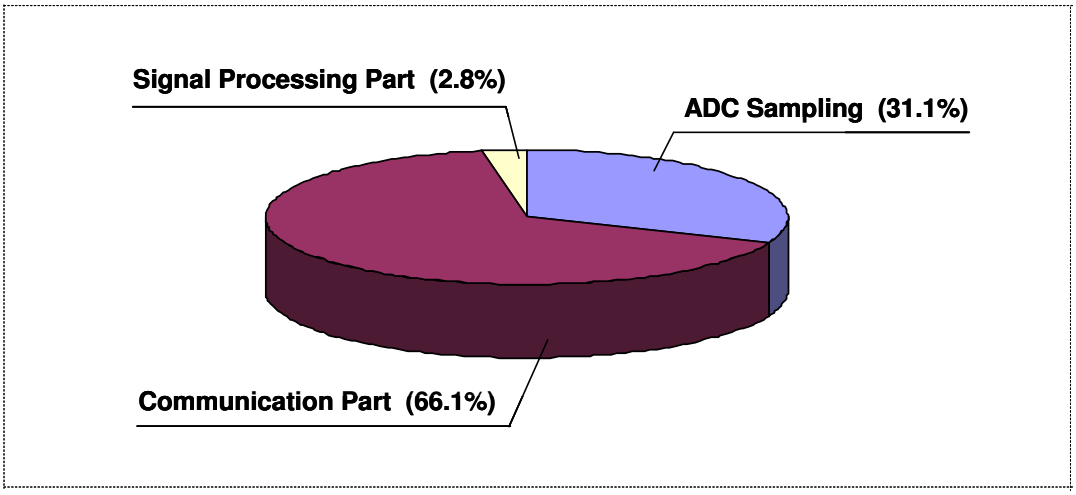


Figure 2.4 Energy Assignment to the Three Main Blocks inside the Front-End Sensor Node – Rene. (Ref. [72_*L. Doherty*])

Normally, there are two ways to process the sampled data from the front-end sensor node in WSNs. One is to deliver the sampling data to the sink node directly without any processing, and then transmits them to a server. Finally, the data are processed in the more computational sever. Another is to process the data inside the sensor locally and only transmit the final classification result.

In this research, the local processing is adopted basing on two following considerations:

- As indicated in Fig. 2.4, the ADC and communication blocks spend most energy, and the remaining energy for MCU processor is limited. Moreover, wireless communication generally consumes more power than computation [61, 72]. Therefore, locally processing the sampling sound data is more power efficient than wirelessly transmitting them to the upper-level server to process.
- The limited communication bandwidth is a crucial problem in WSNs system [6]. Transmitting the sound raw data must occupy more bandwidth than transmitting the

final processing results. Therefore, the local processing is helpful to save some bandwidth resource.

In the following Chapter 3, the necessity of processing the sampling data locally from the hardware and energy aspects is justified in detail. The executed acoustic environmental sound recognition algorithms in the MCU of power-aware WSNs system should be simple and effective. Therefore, it can satisfy both the high recognition accuracy and low power consumption requirements. This is the most challengeable task of our research and our solutions are explicated in the following Chapters 4 and Chapter 5.

2.2.4 Length of Processing Unit: One Second

In the Chapter 1, one of our research targets is defined as: the environmental sound recognition module in the wearable sensor node could continuously work for 24 hours. Therefore, the detection results can capture and understand a person's activities for a whole day without any interruption. Especially, the proposed method sometimes meets some extreme situations. For example, a fire siren alerts when the fire happens. Our sound recognition system should detect the siren sound within a very short time, and inform the carrier immediately. Similar requirement happens in household medical health monitoring and elderly care.

Above two examples require our sound's segmentation method should not adopt "duty cycle" method in reference [19]. Otherwise, 10 seconds interval with 0.1 second sampling "duty cycle" may miss some important activities taken place within the neighboring samplings. For example, this "duty cycle" method may miss sampling abnormal walking

sound if a stroke happens between the neighbor samplings. Therefore, it needs continuous sampling based on our applications.

At the same time, time for the sound recognition processing must not be long. The recognition must be finished within the sampling interval (processing unit) or before the coming sampling. Otherwise the carrier will miss precious reaction time. These require our sound recognition algorithm must be simple and effective. From the previous researches [33, 36, 43, 45, 48], the appropriate processing unit is in "second" scale is clear. And the sound's processing unit decides the final recognition accuracy. Comprehensively considering the energy assigned for the microphone in Fig. 1.3 and accuracy two evaluation factors, "*one second"* of processing unit is adopted in our research.

2.3 Basic Approaches and Principles of Our Solution

Aiming at achieving certain high recognition accuracy with limit assigned power by our environmental sound recognition algorithms, two basic approaches are proposed.

- The sound feature should be simple. In this way, each sound's template after the training stage will occupy little memory. This can save memory which is a precious recourse in the WSNs system. On the other hand, simple sound feature extraction brings low calculation cost in the on-line classification stage. This leads to low power consumption accordingly when executing the algorithm inside the sensor's MCU.
- The sound feature with the classifier should be effective which can achieve certain high level accuracy. Simple sound feature decreases calculation cost, however, it also brings low recognition accuracy problem. However, this disadvantage can be

compensated with high performance classifier [40]. That means the system's comprehensive performance (accuracy and power consumption) is decided by the feature and classifier's combination. For example, in the later Chapter 5, simple and low cost Haar-like feature with high performance HMM classifier can achieve high accuracy with reasonable calculation complexity. This is helpful to implement the Haar+HMM algorithm upon our power-aware sensor node.

2.4 Chapter Summary

In this chapter, the power-aware front-end sensor's hardware platform and software level recognition flow are introduced. The hardware specification used for evaluation of power consumption in the following Chapter 3 is specifically discussed. Important assumptions and constrains for the research are also discussed and explained. Finally, aiming to achieve a certain high accuracy with less power consumption by our sound recognition algorithm, we propose our basic approaches.

Chapter 3 Experiment Setup and System's Performance Evaluation

By utilizing the wearable senor node in previous chapter introduced, our experimental setup is introduced in the first part of this chapter. Test environmental sounds and data sets for the training and test are described. In the second part, we proposed our evaluation approach for the sound recognition algorithm(s) executed on the power-aware wearable sensor. That is our recognition algorithm should both satisfy certain detection accuracy, and less power consumption within the budget simultaneously.

3.1 Experimental Setup

3.1.1 Test Environmental Sounds

There are many activities in our daily life. We can distinguish what people are doing under what kind of environment by analyzing the activity's background sound. Sounds chosen for the experiment are listed below. Most of person's daily activities at home are included.

1: Vacuum cleaner (house cleaning)

2: Washing machine (wash something)

3: Water sound from tap -- Household Clean

4: Brush teeth

5: Shaving (shave beard)

6: Taking shower

7: Hair dryer (dry hair)

8: Urination (man)

9: Flush toilet (use water closet) -- Household Sanitary

10: Chewing cake (eat)

11: Drinking (drink something)

12: Oven-timer (toast some food) -- Household Dietetic

13: Walk inside room

14: Walk outside

15: Run

16: Train start (train accelerates, in train)

17: Train run (train normally runs, in train)

18: Rain hits Umbrella (in the rain) -- Outside Acts

19: Mechanical alarm

20: Telephone rings -- Others.

3.1.2 Experimental Data Collection and Data Sets

Main parameters of sound sensor inside the front-end wearable node shown in Fig. 1.3 are:

1: sampled sounds are mono

2: 16,000 samples/second (16 kHz sampling rate)

3: 16-bit resolution sampling

4: 256 samples/frame

The sampling mode of the wearable sensor node 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 tester's chest. Those sounds are

recorded in real environment, not in a noise-isolated room. The experimental sound data are stored in the sensor node's on-board memory, and used for their templates training and test inputs.

With the wearable sensor node introduced, above introduced 20 sounds were recorded for templates training and input test two data sets. Each of above mentioned 20 sounds were recorded more than three times. The recording experiments were operated on different days by different testers. All these sampling data taken by different testers were mixed. Among many recordings of each sound, one recording was randomly taken as the test input. Those different 20 test input sounds compose the testing data set. Their lengths vary from 14.9 to 170.5 seconds (indicated on the 2nd column of the Table 4-2 and Table 5-4). Left others are used as training data sets, their length vary from 16 to 277 seconds, total length is 1788 seconds. In this way, unfairness in the training and testing steps can be avoided to a great extent.

As Section 2.2.4 discussed, each unit length of detected sound is one second during the recognition process. This means the proposed algorithms for our sound recognition should finish within each one second.

3.1.3 Recognition Flow

The environmental sound recognition diagram is as Fig. 2.3 indicated. It contains following three major stages.

Stage1: Training sound templates

Step1: We have taken 1.0second length sound as a unit to partition same property sounds from the training data sets. By applying the sound feature extraction algorithm to these units, the feature vectors for the following training are prepared.

Step2: With the sound template training process, the *Step 1*'s feature vectors are trained and generate its template.

Step3: Repeat above two steps to other 19 sounds, these remaining sound templates can be completed. Therefore, detected 20 different environmental sounds have their unique template accordingly after the training process. This provides possibility for different test sounds classification and matching.

Stage 2: Sound matching

The sound matching flow is as shown in Fig. 3.1. When the test sound is input, it is also segmented as 1.0-second units and executed feature extraction. Because those 20 sound templates have been prepared in memory after the training stage, the classification process is applied to match the input sound, and then the closest one among the 20 sound templates is selected.

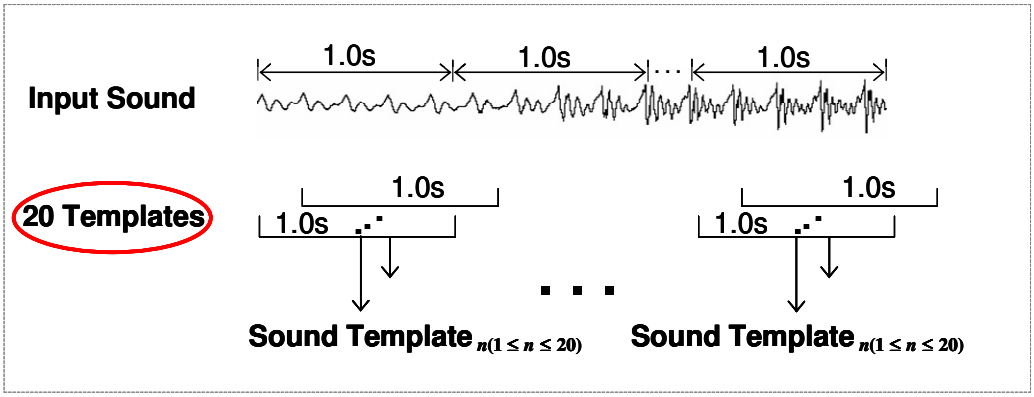


Figure 3.1 Sound Matching and Recognition Flow.

Stage 3: Calculate recognition accuracy and evaluation

The recognition *A*ccuracy *R*ate (*AR*) is defined as:

$$AR = \frac{C_u}{A_u} \times 100\%, \quad \textbf{(3.1)}$$

where C_u is the number of *C*orrectly recognized *u*nits, A_u is the number of *A*ll input sound *u*nits.

3.2 Evaluation Approach: System's Accuracy and Power Consumption

Locally completing the sound recognition inside the wearable sensor node 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 another word, the sound recognition algorithm executed inside the sensor node should simultaneously satisfy both the recognition performance – certain high detection accuracy and less power consumption.

3.2.1 Recognition Accuracy

In order to evaluate the sound recognition system performance, the accuracy is an important evaluation parameter. To decide a benchmark recognition accuracy value of the system, the best and intuitional method is to refer the human's hearing test. How much accuracy human hearing system can correctly recognize the ambient environmental sounds is taken as a reference for our automatic recognition system. In previous works similar to our

research, the human listening test experiences have been carried out and their results have been reported [43, 45, 48].

In Chu's work [43], there are 14 different kinds of environmental sounds as test target. Recognition period unit has 2 seconds, 4 seconds and 6 seconds three different length cases. Through experiments, the author concludes that:

- Different period unit length leads to different recognition accuracy. However, the period length becomes less important as it passes a certain threshold. Among the three cases, the accuracy of 2s length case has less accuracy than the 4s and 6s cases. The accuracy of 4s and 6s length cases is close.
- 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 for the 4s length case.
- Under the condition of 4s length case, the Chu's proposed sound recognition algorithm achieves 83% accuracy. This means that sound recognition can reach comparable accuracy as human ear does.

In Eronen's work [45], there are 18 different kinds of environmental sounds are utilized as the test target. Through the experiments, we can learn that:

- The author proposed a "Higher-level" definition among the 18 different sounds and six "Higher-level" classes are classified. For example, the background environmental sounds are classified as "in the public traffic" if the background sounds is car, bus and train, etc. The "Higher-level" average 88% accuracy is higher than the average 66% accuracy of individual sound's result. This means that human hearing system is

sensitive to different "Higher-level" sounds. However, the recognition capability drops when coming to the specific in a similar sound content group.

- Mono and Stereo two sound configurations achieve similar accuracy results. This means that Mono sound configuration is also fitful for recording sounds and later processing.

Through these previous studies, we find that if the recognition targets are environmental sounds, the listening test experiments 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. For one-second detecting length in our research, 82% average recognition accuracy is a challenging performance to achieve.

3.2.2 Power Consideration and Evaluation

Because the WSNs system and its front-end wearable sensor node are power-aware, energy consideration is an essential factor compared with normal sound recognition algorithm study. To achieve both high recognition accuracy and low power consumption of the sound recognition algorithm, a lot of effort is dedicated to solve this problem in our work.

3.2.2.1 Algorithm's Evaluation in Power Consumption Aspect

The performance of an applied algorithm is normally evaluated by the four methods sequentially. They are explicated in the following and concluded in Table 3-1.

Software Algorithm Based

- Algorithm's General Complexity – calculation cost controlled by the loops of key variables in the algorithm.
- Calculation Cost and Approximate Energy Evaluation – multiplication and addition derived calculation cost and their approximate energy expenditure.

Hardware Platform Based

- Calculation Cycles Count – algorithm's execution cycles on hardware platform's clock.
- Direct Energy Measurement – energy measurement upon hardware platform.

"Algorithm's General Complexity" evaluation method indicates the complexity by counting the calculation amount inside program loops. They are decided by some algorithm's key variables. In initial algorithm level study, this method is convenient and quick to approximately understand how complex the algorithm is. Therefore, it is very commonly adopted to evaluate the algorithm's performance in many acoustic applications' *early* research stage [43, 47]. However, this method does not concern the hardware factors and algorithm's implementation.

"Calculation Cost and Approximate Energy Evaluation" method is to calculate the amount of multiplication and addition which is derived from the algorithm [47, 81]. Multiplication and addition of each functional block in the algorithm can be counted. As we known, the algorithm is executed sequentially inside the CPU of different platforms (MCU, DSP, FPGA) with basic addition and multiplication calculation. If we have known basic values of the processor and hardware's information (such as, working voltage and average working current), the approximate *total energy* of the algorithm can be calculated as:

$$\boldsymbol{TotalEnergy = E_{oneMul} \times Num_{Mul} + E_{oneAdd} \times Num_{Add}}, \quad \textbf{(3.2)}$$

where, $\boldsymbol{E_{oneMul}}$ and $\boldsymbol{E_{oneAdd}}$ stand for the energy consumption for one multiplication and addition respectively in CPU. $\boldsymbol{Num_{Mul}}$ and $\boldsymbol{Num_{Add}}$ stand for the number of multiplication and addition in an algorithm.

The advantage of this method is that the algorithm designer can understand each functional block's calculation complexity. It is evaluated by the number of multiplication and addition inside the block. In *middle* term of the research stage, this method acts as a "bridge" connecting the algorithm and hardware design. If having known some basic hardware information in advance (e.g. how much energy a multiplication and addition calculation spends), the designer can roughly calculate how much energy the algorithm will spend upon the hardware platform.

"Calculation Cycles Count" evaluation method utilizes execution cycles in a unit of hardware platform's clock to indicate the algorithm's complexity [57, 79]. This method can give an insight into each functional block's calculation cost. The more cycles have, the more complex this functional block is, and more energy consumes. It is utilized at the *mid-late* research stage based upon the decided hardware platform. If the cycles of certain block are out of budget, the designer can easily return to modify the algorithm and satisfy the requirement.

"Direct Energy Measurement" is the most accurate to measure the energy consumption of an implemented function which originates from the designed algorithm. It is used to verify the energy performance is qualified the target specification or not at the *final* design stage [16, 36, 72, 80, 82]. Our future research target is to implement our proposed environmental sound recognition algorithm on the hardware platform, and satisfy both the accuracy and low-power consumption's requirements.

Table 3-1: Power Consumption Evaluation Methods in Different Design Stages.

	Research Stage	Disadvantages	Advantages
Algorithm's General Complexity	Algorithm (early stage)	• no hardware, energy consumption information	• quick and approximate evaluation
Calculation Cost and Approximate Energy Evaluation	Algorithm (middle stage)	• need knowing some basic hardware platform information in advance	• quick evaluation • energy assigned for functional blocks can be approximately understood
Calculation Cycles Count	Hardware (mid-late stage)	• function block level, not instruction level.	• hardware based • energy of functional blocks is indicated by the number of tested cycles
Direct Energy Measurement	Hardware (final stage)	• difficult in early algorithm stage	• accurate energy evaluation

Through the Table 3-1, we notice that the calculation complexity and energy evaluation of an algorithm is necessary at different design stages. If the design does not satisfy the energy requirements, the designer can return and modify it. During our sound recognition algorithm study, the "Calculation Cost and Approximate Energy Evaluation" method is convenient for us to evaluate the algorithm's approximate energy consumption before entering hardware implementation.

3.2.2.2 Power Consideration in Previous Sound Recognition Algorithms

In normal sound recognition researches [35, 40, 41, 43, 44, 45, 48, 57, 73-75], recognition accuracy is evaluated [35, 40, 41, 44, 45, 48, 73-75] or with a simple discussion of the algorithm's computational cost [43]. Because these researches are not hardware related, the designers seldom consider their algorithms from a hardware angle.

Some of the sound recognition algorithm(s) is executed on enough power supplied DSP / FPGA hardware platform. In Dong's work [57], the author applies MFCC features with HMM classifier to realize environmental sounds recognition for a digital hearing aids application. Based on the DSP platform, architecture optimized DSP block for complex Viterbi classification and accelerator for some specific calculation FFT, Hamming filters, discrete cosine transform (DCT) are utilized. These methods are helpful to accomplish the real-time sounds recognition. Each functional block's cycle count can be known and analyzed, which helps optimize the design and improve the system's performance. Similar methodology is also adopted in Hwang's speech codec work [79].

Researches of sound-based context recognition upon a wearable sensor have been studied in work [36, 53]. In Stager's work [36], the author studied the trade-offs between system's power consumption and accuracy. With only little degradation in recognition accuracy performance, the power consumption is decreased obviously and battery lifetime of the system gets longer. The power evaluation is executed on the sensor node. Similar wearable system is also introduced in Wichern's work [53].

3.2.3 Our Evaluation Approach

One of the evaluation parameter for our sound recognition performance is the system's average accuracy. Referring to some similar researches, 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 our applied recognition algorithm(s) can fulfill the sound recognition or not with limited power assigned for MCU inside the wearable sensor node. Among four methods, the "Calculation Cost and Approximate Energy Evaluation" is adopted in our research. The environmental sound recognition algorithm is decomposed into instruction-level multiplication and addition calculations. Approximate needed energy is summed up as Eq. (3.2) indicated.

The MCU inside our wearable sensor node in Fig. 1.3 is Renesas Technology's H8S/2218 chip [16, 19, 64]. The Chip detail specification has been introduced in Section 2.1.1.3. 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 20MHz (50ns per cycle), 1.8V input voltage with 4mA working current. Main parameters of the MCU and CPU core are summarized in Table 3-2. From the specification [64], it can be calculated:

- For one cycle commands, such as add and subtract operation consumes 4mA×1.8V×50ns = 0.36nJ energy.
- For four cycles command, multiply operation consumes 4mA×1.8V×4×50ns = 1.44nJ energy.

Table 3-2: Main Electronic 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*

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

Therefore, the *total energy* of the algorithm in Eq. (3.2) can be calculated as:

$$TotalEnergy = 1.44nJ \times Num_{Mul} + 0.36nJ \times Num_{Add}, \quad (3.3)$$

where, Num_{Mul} and Num_{Add} stand for the number of multiplication and addition in an algorithm.

We aim that the sound module in the sensor node could continuously work for 24 hours (3,600×24=86,400 seconds), and the CPU core can finish the sound recognition algorithm within each one-second 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.

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

Therefore, based on the 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. 3.2 for performance comparison. If the performance marks of the applied algorithms drop into the top left shaded region of the figure, it can be concluded that the algorithms are suitable for our sound recognition application. Other three tri-angle marks indicate unqualified situations. Our research effort should be as the arrow in the Fig. 3.2 indicates. That means the applied sound recognition algorithm can achieve high accuracy with less power consumption.

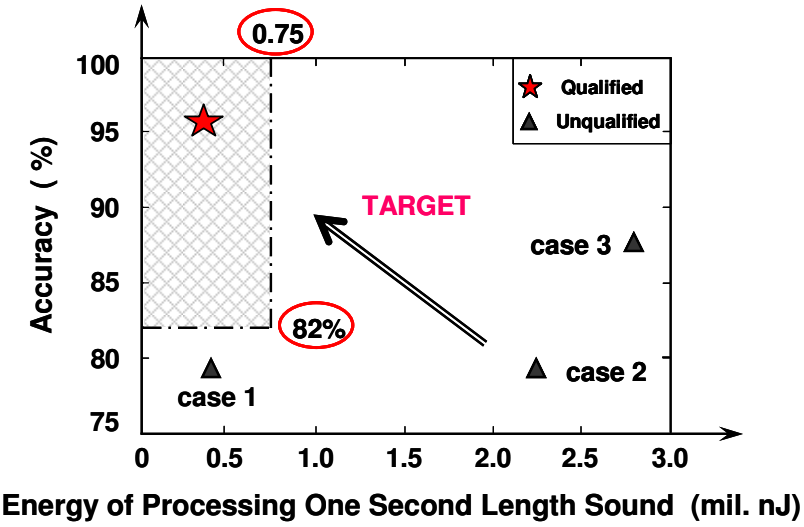


Figure 3.2 Our Sound Recognition Performance Evaluation Approach – Average Accuracy and Power Benchmarks.

3.3 Chapter Summary

In the first part of this chapter, our experimental setup is introduced. Test environmental sounds and data sets for the training and test are described.

In the second part, the evaluation approach for our sound recognition algorithm executed on the power-aware wearable sensor is proposed. Based on the 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 our sound recognition algorithms. The two benchmarks are concluded and indicated in Fig. 3.2. In the following chapters, this approach is taken as a reference to evaluate our proposed environmental sound recognition algorithms.

Chapter 4 Mel-Scale Feature with LBG Classification for Environmental Sound Recognition

In this chapter, sound feature extraction Mel-frequency cepstral coefficients (MFCC) and vector quantization (VQ) classification Linde-Buzo-Gray algorithm (LBG) algorithms are applied for recognizing the background sounds in the human daily activities. Applying these algorithms to 20 typical daily activity sounds, average recognition accuracy of 93.8% can be achieved. In these algorithms, how three parameters (i.e., Mel filters number, frame-to-frame overlap and LBG codebook cluster number) affect system's calculation burden and accuracy is also investigated. By adjusting these three parameters to an optimized combination, the multiplication and addition calculation burden can be reduced by 87.0% and 87.1% individually while maintaining the system's average accuracy rate at 92.5%. Based on the performance evaluation method decided in previous Chapter 3, the comprehensive performance of proposed MFCC+LBG algorithm is evaluated.

4.1 Introduction and Related Work

As acoustic signal processing study, extensive effort has been made to develop systems that can recognize such specialized audio sources as speech, music. To the best of our knowledge, only few frameworks [35, 40-45, 80] have been directed to the daily sounds happening around us. Presenting the spectral envelope characteristic of a sound signal, linear prediction cepstral coefficients (LPCC) [46, 47] is a typical sound feature extraction method. However, it can clearly be concluded that MFCC outperforms the LPCC algorithm in sound recognition experiments from previous work [43, 44]. Recently, matching pursuit (MP) algorithm is introduced to decompose sound's time-frequency feature in Chu's work [43]. During each step, the best decomposed matching atom from a redundant dictionary (such as, Gabor dictionary) is searched and the sound can be presented by linearly combination with

those atoms. Problem of the MP algorithm is that calculation cost for the searching enlarges dramatically with the number of the atoms in dictionary increase. At the same time, after comprehensive comparison of different sound features extraction methods, Cowling's work [40] concludes that MFCC is one of suitable algorithms for non-speech environmental sound recognition. Chen [35], Goldhor [41], and Davis' work [84] also prove that good sound recognition result has been achieved by using Mel cepstrum. These provide the essential motivation for us to apply MFCC as the sound feature.

In addition, power resource of the WSNs system is very limited. Therefore, applied sound recognition algorithm for our front-end wearable sensor should not only achieve high recognition accuracy as traditional sound recognition method, but also with small calculation cost which is evaluated by amount of multiplication and addition calculation in our research. After studying the MFCC algorithm, we find that the number of Mel filters and sound frame-to-frame overlap are two important factors that decide the system's accuracy and calculation amount. In work [35, 83], how Mel filters number affects accuracy has some discussion. However, research so far has rarely been discussed on how frame-to-frame overlap affects the final result. Moreover, with advantage of fewer calculation and higher recognition rate, LBG classifier takes place of the previous dynamic time wrapping (DTW) classifier which is reported in our previous work [76]. Therefore, much effort is on studying how these factors affect and improve the system's accuracy and calculation burden in this chapter.

In order to classify environmental background sound in our everyday life, the Chu's work [43] employs combined MP and MFCC features of 14 sounds which normally happen in our daily environment and achieves 83% average accuracy. However, MP is a computationally complex method. This shortcoming limits it to be applied in our power-aware wearable

sensor node. Another, if the recognition targets are environmental sounds, the listening test experiment executed indicates that people hearing normally can achieve around 82% accuracy. This conclusion provides a benchmark for deciding the accuracy level of similar environmental sound recognition system. Chen's work [35] is close to our research. Seven sounds often happening in bathroom, such as showering and brush tooth sounds, etc., are recognized by utilizing MFCC+HMM and 83.5% average accuracy is achieved. However, as sound template's training process concerned, HMM is a complicated algorithm compared with the LBG. By considering the system's accuracy and calculation burden comprehensively, MFCC+LBG algorithm is decided and applied in our system.

4.2 Sound Recognition Algorithm's Flow

Sound recognition flow is depicted in Fig. 4.1. It has two steps: off-line sound templates generation and on-line sound classification sequentially.

Sound templates characteristics can be extracted off-line and stored in memory in advance. When input test sound comes, by using the same feature extraction method, its feature can be on-line extracted and compared with stored templates to get recognition result. Finally, recognition accuracy is calculated.

The better the sound characteristics can be extracted, the better result can be achieved from the sound classification. In this project, MFCC is applied to extract the sound feature. Based on the extracted templates and input test sounds MFCC matrixes, LBG is applied to do sound classification.

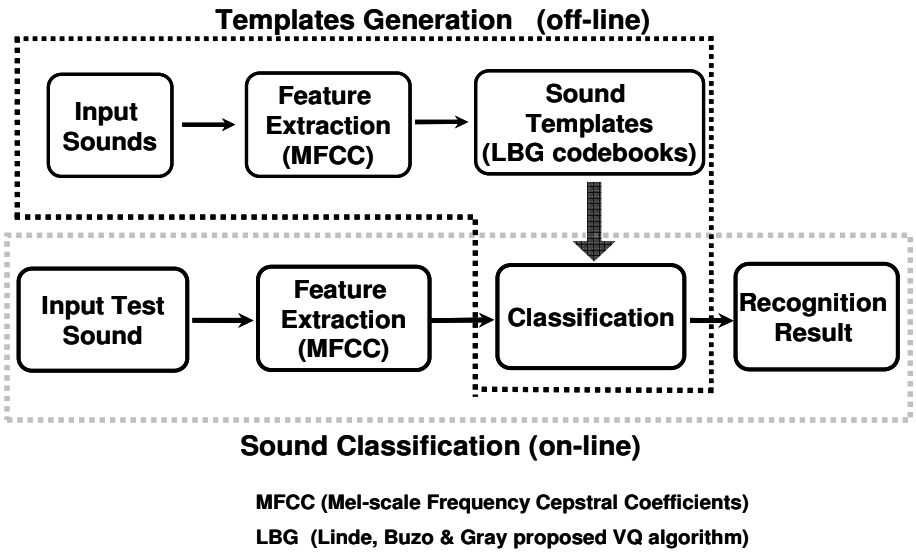


Figure 4.1 Sound Recognition Flow with the MFCC+LBG Algorithm.

4.2.1 Why Mel-Scale?

In order to simplify the system and make the system low cost, our ideal system's sampling rate is less than 16k Hz. Therefore, Nyquist-Shannon Sampling Theory decides what we concern sounds are within [0Hz, 8k Hz]. Fortunately, this range is similar to voice range [0Hz, 4 kHz]. Therefore, we can refer to some achieved voice recognition methods to sounds in our project.

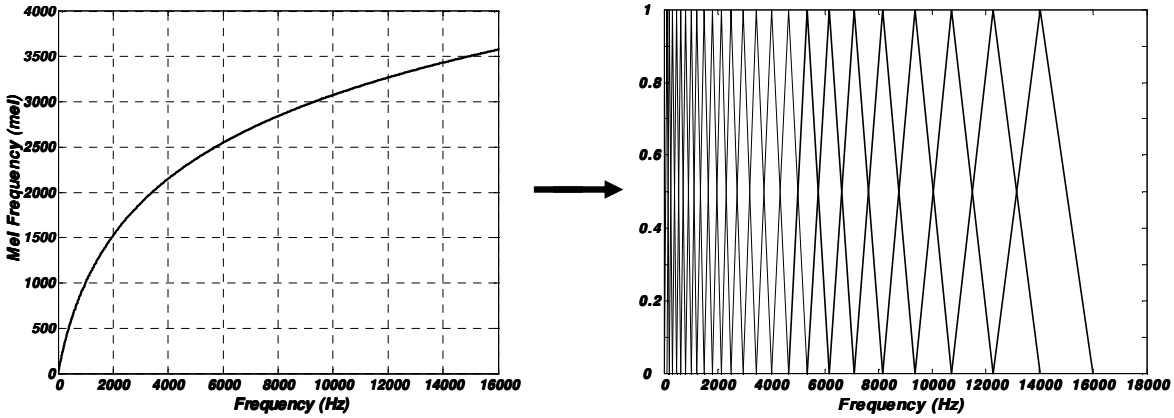
In speech recognition, MFCC is classical algorithm to do feature extraction. How about borrow this algorithm to our sounds recognition project? This motivates us to take a tryout. In fact, the recognition accuracy result shown in Section 4.4.4 by using this algorithm to do sound feature extraction is quite good.

"Mel" comes from word "melody" which is people's feeling based on different pitch values [85, 86]. Scientist found that the pitch comparison scale is not linear with sound

frequency *f*, but linear with called "Mel frequency"- *Mel (f)*. That means sound loudness or pith is linear with "Mel frequency", the higher Mel frequency, the higher pitch or loudness.

Sound frequency *f* and *Mel (f)* relationship is as Eq. (4.1) indicated:

$$Mel(f) = 2595 * \log_{10}(1 + f / 700). \quad \textbf{(4.1)}$$



(a) **(b)**

Figure 4.2 Relationship of the Frequency *f* and Mel Frequency *Mel (f)*.

Based on the Eq. (4.1), a group of triangle band-pass filters in Fig. 4.2 (b) is designed to imitate the curve's function as in Fig. 4.2 (a). We can see that:

- In Fig. 4.2 (a), if the frequency *f* is less than 1 kHz, Mel (*f*) is linear with *f* and logarithmic spacing above 1k Hz.
- Accordingly, in Fig. 4.2 (b), neighbor triangle filters summit distance is same (linear) if the *f* is less than 1k Hz. However, neighbor triangle filters summit distance is increased (no linear) when the *f* is over 1k Hz.

If the horizontal axis in Fig. 4.2 (b) is changed to *Mel (f)*, neighbor triangle filters summit distance will be same, which means sound loudness is linear in *Mel (f)* scale.

4.2.2 Feature Extraction – MFCC Flow

Seven steps in MFCC algorithm are shown in Fig. 4.3:

1: Pre-emphasis

2: Frame blocking

3: Passing Hamming window

4: FFT

5: Mel triangle bandpass filters

6: log

7: DCT (discrete cosine transform)

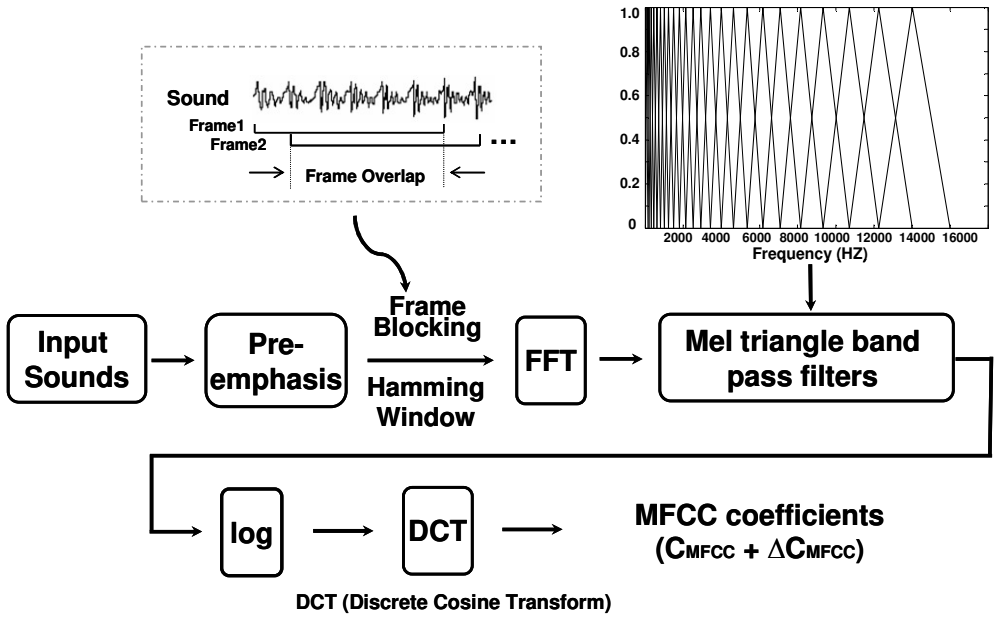


Figure 4.3 MFCC Algorithm Flow.

Some of the steps are described in detail as follows.

Step 1: Pre-emphasis high-pass filter is designed to enhance the higher frequency sound information. Another issue is to do normalization: sound's new value is normalized by using each element to divide the maximum value in the segment.

Step 2: In Fig. 4.3's top left "Frame Overlap", if we can guarantee high recognition accuracy without frame overlap or less frame overlap, the calculation burden can be reduced obviously because of the less calculated frames in certain sound segment.

Step 5: In previous work [76], two experiments prove that Mel-scale triangle filters can extract sound feature well. When a bank of 24 Mel bankpass triangle filters is applied to the system, the 256 points sound frame is condensed into a Mel spectrum matrix ([24×1]) after passing through it.

Step 6, 7: MFCC's coefficients $C_{MFCC}(n)$ is expressed as:

$$C_{MFCC}(n) = \sum_{k=1}^{K} (\log m_k) \cos[n(k - \frac{1}{2}) \frac{\pi}{K}] \qquad n = 1,2,..., L, \tag{4.2}$$

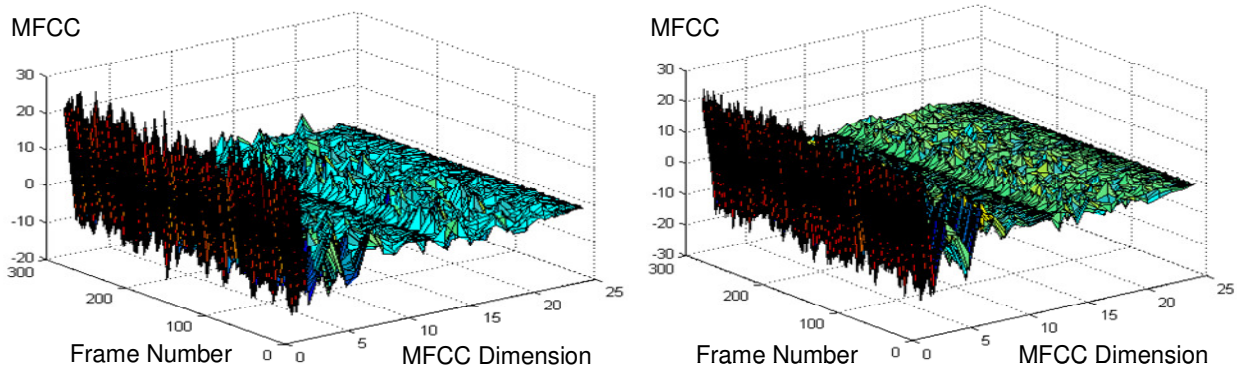
where Mel spectrum is m_k ([24×1]), k=1,2,…K. K is Mel filters number (K =24 in the experiment). L is desired length of cepstrum. Normally, $L < K$ is helpful to compress data (L=12 in the experiment). Therefore, *Step 5*'s achieved [24×1] matrix is further condensed into [12×1] matrix.

Delta values of the [12×1] matrix that means how quickly the $C_{MFCC}(n)$ changes are also used to denote sound characteristics. Therefore, newly built [24×1] matrix combined MFCC and delta MFCC is obtained to feature a frame sound's information. The 0^{th} MFCC coefficient representing the test sound energy is excluded because it is regarded as somewhat unreliable [83].

Through studying the MFCC sound feature algorithm flow, the input sound signal is compressed into a low-dimensional MFCC. It can also be found that two factors play an important role to decide the system's accuracy and calculation cost, they are:

- Frame overlap in "Frame blocking" in *Step 2*
- Mel triangle filters number in *Step 5*

After sound feature extraction, triangle Mel band pass filters can successfully extract sound characteristics which are presented by the MFCC coefficients. Fig. 4.4 is a good example to illustrate its effect. In this figure, 1.5 second sounds of train start and train running's MFCC are calculated. In this example, 24 Mel filters are used, sampling rate is 16k samples per second, 256 points per frame, frame to frame overlap is 128 points. Size of the MFCC coefficients is [297×24] and their values are as Fig. 4.4 shown. We can see the difference between these two sounds' MFCC value, train start's MFCC amplitude is higher, train run's MFCC amplitude is flatter compared with train start's.



(a) Train Start **(b) Train Running**

Figure 4.4 Mel Domain Diagram of Two Sounds - Train Start and Train Running (1.5-second length).

4.2.3 Why MFCC Can Have Less Overlap?

Explanation by an experimental example is shown in Fig. 4.5.

MFCC feature vectors of 1.5 seconds Sound "oven-timer" (item 12 in Section 3.1.1) and Voice of 'Sensor Network' are extracted with overlap=0 and 50%. The Mel vectors of these two conditions are [24×89] and [24×182]. Mel vectors average value of each case is defined as $\overline{MFCC}$. Euclidean distance between each vector and $\overline{MFCC}$ is defined as $\left\| MFCC_j - \overline{MFCC} \right\|$ which indicates how much deviation from its $\overline{MFCC}$. Results of four cases are shown in Fig. 4.5.

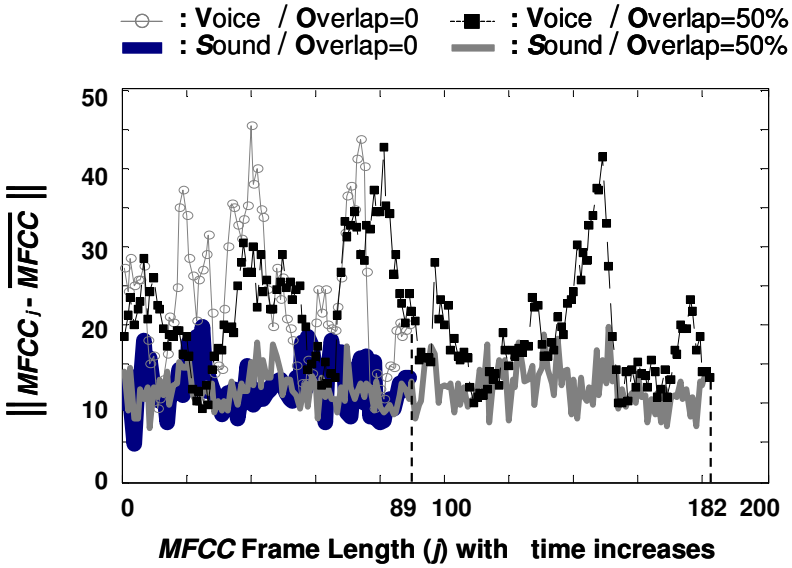


Figure 4.5 An Experimental Waveform That Explains Less Overlap in Sound Process Is Available.

From this figure, it can be clearly noticed that Sound fluctuation is less than Voice's as time increases in two cases of "0 overlap". Additionally, the fluctuations of Sound with 0

overlap and Sound with 50% overlap are nearly same. The reason is that, in speech recognition, frame-to-frame overlap is applied to decrease discontinuity of consecutive frames. However, in sounds recognition, contents in near two frames do not change too much and quickly as the voice frames do. Therefore, frame overlap is not compulsory in sound recognition.

Later part of experimental results verifies that less overlap of neighbor frames is an effective way to reduce calculation cost with little recognition accuracy degradation.

4.2.4 Classification – LBG Algorithm

In 1980, Linde, Buzo, Gray proposed the LBG algorithm which provides a new, quick and simple multi-dimensional integration VQ method. LBG uses an iterative way to generate a codebook and a partition from given training vectors. The codebook can represent original vectors with smallest average distortion. References [87, 88, 89] introduce the algorithm flow in detail.

Suppose $X = \{x_1, \ldots, x_i, \ldots, x_n\}$is input test sound's MFCC feature vectors within certain period, n is number of frames in the segment.

With LBG algorithm, sound j's previously extracted MFCC template vectors can be further condensed into k clusters codebook Y_j (k is in power of 2). Y_j is denoted as $Y_j = \{\lambda_{j1}, \ldots, \lambda_{jq}\ldots, \lambda_{jk}\}$, Different m sound templates can be presented by m codebooks as $Y_1, \ldots, Y_j, \ldots, Y_m$.

Classification process can be described as:

Step 1: calculates minimum distance between X and codebook Y_j for matching.

$$D(X, Y_j) = \sum_{i=1}^{n} e(X_i, Y_j), \quad \textbf{(4.3)}$$

where $e(x_i, Y_j) = \min_{1 \le q \le k} \left\| x_i - \lambda_{jq} \right\|^2$), and $\|.\|$ denotes the *L*2 norm.

Step 2: ***C***lassification ***R***esult (***CR***) is the smallest distortion between test sounds *X* and stored $Y_1, \ldots, Y_j, \ldots, Y_m$.

$$CR = \min_{Y_j (1 \le j \le m)} D(X, Y_j). \quad \textbf{(4.4)}$$

In this research, x_i and all codewords in the codebooks $Y_1, \ldots, Y_j, \ldots, Y_m$ are [24×1] MFCC vectors. Each codebook contains same *k* clusters.

When test sound *X* is input, because *k* clusters ($k<n$) comprises the trained LBG codebook, just $n \times k$ distance calculation are executed when doing test sound and a template matching in *Step 1*. However, with previous work's DTW classifier, only distance matrix between test sound and a template needs $n \times n$ times distance calculation in [76]'s *Step 1*. Therefore, LBG classifier can reduce calculation burden dramatically compared with DTW if $k << n$.

In Eq. (4.3), we notice that the calculation burden increases with codebook cluster's value *k* getting larger. For this reason, under premise of certain high recognition accuracy rate, the less value *k* can lead to less calculation cost. Relation between the multiplication, addition calculation and value *k* will be shown in Fig. 4.7.

4.3 Experimental Process and Consideration of Some Parameters

4.3.1 Experimental Setup and Details

The experimental setup and details has been introduced in previous Section 3.1. In order to make a comparison, another sound recognition method - Haar+HMM algorithm in the following Chapter 5 also adopts the same experimental setup and data sets as Section 3.1 described.

The test target 20 different type of environmental sounds is same as Section 3.1.1 described. In our experiments, some basic person's daily activities are covered. For example, such as inside house activities, household clean, sanitary, dietetic, outside activities, and so on. The background environmental sounds happening with these activities are recorded with the wearable sensor node in Fig. 1.3.

The experimental parameters, process, and experimental data sets for the sound's template training and testing input have also been introduced in the previous Chapter 3. When doing the experiments, we follow these introductions in Section 3.1.2. During the templates' training and recognition process, each unit length of sound is *one* second. This means the algorithm(s) for our sound recognition upon the wearable sensor platform must finish the detection within each second as the Section 2.2.4 discussed.

4.3.2 Recognition Flow

Following three major stages comprise the recognition flow as in Section 3.1.3 introduced.

Stage 1: Training sound templates with LBG clustering

Step1: We have taken 1.0 second length sound as a unit to partition same property sounds from the training data sets. By applying the MFCC algorithm to these training units, sound features vectors for the following LBG training are extracted.

Step2: With the introduced LBG clustering, the Step 1's MFCC vectors are trained and generate a sound codebook.

Step3: Repeat above two steps to the other 19 sounds, we can get different codebooks of those 20 testing sounds. The different codebooks occupy different 'domain' of whole vector space. This provides possibility for test sounds matching.

Stage 2: Sound matching/classification

The matching flow is same as Fig. 3.1.3 shown. When the input test sound comes, it is also segmented as 1.0 second units, and its MFCC feature is extracted. Because the completed training template codebooks of those 20 sounds have been stored in memory, distance between the test MFCC vector and the 20 different sound's codebook is calculated. The closest one among those 20 sound templates is recognized as the most similar to the test sound.

Stage 3: Calculate recognition accuracy rate and algorithm's calculation cost

The final recognition *A*ccuracy *R*ate (*AR*) of our sound recognition system is defined as the Eq. (3.1) in Section 3.1.3. 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.

4.3.3 Consideration of Some Parameter Values

Some parameters decide the system's performance. Before the final evaluation of our system, these parameters must be decided through some experiments.

When the LBG codebook cluster number *k* is 2, 4, 8, 16, 32, detection length is 0.5, 1.0, 1.5, 2.0, 3.0, 5.0 second(s), and with frame-to-frame 128 and 0 overlap respectively (50% and 0% overlap, a frame length=256), the average recognition accuracy are calculated. Their results are indicated in Fig. 4.6. The reason to select these two overlap is that proportion 50% is commonly used in sound and speech signal processing, and 0% is an extreme case through which we can understand how much margin the worst case has.

In Fig. 4.6, the accuracy doesn't differ obviously with respect to each template length cases with the *k* increasing. For example, if the template length is 1.0s, the accuracy of all cases ranges from 90.0% to 96.1%. However, multiplication and addition amount in the classification stage in Fig. 4.7 is doubled as *k* increases. These mean, if a proper *k* is chosen, satisfying recognition accuracy (higher than 82% decided in Section 3.2.1) can be achieved with small calculation cost. Finally, the appropriate cluster number *k* is decided as 4 after thoroughly studying the Fig. 4.6.

Among Fig. 4.6 B's cases, when the k =4 and template length is 1.0s, the recognition accuracy is 92.5%. The result is very close to the accuracy 94.3% of 1.5s case and better than 85.6% of 0.5s case. Compared with the 1.0s template length case, the accuracy of 2.0s, 3.0s and 5.0s three cases have no obvious improvement. Comprehensively considering these facts, 1.0s is taken as the template unit. This experiment also proves that the constraint - "Length of Processing Unit is One Second" in Section 2.2.4 is reasonable.

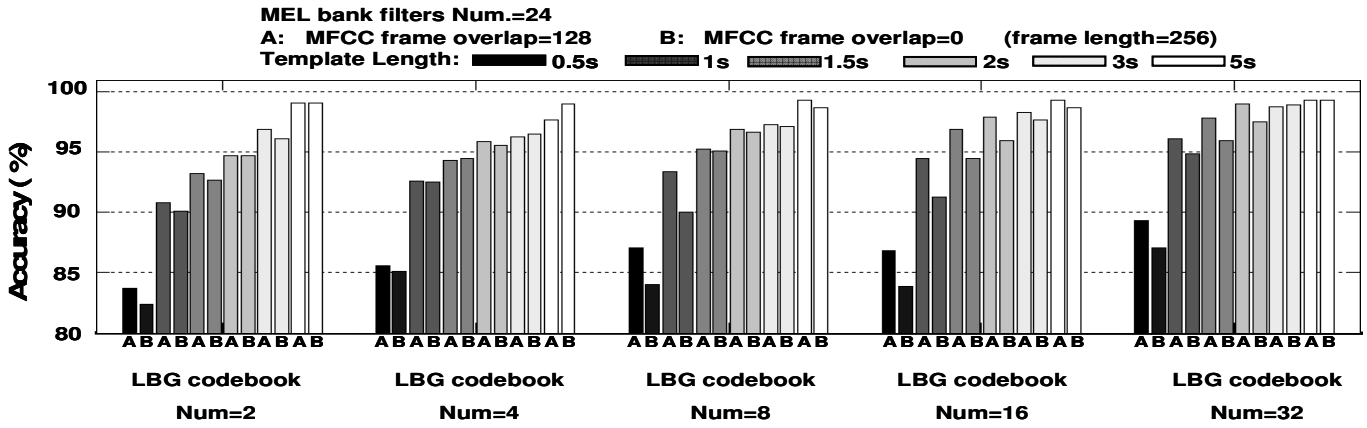


Figure 4.6 Average Recognition Accuracy as a Function of the Template Length and LBG Codebook Cluster Number *k*.

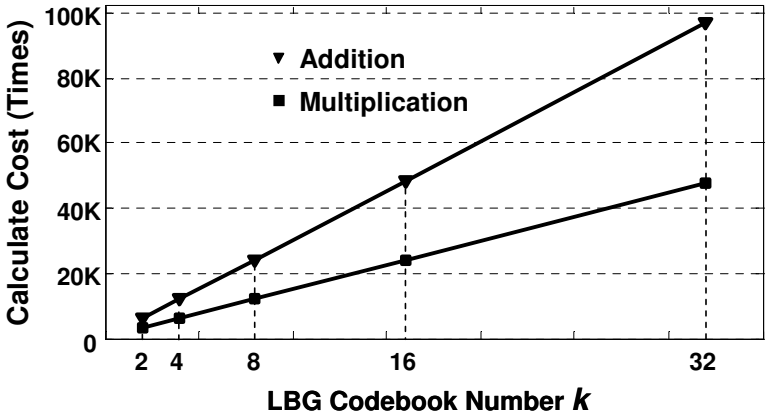


Figure 4.7 Multiplication and Addition Calculation Cost as a Function of the LBG Codebook Cluster Number *k*.

4.4 Experimental Results and Discussion

For sound recognition algorithm applied for the power-aware WSNs system, we expect the calculation cost could be decreased much on premise of certain high accuracy rate. Therefore, it is necessary to do trade-off between the system's recognition accuracy and

calculation cost. As the analyses of MFCC algorithm flow in Section 4.2.2, the Mel filters number and frame overlap are two parameters that play important roles to determine the system's performance.

Most research sets the frame overlap as a default value (e.g. 50% overlap), and has not too much discussion about how this variance affects the system performance. However, through our study, we find that tuning appropriate variance of the frame overlap and proper cooperation with other parameters are effective methods to achieve low calculation cost with little accuracy deterioration.

4.4.1 Recognition Accuracy (Mel-filter Number, Frame Overlap)

The Mel-filter Number is set as seven values 12, 14, 16, 18, 20, 22, 24, frame overlap is set as eight values 0, 32, 64, 96, 128, 160, 192, 224, and every frame's length is 256 points. The template length is as previously decided 1.0 second. Among these 56 different combinations of Mel-filter Number and Frame Overlap, every combination's average recognition accuracy of the test 20 sounds is illustrated in Fig. 4.8 with the proposed MFCC+LBG algorithm.

From the figure, some trends can be concluded:

- More the filters, higher the system's accuracy becomes.
- As frame overlap increases, accuracy becomes higher.

Two extreme cases in the figure - case A (filter number=24, frame overlap=224) with highest accuracy rate 93.8% and case C (filter number=12, frame overlap=0) with accuracy rate 88.8% are taken as references.

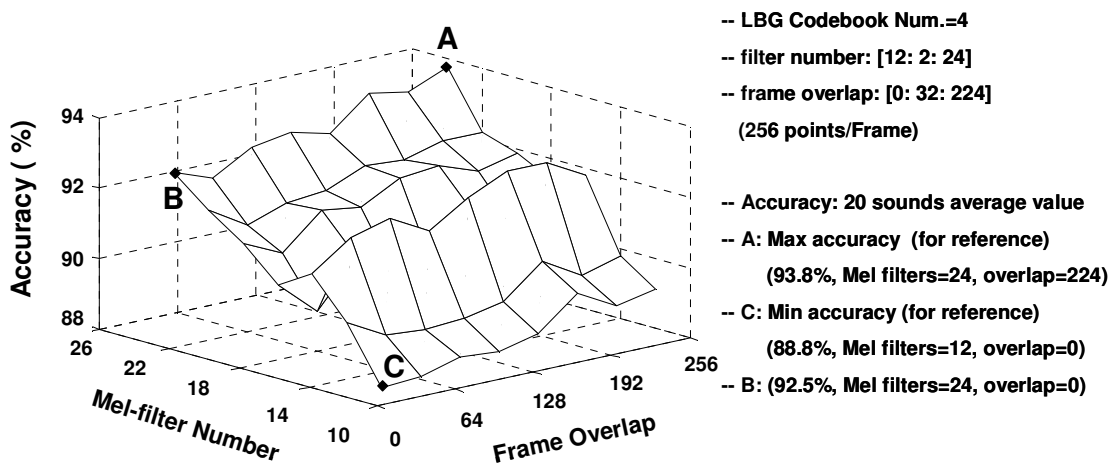
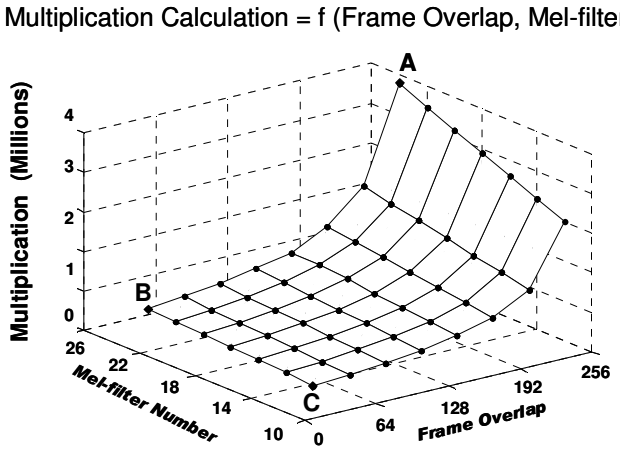


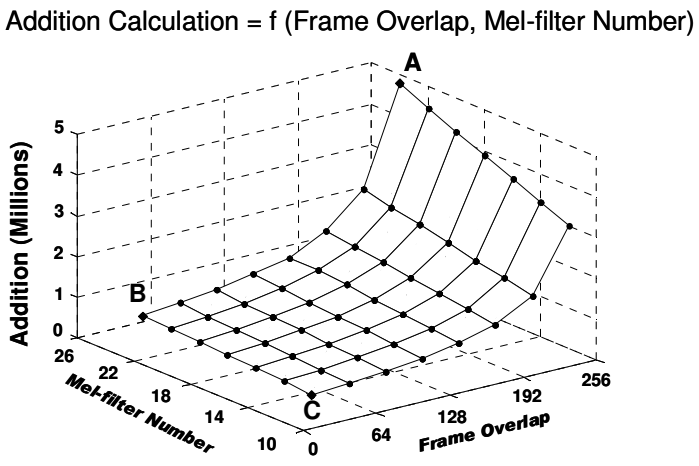
Figure 4.8 Accuracy Rate in Function of Mel-filter Number and Frame Overlap.

4.4.2 Calculation Cost (Mel-filter Number, Frame Overlap)

Multiplication and addition calculation of the algorithm versus the Mel-filter Number and the Frame Overlap are shown in Fig. 4.9 (a) and 4.9 (b) individually. We can see that:



(a)



(b)

Figure 4.9 Calculation Cost of Multiplication and Addition in Function of the Mel-filter Number and Frame Overlap.

- Frame Overlap is a main parameter to affect the amount of multiplication and addition calculation. With its increase, both of the calculations increase quickly.
- In contrast, Mel-filter Number has little effect on the calculation cost. It is because Mel filtering is just one step among seven steps of the MFCC flow, and does not add any calculation burden to the latter LBG classifier.

Therefore, aiming to achieve high accuracy with low calculation cost, increasing the Mel-filters and decreasing the neighbor frame to frame's overlap is a good applicable method. Under this consideration, case B in Fig. 4.8 and Fig. 4.9 satisfies these requirements and taken as a candidate case in the following experiments.

4.4.3 Experimental Results

With reference cases A, C in Fig. 4.8 and Fig. 4.9, case B's average accuracy, multiplication and addition calculation with same MFCC+LBG algorithm are listed in Table 4-1, bar graph is illustrated in Fig. 4.10. Twenty sounds recognition confusion matrix result (only case B) is contained in Table 4-2.

From those figures and tables, we can see that case B sacrifices 1.3% of accuracy compared with the highest accuracy case A (93.8%-92.5%). However, this help decrease 87.0% of multiplication and 87.1% addition calculation compared with the reference case A. Case C displays best performance in terms of computational complexity, followed by case B. However, the recognition accuracy of some test items is not satisfying as that of point B in Table 4-2. Therefore, comprehensively trading off accuracy and calculation cost, case B's condition (template sound length =1.0s, Mel filters number=24, frame overlap=0) has the best performance among those 56 combinational cases. 92.5% high recognition accuracy rate with less calculation cost can be achieved.

Table 4-1: Performance Comparison of Optimized Case B, Reference Cases A and C with the Same MFCC+LBG Algorithm.

	Average Accuracy (%)	Multiplication (Millions)	Addition (Millions)
A (24 Mel filters/224 overlap)	93.8% (max accuracy)	3.544	4.779
B (24 Mel filters/0 overlap)	92.5%	0.459 (A's 13.0%)	0.615(A's 12.9%)
C (12 Mel filters/0 overlap)	88.8%	0.294 (A's 8.3%)	0.391(A's 8.2%)

Table 4-2: Twenty Sounds Recognition Accuracy Confusion Matrix of Optimized Case B, Reference Cases A and C.

		20 Test Sounds' Recognition Confuse-Matrix for Point B																				LBG B(=4)	LBG A	LBG C	GMM (=4)	DTW
	Len. (S)	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Cu*	Au*	Acc.	Acc.	Acc.	Acc.	Acc.
A1	48.3	47	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	47	48	**97.9%**	97.9%	93.8%	100%	97.9%
A2	170.5	0	127	0	0	0	0	0	0	0	4	13	0	1	21	3	0	1	0	0	0	127	170	**74.7%**	78.2%	64.1%	94.1%	22.4%
A3	141.4	0	0	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	141	141	**100%**	100%	100%	100%	100%
A4	26.1	0	0	0	22	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	22	26	**84.6%**	92.3%	80.8%	96.2%	69.2%
A5	71.3	1	0	0	0	60	0	0	0	0	0	0	0	0	1	0	0	0	9	0	0	60	71	**84.5%**	87.3%	93.0%	88.7%	85.9%
A6	105.7	0	0	0	0	0	104	0	1	0	0	0	0	0	0		0	0	0	0	0	104	105	**99.1%**	100%	82.9%	98.1%	98.1%
A7	25.0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	25	25	**100%**	100%	100%	100%	100%
A8	14.9	0	0	3	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	11	14	**78.6%**	71.4%	78.6%	100%	57.1%
A9	17.5	0	0	0	0	0	0	0	0	17	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	62	1	0	0	3	0	0	0	0	0	0	62	66	**93.9%**	87.9%	90.9%	100%	74.2%
A11	20.4	0	0	0	0	0	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0	19	20	**95.0%**	95.0%	95.0%	40.0%	75.0%
A12	51.5	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	0	0	51	51	**100%**	100%	100%	100%	100%
A13	35.6	0	0	0	0	0	0	0	0	0	4	3	0	27	0	1	0	0	0	0	0	27	35	**77.1%**	85.7%	71.4%	88.6%	88.6%
A14	21.9	0	0	0	0	0	0	0	0	0	0	0	0	1	20	0	0	0	0	0	0	20	21	**95.2%**	100%	76.2%	100%	81.0%
A15	36.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	36	36	**100%**	100%	97.2%	77.8%	88.9%
A16	22.8	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	18	1	0	0	0	18	22	**81.8%**	90.9%	59.1%	68.2%	77.3%
A17	40.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	37	0	0	0	37	40	**92.5%**	92.5%	92.5%	100%	90.0%
A18	142.3	5	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	135	0	0	135	142	**95.1%**	96.5%	100%	98.6%	89.4%
A19	24.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	24	24	**100%**	100%	100%	100%	100%
A20	21.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	21	21	**100%**	100%	100%	100%	100%
		Average Accuracy																						**92.5%**	**93.8%**	**88.8%**	**92.5%**	**84.8%**

***Cu*: C**orrectly* recognized ***u**nits* **Au*: *A**ll* input sound ***u**nits*

***** :** for space limitation, above Confuse-Matrix is for Point B, LBG A/LBG C/GMM/ DTW just has accuracy.

A1: Vacuum cleaner (house cleaning)
A2: Washing machine (wash something)
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 outside
A15: Run
A16: Train start (train accelerates, in train)
A17: Train run (train normally runs, in train)
A18: Take umbrella in the rain
A19: Mechanical alarm
A20: Telephone ring (telephone comes)

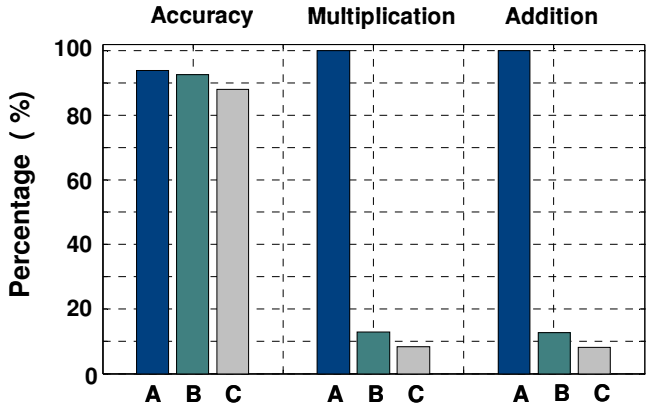


Figure 4.10 Comparison of Optimized B, Referenced A and C's Accuracy and Calculation Cost.

4.4.4 Performance Comparison and Whole System's Evaluation

In this research, the MFCC+LBG algorithm achieves better result than the MFCC+DTW [76] and MFCC+GMM (Gaussian mixture model) [90] algorithms as recognition accuracy and calculation burden concerned. Performance of these three algorithms with same experimental parameters (Section 4.4.3 point B's case) is indicated in Fig. 4.11, Table 4-2 and Table 4-3.

In order to have a fair performance comparison, the number of LBG codebook cluster and GMM mixture-components' number are all set as 4. One division and logarithm calculation in the GMM classifier are assumed to take a single multiplication calculation.

Under same circumstance, 20 sounds average recognition rate of MFCC+LBG and MFCC+GMM can be up to 92.5% which is higher than that of MFCC+DTW 84.8%. In addition, the most attractive is that MFCC+LBG can reduce multiplication by 79.0% and addition calculation by 85.7% compared with the reference MFCC+DTW. As the amount of multiplication calculation shown in Fig. 4.11, MFCC+LBG is also less than MFCC+GMM and need not calculate complicated logarithm and division as GMM does.

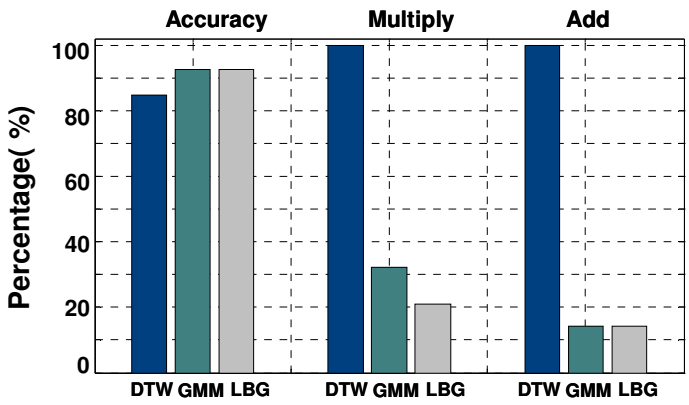


Figure 4.11 Accuracy Comparison of the MFCC+LBG, MFCC+DTW, and MFCC+GMM Algorithms.

Table 4-3: Performance Comparison of the MFCC+DTW, MFCC+GMM, and MFCC+LBG Algorithms.

	Average Accuracy	Multiply (mil.)	Add (mil.)	Energy * (mil. nJ)
MFCC+DTW (24 Mel filters/0 overlap)	84.8%	2.184	4.293	5.721
MFCC+GMM (24 Mel filters/0 overlap)	92.5%	0.705 (DTW's 32.3%)	0.610 (DTW's 14.2%)	1.235
MFCC+LBG (24 Mel filters/0 overlap)	92.5%	0.459 (DTW's 21.0%)	0.616 (DTW's 14.3%)	0.883

* The whole Energy = 1.44nJ × Mul. + 0.36nJ × Add (mil. nJ) based on Section 3's discussion.

Based on the previous discussion, the whole system's performance comparison is illustrated in Fig. 4.12. Three sound recognition algorithms recognition accuracy are all over the accuracy benchmark 82% decided in previous Section 3.2.1. Among them, the proposed MFCC+LBG algorithm can achieve 92.5% accuracy. In order to decrease the algorithm's calculation cost and satisfy the system's power consumption requirement, the "Mel-filter Number" and "Frame Overlap" two parameters in the MFCC feature extraction have been discussed. The optimization of these two parameters decreases the algorithm's calculation cost greatly. If the algorithm executed on our wearable sensor node introduced in Chapter 2, as Table 4-3 indicated 0.883mil. nJ/s energy is need. However, the proposed MFCC+LBG algorithm with lest calculation cost does not satisfy the power consumption's requirement still.

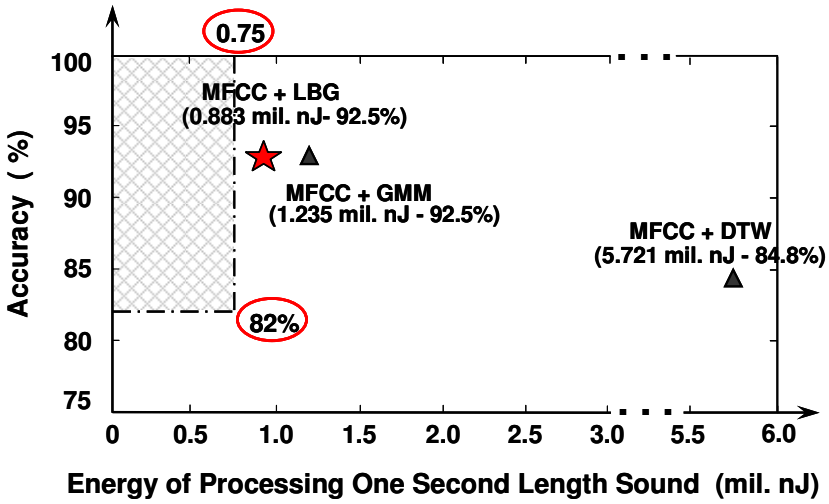


Figure 4.12 Performance Comparisons of MFCC+LBG, MFCC+DTW, and MFCC+GMM Algorithms.

4.5 Chapter Summary

In this chapter, MFCC+LBG algorithm is applied to recognize the background environmental sound. Twenty typical daily activities sounds are recognized. As the system's two important factors – recognition accuracy and algorithm's calculation cost, we have a thorough study. Finally, the approximate power consumption of executing the algorithm upon our wearable sensor node has been evaluated.

Compared with our previous study by utilizing MFCC+DTW method, the optimized MFCC+LBG sound recognition algorithm improves the recognition accuracy to 92.5%, obvious improvement than 84.8% with the MFCC+DTW. If the algorithm executed on our wearable sensor node, the power consumption is much less than the MFCC+DTW method. However, the MFCC+LBG's power consumption (0.883mil. nJ/s) does not qualify the benchmark requirement (0.75mil. nJ/s) through the power consumption evaluation.

Therefore, there is a need to have a new method to satisfy the system's power consumption requirement while maintaining certain level recognition accuracy. In the following Chapter 5, a novel sound Haar-like feature with high performance HMM classifier is proposed to solve these problems.

Chapter 5 Low-Complex Haar-Like Feature with HMM Classification for Environmental Sound Recognition

In this chapter, a low cost sound feature extraction Haar-like filtering with hidden Markov model (HMM) classification algorithm is newly proposed and applied to recognize the environmental sounds. Average recognition accuracy 96.3% of 20 typical daily activity sounds by the proposed algorithm can be achieved, which outperforms normal personal hearing capacity 82% accuracy. At the same time, it also satisfies the amount of calculation cost decided by the wearable sensor node's energy resource. Through experimental comparison, the proposed method outperforms other normally utilized sound recognition algorithms as the recognition accuracy and calculation cost two evaluation parameters concerned.

5.1 Introduction and Related Work

Some environmental sound recognition researches have been reported previously [35, 38, 40, 41, 43, 44, 48, 57].

At the feature extraction stage, conventional state-of-the-art Mel-frequency cepstral coefficients (MFCCs) filtering is used to extract the sound feature and obtain good recognition accuracy [35, 41, 48, 57]. 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. Recently, in study [43], 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. At the classification stage, performance of the Gaussian mixture model (GMM), support

vector machine (SVM), Linde-Buzo-Gray algorithm (LBG), *k*-means, and HMM classifiers has been studied and compared in work [40]. Through the work, we have learned that the HMM [48, 81] classifier can achieve high recognition accuracy with an acceptable increment of calculation cost compared with other classifiers.

Energy efficiency plays an important role for mobile and wearable devices in the WSNs system [61]. 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 [57, 62] platforms may not perform well on our wearable sensor node. In work [57], 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 [38], 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. In work [35], 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. Therefore, a major challenge for our research is development of a new sound recognition algorithm for achieving high accuracy with low calculation cost to meet the energy requirement.

Among the similar researches, Nishimura's studies [82, 93, 104] are the most close to this research. However, there are some fundamental difference between his work and this research as Table 5-1 shown. One is the research target is different, speech/non-speech detection is one of Nishimur's main focus [82, 104]. In this research, what is the specific environmental background sound is our main focus. Another, in this research, approximate power consumption evaluation upon the power-aware wearable sensor platform is comprehensively discussed and studied; however, there is not too much consideration and power consumption evaluation on a certain hardware platform in Nishimura's work [104]. Specifically speaking, to calculate the total number of multiplication and addition of the sound algorithm separately is adopted in this work; however, the total number of multiplication and addition is counted as the applied algorithm's computational complexity in Nishimura's studies [82, 93]. In fact, the power used for a multiplication is about as 4 times as an addition operation needs [58, 64]. Thirdly, because the difference of an environmental sounds with time variation can be well modeled by a specific statistical HMM model; therefore, the HMM classifier is used in this research. The performance comparison of the Haar+HMM and Haar+LBG algorithms is illustrated in later Fig. 5.9 of Section 5.4.4.

Table 5-1: Comparison of Nishimura's Studies (Ref. [82, 93, 104]) **and This Work.**

	Nishimura's Studies [82, 93,104]	**This Research**
Research Target	**Speech/Non-speech**	**Sound recognition**
Hardware Consideration	■ **Does not concern too much about hardware platform.** ■ **Does not give comprehensive power consumption evaluation on hardware platform.**	■ **Approximate power consumption evaluation consider hardware sensor platform.**
Power Consumption Evaluation	■ **Total number of multiplication and addition → calculation cost** [93]. ■ **Booth's multiplication** [104].	■ **Power consumption based on hardware sensor platform.** ■ $E_{Multiplication} = 4\ E_{Addition}$ [64, 58]
Classifier	■ **LBG** [93] ■ **AdaBoost** [104]	■ **HMM**

In this chapter, a novel Haar+HMM algorithm is proposed to recognize the environmental background sounds. Haar-like filtering is a commonly used feature extraction method for two-dimensional (2-D) image processing fields. This method was first used in 2-D face detection and yielded good performance [92]; it was also applied to speech and non-speech detection [93]. 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 [95] 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.

5.2 Implementation of Sound Recognition by the Haar+HMM Algorithm

The proposed sound recognition flow is shown in Fig. 5.1. It follows two sequential steps: generation of off-line sound's classifier training 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's training classifier 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 [48, 81, 100].

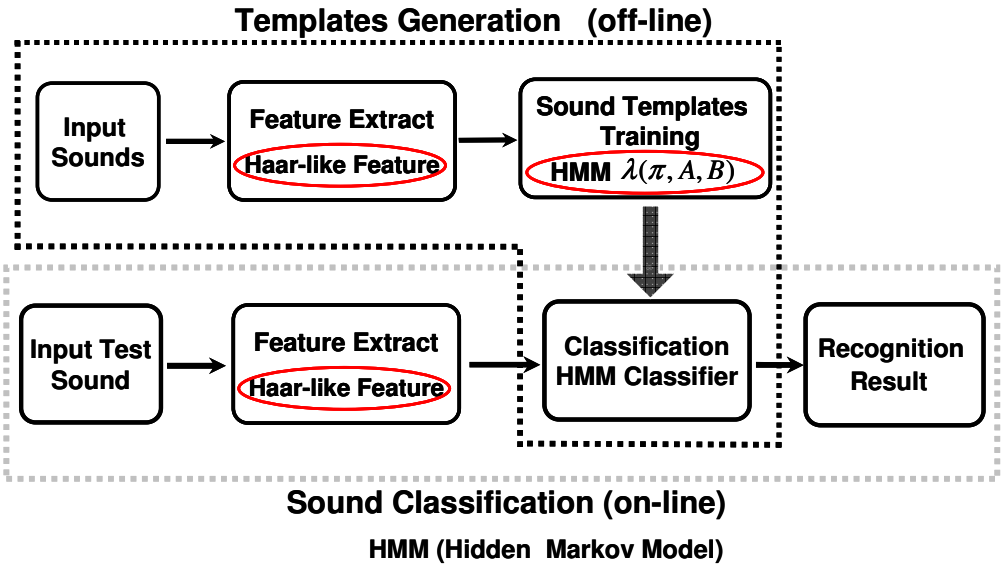


Figure 5.1 Sound Recognition Flow with the Haar+HMM Algorithm.

5.2.1 Why Employ Haar-like Sound Feature with HMM Classification?

There are many sound features which are commonly used in the feature extraction stage,

such as previously introduced linear prediction cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), and matching pursuit (MP), etc. Haar-like sound feature is one of them. Inspired by the low cost and efficient feature extraction of Haar-like filtering used in 2-D image signal processing, this filtering method is also applied to some 1-D signal, for example speech/non-speech detection [82, 93, 104], acceleration processing [94]. This idea was firstly proposed and achieved satisfying result in Nishimura's work in year 2008 [93]. Similar to the formerly introduced Mel-filters group in MFCC algorithm's sound feature flow, Haar-like filters group with different scaling patterns is used to extract the sound feature. An appropriate filters group for specific sound can be achieved after the training stage (the training process is specified in Section 5.2.3). Therefore, the sound's unique characteristics can be extracted and get more extinguished after the specific Haar-like filters group filtering. Another, the Haar-like filter possesses simple structure with extremely low computational cost advantage. This is helpful to realize less power consumption when executing the sound feature extraction inside the sensor node's MCU.

Most of the environmental sounds are "quasi-stationary" which has been introduced in Stager's work [37]. That means there are certain difference and regular stationary part in the sound's spectrum with time variation. Those different statuses and the status transition between each other in time-series sound can be modeled by a statistical HMM. After training the extracted Haar-like features, each sound has its own HMM classifier. The classification result is decided by selecting the class with the largest posteriori probability. In references [107, 108], applicability of application the statistical HMM to environmental sound processing and recognition has been discussed and proved.

Employing the HMM as the classifier to recognize environmental sounds has been

developed in researches [84, 105, 106, 107, 108]. However, the sound feature in these researches is traditional MFCC [105], LPCC [107], combinational time and frequency domain features [84, 108]. The computational complexity of these sound features is high. This inevitably increases the power consumption when executing them inside the sensor's MCU. Therefore, in this chapter, a low computational cost Haar-like sound feature with high performance HMM classification is used for the environmental sound recognition.

5.2.2 Haar-like Sound Feature Extraction

A: 1D Haar-like filtering

A basic Haar-like filter $h_{filter}(j)$ is denoted by Eq. (5.1) and shown in Fig. 5.2.

$$h_{filter}(j) = \begin{cases} -1 & -W_{filter}/2 < j \le 0 \\ +1 & 0 < j \le W_{filter}/2 \end{cases}, \tag{5.1}$$

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

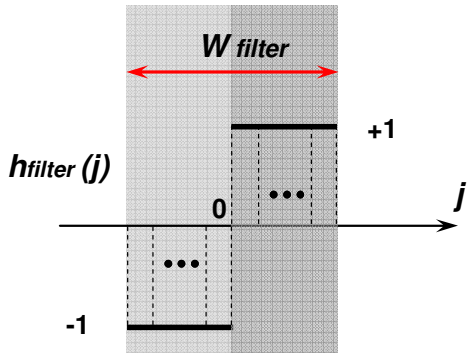


Figure 5.2 One-Dimension (1-D) Haar-like Filter $h_{filter}(j)$.

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. 5.3, 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. 5.3. The Haar-like feature x_m is calculated by the sum of the absolute outputs of Haar-like filtered signals:

$$x_m = \sum_{n=0}^{N-1} \left| \sum_{k=1}^{W_{filter}} h_m(k) * s(n.W_{shift} + k) \right|, \qquad \textbf{(5.2-a)}$$

$$= \sum_{n=0}^{N-1} oneFilterValue(n), \qquad \textbf{(5.2-b)}$$

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:

$$N = (W_{frame} - W_{filter}) / W_{shift} + 1. \qquad \textbf{(5.3)}$$

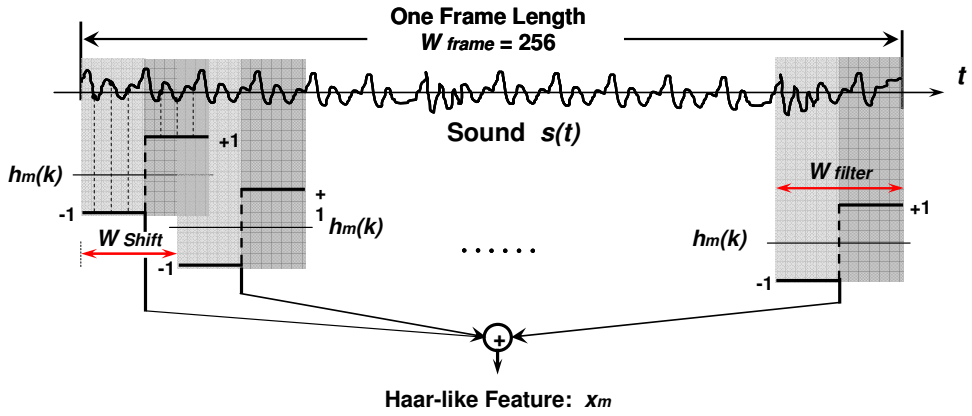


Figure 5.3 One-Dimension (1-D) Haar-like Filtering for One Frame's Sound Signal.

Parameter W_{shift} is adjustable as α change (α is defined in Eq. (5.4)). 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 α=0, W_{shift} is set to 1.

$$\alpha = W_{shift} / W_{filter} \quad \textbf{(5.4)}$$

B: Integral Signal (*IS*)

From Eq. (5.1) and Fig. 5.2, it follows that the coefficients of the Haar-like filter are -1 when $j \leq 0$, 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 [92], a novel concept called *I*ntegral *S*ignal [95] is newly utilized in this work. The *I*ntegral *S*ignal of each sound frame has been calculated and stored in memory as a preprocessed intermediate signal for later use. It is defined as follows:

$$IS(n) = \sum_{t \leq n} s(t). \quad \textbf{(5.5)}$$

Therefore, the filtered sound signal calculation can be denoted as:

$$oneFilterValue = IS(t + W_{filter}) - 2 * IS(t + W_{fitler} / 2) + IS(t). \quad \textbf{(5.6)}$$

In Eq. (5.2-a), 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. (5.6), the calculations are reduced to one multiplication and two addition calculations. Therefore, it is obvious that the computational complexity of x_m in Eq. (5.2-b) decreases. At

the same time, the accuracy does not deteriorate.

C: 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 Section 5.2.2_A 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:

$$M = \begin{pmatrix} HaarFilNum \\ HaarWidMax/2 \end{pmatrix}. \quad \textbf{(5.7)}$$

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:

$$X_v = \{x_{v1}, x_{v2},\ldots, x_{vi},\ldots, x_{vn}\}, \quad \textbf{(5.8)}$$

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.

5.2.3 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 [93]. 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:

- 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. (5.8) defines (n=*H*aarFilNum).
- 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. (5.7) 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 5-2 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 [87] is employed to develop new cluster centers in work [93]. In this research, k-means cluster [96, 97] 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, 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.

Table 5-2: Training Haar-like Filters Pool Size with Relation to the Two Parameters - "*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]					

* Below grey columns are impossible cases. Middle white columns are inexecutable cases because the *M* value is less than our target 20 testing sounds. Top grey columns are our experimental cases.

5.2.4 HMM Classification

As shown in Fig. 5.1 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 HMM 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 [47, 81, 98].

The block diagram of an on-line test sound HMM classifier is shown in Fig. 5.4. 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 [46, 47], the result with the highest likelihood among all the templates is recognized as the most similar to the test sound.

$$l^* = \arg\max_{1 \le l \le p}[P(O_l \mid \lambda^l)]. \qquad \textbf{(5.9)}$$

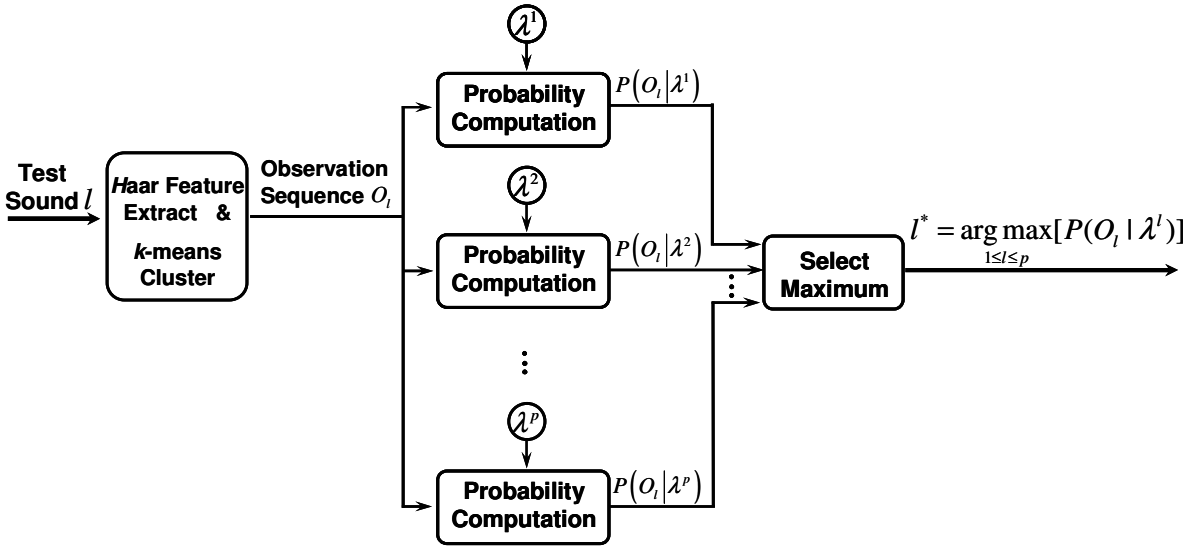


Figure 5.4 Block Diagram of a Test Sound's HMM Classification.

After analyzing Eq. (5.9), 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 [47, 81].

5.3 Experimental Process and Consideration of Some Parameters

5.3.1 Experimental Setup and Details

The experimental setup and details has been introduced in previous Section 3.1. In previous Chapter 4, the MFCC+LBG algorithm for environmental sound recognition also adopts the same experimental setup and data sets.

The test target 20 different type of environmental sounds is same as Section 3.1.1 described. In our experiments, some basic person's daily activities are covered. For example, such as inside house activities, household clean, sanitary, dietetic, outside activities, and so on. The background environmental sounds happening with these activities are recorded with the wearable sensor node introduced in Fig. 1.3.

The experimental parameters, process, and experimental data sets for the sound's template training and testing input have also been introduced in the previous Chapter 3. When doing the experiments, we follow these introductions in Section 3.1.2. During the templates' training and recognition process, each unit length of sound is *one* second. This means the algorithm(s) for our sound recognition upon the wearable sensor platform must finish the detection within each second as the Section 2.2.4 discussed.

5.3.2 Recognition Flow

Following three major stages comprise the recognition flow as in Section 3.1.3 introduced.

Stage 1: Training and getting the HMM classifier $\lambda^{v}(\pi, A, B)$ for individual sound $s_v(t)$

Step1: We have taken 1.0 second length sound as a unit to partition same property sounds from the training data sets. After Haar-like filtering these training units (Section 5.2.3 introduces how to select a suitable Haar-like filters group), pool of sound features vectors for the following *k*-means clustering are prepared.

Step2: After obtaining the updated centroids a sound $s_v(t)$ by *k*-means clustering, an observation sequence of the sound can be formed and denoted as O_v.

Step3: With the composed training sound's observation sequence O_v and initial HMM parameter $\lambda^{v}(\pi, A, B)_0$, the model $\lambda^{v}(\pi, A, B)$ can be refined and completed with the Baum-Welch algorithm's training.

Step4: Repeat above two steps to the other 19 sounds, we can get different HMM classifier $\lambda^{v}(\pi, A, B)$ of those 20 testing sounds.

Stage 2: Sound matching/classification

The matching flow is same as Fig. 3.1 shown. When the input test sound comes, it is also segmented as 1.0 second units, and its Haar-like feature is extracted. Because the completed training HMM classifiers $\lambda^{v}(\pi, A, B)$ of those 20 sounds have been stored in memory, the probabilities of the test sound's observation sequence O_l with those 20 different sound's classifiers are calculated by employing the *Viterbi* algorithm individually. The smallest one is recognized as the most similar to the test sound.

Stage 3: Calculate recognition accuracy rate and algorithm's calculation cost

The final recognition *A*ccuracy *R*ate (*AR*) of our sound recognition system is defined as the Eq. (3.1) in Section 3.1.3 introduced. 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.

5.4 Experimental Results and Discussion

As analyzed in previous Chapter 3, 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.

5.4.1 Parameters Tuning and Recognition Accuracy Rate

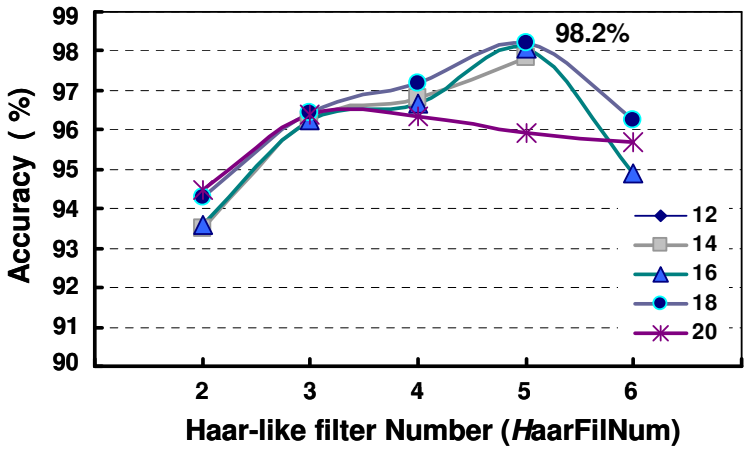


Figure 5.5 Average Accuracy in Function of the Parameters: HaarFilNum and HaarWidMax.

Figure 5.5 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, α=0 (W_{shift}=1), number of HMM states=7, number of HMM observe symbol=15, and ε=0.01, the average accuracy of the 20 sounds can reach highest at 98.2%. Even with *H*aarFilNum =2 (other parameters are identical), it can yield accuracy of more than 93.0%. These results greatly outperform the required minimum accuracy of 82% decided in Chapter 3, and also prove that our proposed Haar+HMM environmental sound recognition algorithm with the proposed training method is effective.

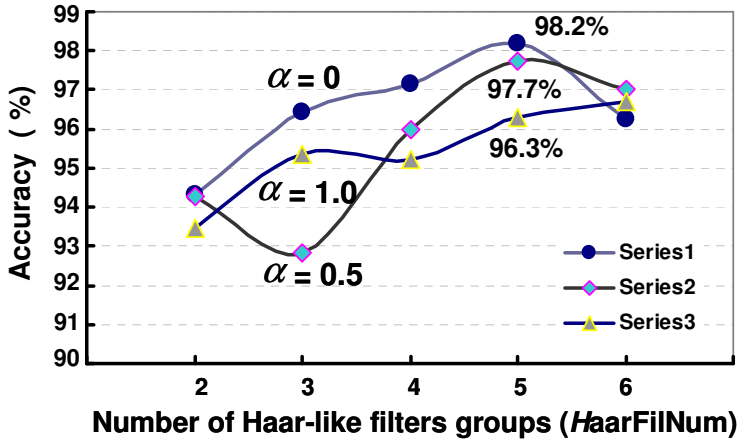


Figure 5.6 Average Accuracy in Function of the Parameter: *H*aarFilNum and α.

Besides α=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. 5.6. Except for the value of α, the parameters are set as in the previous experiment with a maximum accuracy 98.2%. From this figure, it can be observed 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 1.9%. So the influence of the value of α on accuracy is not significant if the appropriate *H*aarFilNum is chosen.

5.4.2 Comparison of Different Sound Features' Performance

Different sound features yield different performances. With the same HMM classifier utilized in Section 5.4.1, the accuracy and calculation cost of the MFCC [84] and three Haar-like features (α=0, 0.5, 1.0, *H*aarFilNum=5, *H*aarWidMax=18) are compared. The process of the MFCC feature extraction is complex which contains 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. (5.6) denotes. The experimental results shown in Fig. 5.7 and Table 5-3 prove that our proposed Haar+HMM outperforms MFCC+HMM in terms of both accuracy and calculation cost. The most aggressive case with α=1.0 can obtain 96.3% 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. 5.6, Fig. 5.7 and Table 5-3, 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 1.9% 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 α=0 case. It is because the filters number N in Eq. (5.3)

deceases with increasing α and further reduces the calculation cost in sounds 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.

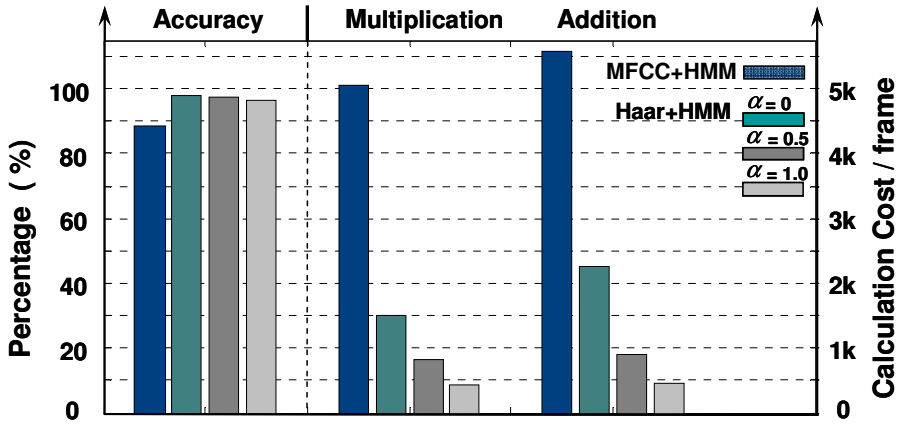


Figure 5.7 Performance Comparison of Proposed Haar-like and Traditional MFCC Sound Features with Same HMM Classifier – Average Accuracy and Multiply / Addition Calculation Cost (256 samples/frame).

Table 5-3: Different Sound Feature – MFCC and Haar-like Feature (α=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 (α = 0)	98.2%	1,510 (MFCC's 29.9%)	2,255 (MFCC's 40.4%)
Haar + HMM (α = 0.5)	97.7%	0,834 (MFCC's 16.5%)	0,903 (MFCC's 16.2%)
Haar + HMM (α = 1.0)	96.3%	0,420 (MFCC's 8.3%) (α =0's 27.8%)	0,456 (MFCC's 8.2%) (α =0's 20.2%)

5.4.3 Performance Comparison of Different Classifiers

With the same α=1.0 Haar-like feature configuration used in Section 5.4.2, the performance of the HMM classifier is investigated with the referenced *k*-means and LBG classifiers. The comparison results are shown in Fig. 5.8, Table 5-4 and Table 5-5. The clusters number of the HMM and the *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 96.3% among these three cases.

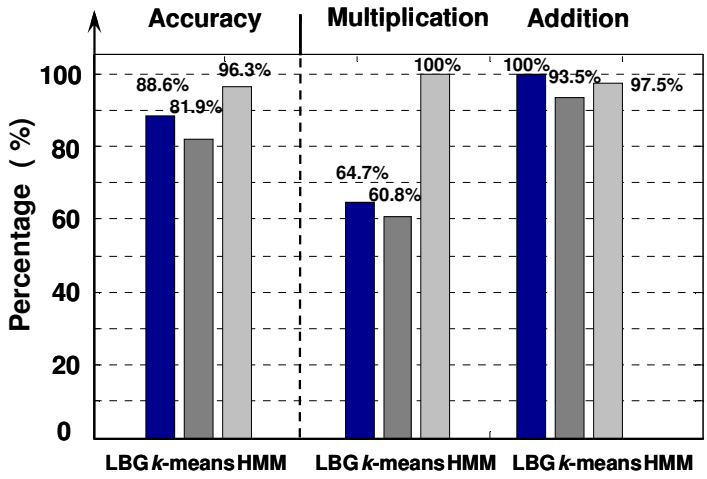


Figure 5.8 Performance Comparison of LBG, *k*-means and HMM Classifiers with Same Haar-like Sound Feature (Haar-like Feature's α=1.0).

During the classification, the HMM classifier needs more computation than the *k*-means classifier does. As in Section 5.2.4 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 which 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. (5.9) indicated. These obviously lead to an increase of multiplication calculation compared with *k*-means cluster in Fig. 5.8.

Table 5-4: Recognition Accuracy Confusion Matrix of 20 Different Tested Sounds with Haar+HMM Algorithm (α=1.0); Accuracy Comparison with Other Haar+HMM Two Cases (α=0/0.5), Haar+*k*-means and Haar+LBG.

	Sound Length	20 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	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	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	169	170	**99.4%**	100%	100%	68.2%	85.9%
A3	141.4	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	140	141	**99.3%**	94.3 %	100%	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	26	26	**100%**	100%	96.2%	88.5%	76.9%
A5	71.3	1	0	0	2	54	5	0	0	0	0	0	5	0	0	0	0	0	2	2	0	54	71	**76.1%**	94.4%	95.8%	12.7%	56.3%
A6	105.7	3	0	0	0	0	99	0	2	0	0	0	0	0	0	0	0	0	1	0	0	99	105	**94.3%**	97.1%	99.1%	30.5%	96.2%
A7	25.0	0	0	0	0	0	0	20	0	1	0	0	0	0	1	0	0	3	0	0	0	20	25	**80.0%**	84%	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	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	17	17	**100%**	94.1%	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	66	66	**100%**	100%	95.5%	100%	92.4%
A11	20.4	0	0	1	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	19	20	**95.0%**	100%	100%	70.0%	75.0%
A12	51.5	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	0	0	51	51	**100%**	96.1%	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	34	35	**97.1%**	100%	97.1%	54.3%	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	20	21	**95.2%**	100%	100%	100%	90.5%
A15	36.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	36	36	**100%**	94.4%	100%	88.9%	97.2%
A16	22.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	22	22	**100%**	100%	81.8%	90.9%	81.8%
A17	40.2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	39	40	**97.5%**	100%	100%	100%	100%
A18	142.3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	140	0	0	140	142	**98.6%**	95.1%	95.1%	99.3%	99.3%
A19	24.5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	23	24	**95.8%**	100%	100%	100%	100%
A20	21.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	21	21	**100%**	90.5%	95.2%	100%	100%
		Average Accuracy																						**96.3%**	97.7%	98.2%	**81.9 %**	**88.6%**

***Cu*: C**orrectly* recognized ***u**nits* **Au*: *A**ll* 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 something) 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 outside A15: Run
A16: Train start (train accelerates, in train) A17: Train run (train normally runs, in train) A18: Take umbrella in the rain
A19: Mechanical alarm A20: Telephone ring (telephone comes)

5.4.4 Performance Comparison of Whole System

Performance comparison of the recognition algorithms of different environmental sounds is illustrated in Table 5-5 and Fig. 5.9. Results of four algorithms – MFCC+HMM, Haar+LBG, Haar+*k*-mean, and Haar+HMM are compared. Among them, accuracy of the three algorithms: MFCC+HMM, Haar+LBG, Haar+HMM outperforms the 82% benchmark decided in Chapter 3. The highest accuracy is achieved by the Haar+HMM algorithm.

In Fig. 5.9, 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 96.3%–81.9%=14.4% due to the effective HMM classification. These results prove that the HMM classifier has better accuracy performance than the *k*-mean classifier with more calculation cost.

Within the top left region confined by the two benchmarks in Fig. 5.9, the Haar+HMM algorithm achieves better comprehensive performance. As a baseline for comparison, the Haar+LBG algorithm proposed by Nishimura is employed [93]. Our proposed Haar+HMM method 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 96.3% accuracy than the Haar+LBG's 88.6%. When the requirement of the calculation energy becomes stricter, Haar+LBG can be a candidate solution.

Table 5-5: Comprehensive Performance Comparison of Four Different Sound Recognition Algorithms - MFCC+HMM, Haar+LBG, Haar+*k*-means, Haar+HMM (1 Second / *u*nit =124 Frames in Each Second Sound *u*nit, Haar-like Feature's α=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)	88.6%	0.052	0.056	0.198	0.358	0.250	0.414	0.509
Haar + *k*-means (*H*aarFilNum=5, training centroids=15)	81.9%	0.052	0.056	0.186	0.334	0.238	0.390	0.483
Haar + HMM (*H*aarFilNum=5, training centroids=15)	**96.3%**	**0.052**	**0.056**	**0.306**	**0.349**	**0.358**	**0.405**	**0.661**

* The whole Energy = 1.44nJ × Mul. + 0.36nJ × Add (mil. nJ) based on Section 3's discussion.

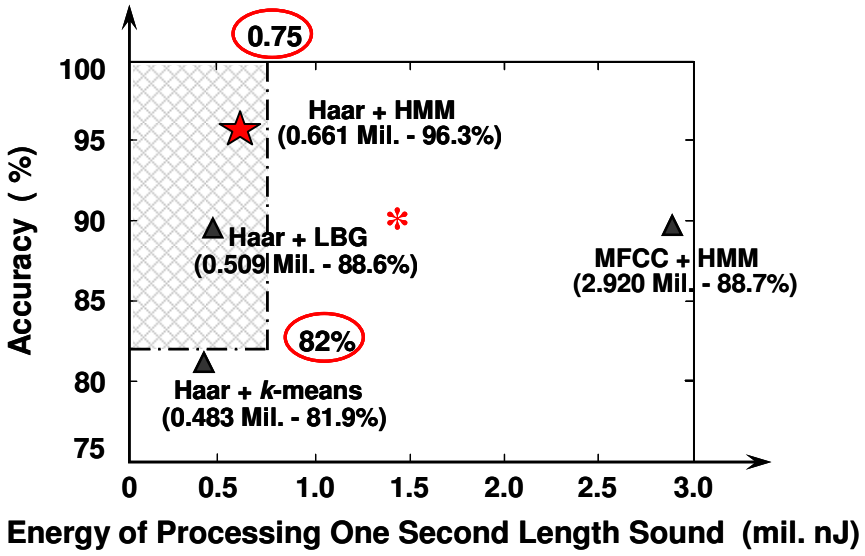


Figure 5.9 Performance Comparison of MFCC+HMM, Haar+LBG, Haar+*k*-means, and Haar+HMM (Haar-like Feature's α=1.0). (* Ref. [93_*J. Nishimura*_ICSP'2008])

5.5 Chapter Summary

As in Chapter 3 discussed, two factors benchmark values for evaluation of our proposed sound recognition algorithms have been decided. Because Haar-like filtering possesses low cost calculation character in sound feature extraction stage and HMM classifier possesses high recognition performance character in classification stage, our newly proposed Haar+HMM algorithm is utilized for the sounds recognition in this chapter.

Twenty different typical daily activities' background sounds are recognized. The recognition accuracy reaches 96.3% which outperforms required 82%. At the same time, the energy spent on our proposed sound recognition algorithm is within the energy budget. These results prove that our proposed Haar+HMM algorithm can successfully solve the problems and satisfy the settled both accuracy and calculation cost requirements.

Chapter 6 Conclusions

In this chapter, conclusion of our research - “Low-Complex Environmental Sound Recognition Algorithms for Power-Aware Wireless Sensor Networks” is delivered. Some potential future working directions for the next research stage will also be proposed.

6.1 Conclusions

With the rapid development of the wireless sensor networks (WSNs), wearable sensing and computation technologies, understanding individual’s activities, social interaction, and group dynamics of a certain society becomes possible and plays an important role for creation a ubiquitous information society around us. This will inevitably enrich our life’s content and improve our society’s efficiency.

Environmental background sound is a good context indicator for human activities, and contains rich information for identifying individual and social behaviors. Therefore, many front-end wearable devices in the WSNs system with sound recognition function are widely used to trace and understand human activities. Because those front-end sensor nodes are low powered and the WSNs system has limited resource, design of these sound-based context recognition algorithms has two major challenges: limited computation resources and a strict power consumption requirement. Therefore, we address to develop a new sound recognition algorithm which can achieve high recognition accuracy while still meeting the wearable sensor’s power requirement in the dissertation.

In Chapter 2, our power-aware front-end sensor node’s hardware platform and software level recognition flow are introduced. The hardware parameters will be used for the evaluation of power consumption in the following Chapter 3 is specified. Important assumptions and constrains for the research are also discussed and explained. Finally, aiming

at achieving certain high recognition accuracy with less power consumption, we propose our basic approaches to satisfy this requirement in algorithm level.

In Chapter 3, the test sounds, experimental setup and details for our research are firstly introduced. Because the power budget of the wearable sensor node for our sound processing is limited, evaluation to the recognition algorithms must consider both accuracy and calculation cost. As the system's accuracy rate concerned, it has been decided as 82% by referencing relative researches. Another, the "Calculation Cost and General Energy Evaluation" method is utilized to calculate the power consumption of the algorithm executed inside the MCU of the node. The evaluation benchmarks of these two factors are concluded and illustrated in Fig. 3.2. To accomplish a proper sound recognition upon the power-aware sensor node, the algorithm must satisfy these two benchmarks.

In Chapter 4, MFCC+LBG algorithm is proposed to recognize the background environmental sound. Twenty typical daily activities sounds are recognized. As the system's two important factors – recognition accuracy and algorithm's calculation cost, we have a thorough study. Finally, the approximate power consumption of executing the algorithm upon our wearable sensor node has been evaluated. Compared with our previous study by utilizing MFCC+DTW method, the optimized MFCC+LBG sound recognition algorithm improves the recognition accuracy to 92.5%, better than 84.8% with the MFCC+DTW. If the algorithm is executed on our wearable sensor node, the power consumption is much less than that of MFCC+DTW method. However, the MFCC+LBG's power consumption (0.883mil. nJ/s) does not qualify the benchmark requirement (0.75mil. nJ/s) through the power consumption evaluation.

In Chapter 5, because Haar-like filtering possesses low cost calculation character in sound feature extraction stage and HMM classifier possesses high recognition performance

character in classification stage, our newly proposed Haar+HMM algorithm is utilized for the sounds recognition in this chapter. Twenty different typical daily activities' background sounds are recognized. The recognition accuracy reaches 96.3% which outperforms required 82%. At the same time, the energy spent on our proposed sound recognition algorithm is within the energy budget. These results prove that our proposed Haar+HMM algorithm can successfully solve the problems and satisfy the settled both accuracy and power consumption requirements.

6.2 Scope of Future Work

In this dissertation, the low-complex environmental sound recognition algorithms have been proposed. Based on the wearable sensor node electrical parameters, the power consumption of execution those algorithms on the sensor's platform have been approximately evaluated. The future researches can be continued as the following research directions described.

In a software level point of view

1: In this research, our test target 20 environmental background sounds are mainly produced from household activities. Sound-context recognition targets can be extended to more complex social activities, such as meeting and discussion, shopping, etc.

2: From the Table 5-5 of Chapter 5, we notice that the HMM classifier occupies much proportion of the total computational cost compared with the Haar-like sound feature. In order to further decrease the whole algorithm's calculation cost while without compromising

the accuracy, there might be some improvement for the HMM classifier. References [Chapter 5 of Ref. 61, 101, 102] have already had certain study and discussion.

3: In this research, our proposed solutions focus on algorithm level study. By using the proposed environmental sound recognition algorithms, it can achieve satisfying results by employing the low complexity Haar-like sound feature with high performance HMM classifier. After the sound recognition algorithm has been decided, how to optimize the detection system to achieve a better performance can be another research direction. In Stager's research [36, 37], the author proposed some methods to trade off and optimize the two important parameters - "recognition accuracy" and "power consumption" of a sound-context recognition system. Similar methodology can be a potential research direction of the future work.

In a hardware level point of view

1: Through the accuracy and power consumption evaluation, the results prove that our proposed algorithms are valid to be implemented on the power-aware wearable sensor node (Fig. 1.3 and Fig. 2.1), and ideal detection performance can be achieved. Therefore, to implement the sound-context detection algorithms on the senor node will be one of our future research directions. With the WSNs system platform, the implemented algorithm's performance is to be evaluated.

2: Integrate with other type of sensor(s) [12, 15, 17, 109], such as accelerometer, IR sensor, thermo sensor, etc. to enhance daily activity recognition function by using wireless sensor networks (WSNs) system is also a potential research direction.

Bibliography

[1] http://en.wikipedia.org/wiki/WSN. (Accessed Jun. 2012)

[2] F. Zhao and L. J. Guibas, "Wireless Sensor Networks: An Information Processing Approach," *Elsevie-Morgan Kaufmann Publishers*, 2004.

[3] D. Culler, D. Estrin and M. Srivastava, "Overview of Sensor Networks," in *Computer*, Vol. 37, No. 8, pp. 41-49, Aug. 2004.

[4] I. F. Akyildiz, W. Su, et al., "A Survey on Sensor Networks", in *IEEE Communication Magazine*, Vol. 40, No. 8, pp. 102-114, Aug. 2002.

[5] J. Heidemann and R. Govindan, "An Overview of Embedded Sensor Networks," in *a Chapter in Handbook of Networked and Embedded Control System*, Springer-Verlag, 2004.

[6] G. Z. Yang, "Body Sensor Networks," *Springer*, 2006.

[7] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," in *Proceeding 1st ACM International Workshop on Wireless Sensor Networks and Applications*, pp. 88-97, Atlanta, 2002.

[8] B. F. Spencer, Jr., M E. Ruiz-Sandoval, and N. Kurata, "Smart Sensing Technology: Opportunities and Challenges," in *Journal of Structural Control and Health Monitoring*, Vol. 11, No. 4, pp. 349 – 368, Sept. 2004.

[9] J. P. Lynch and K. J. Loh, "A Summary Review of Wireless Sensors and Sensor Networks for Structural Health Monitoring," in *The Shock and Vibration Digest*, Vol. 38, No. 2, pp. 91–128, Mar. 2006.

[10] A. Pentland, T. Choudbury, N. Eagle, and P. Singh, "Human Dynamics: Computation for Organizations," in *Pattern Recognition Letters*, Vol. 26, No. 4, pp. 503-511, 2005.

[11] A. Pentland, "Socially Aware Computation and Communication," in *IEEE Computer*, Vol. 38, No. 3, pp. 33-40, 2005.

[12] T. Choudbury, "Sensing and Modeling Human Networks," *MIT Ph.D Thesis*, 2004.

[13] M. Laibowitz, J. Gips, R. Aylward, A. Pentland, and J. Paradiso, "A Sensor Network for Social Dynamic," in *IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 483-491, 2006.

[14] B. Clarkson and A. Penland, "Extracting Context from Environmental Audio," in *2nd International Symposium on Wearable Computers*, pp. 154-155, 1998.

[15] K. Yano, K. Ara, N. Moriwaki, and H. Kuriyama, "Measurement of Human Behavior: Creating a Society for Discovering Opportunities" in *Hitachi Review*, Vol.58, No. 4, pp. 139-143, 2009.

[16] S. Yamashita, et al., "A 15×15mm, 1μA, Reliable Sensor-Net Module: Enabling Application-Specific Nodes," in *IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 383-390, Nashville, Tennessee, USA, Apr. 2006.

[17] K. Yano, N. Sato, Y. Wakisaka, S. Tsuji, N. Ohkubo, M. Hayakawa and N. Moriwaki, "Life Thermoscope: Integrated Microelectronics for Visualizing Hidden Life Rhythm," in *IEEE ISSCC Digest of Technical Papers*, pp. 136-137, 2008.

[18] T. Tanaka, S. Yamashita, K. Aiki, H. Kuriyama, and K. Yano, "Life Microscope: Continuous Daily-Activity Recording System with Tiny Wireless Sensor", in *IEEE International Conference on Networked Sensing Systems*, pp. 162-165, Jun. 2008.

[19] J. Nishimura, N. Sato, and T. Kuroda, "Speech 'Siglet' Detection for Business Microscope," in *Proceeding of IEEE Pervasive Computing and Communications (PerCom)*, pp. 147-152, Mar. 2008.

[20] T. Allen, "Architecture and Communication among Product Development Engineers," in *MIT Press*, Cambridge, MA, pp. 1–35, 1997.

[21] K. Ara, T. Akitomi, N. Sato, S. Tsuji, M. Hayakawa, Y. Wakisaka, N. Ohkubo, R. Otsuka, F. Beniyama, N. Moriwaki, and K. Yano, "Healthcare of an Organization: Using Wearable Sensors and Feedback System for Energizing Workers," in *IEEE 16th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 567-572, 2011.

[22] N. Sato, S. Tsuji, K. Yano, R. Otsuka, N. Moriwaki, K. Ara, Y. Wakisaka, N. Ohkubo, M. Hayakawa, and Y. Horry, "Knowledge-Creating Behavior Index for Improving Knowledge Worker's Productivity," in *IEEE Sixth International Conference on Networked Sensing Systems (INSS)*, 2009.

[23] S. Tsuji, N. Sato, K. Yano, R. Otsuka, N. Moriwaki, K. Ara, Y. Wakisaka, N. Ohkubo, M. Hayakawa, and Y. Horry, "Visualization of Knowledge-Creation Process Using Face-to-Face Communication Data," in *IEEE Sixth International Conference on Networked Sensing Systems (INSS)*, 2009.

[24] Y. Wakisaka, K. Ara, M. Hayakawa, Y. Horry, N. Moriwaki, N. Ohkubo, N. Sato, S. Tsuji, and K. Yano, "Beam-Scan Sensor Node: Reliable Sensing of Human Interactions in Organization," in *IEEE Sixth International Conference on Networked Sensing Systems (INSS)*, 2009.

[25] K. Ara, et al., "Sensible Organizations: Changing Our Businesses and Work Styles through Sensor Data," in *Journal of Information Processing*, Vol. 16, pp. 604-615, 2008.

[26] T. Choudhury, et al. "The Mobile Sensing Platform: an Embedded Activity Recognition System," in *Pervasive Computing*, pp. 32-41, 2008.

[27] S. Basu, "A Linked-HMM Model for Robust Voicing and Speech Detection," in *ICASSP 2003*, pp. 816-819, 2003.

[28] http://en.wikipedia.org/wiki/Activity_recognition. (Accessed Jun. 2012)

[29] Karen Z. Haigh, "Automation as Caregiver: a Survey of Issue and Technologies," in *American Association for Artificial Intelligence (AAAI)*, pp. 39-53, 2002.

[30] L. Bao and S. S. Intille, "Activity Recognition from User-Annotated Acceleration Data," in *PERVASIVE 2004, LNCS 3001*, pp. 1–17, 2004.

[31] J. Yin, "Sensor-Based Abnormal Human-Activity Detection," in *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 8, pp. 1082-1090, 2008.

[32] J. Yin, "Probabilistic Activity Recognition from Low-Level Sensors," *HKUST Ph.D Thesis*, 2006.

[33] A. Krause, et al., "Trading off Prediction Accuracy and Power Consumption for Context-aware Wearable Computing," in *Proceeding of the 9th IEEE International Symposium on Wearable Computers (ISWC)*, pp. 20–26, 2005.

[34] N. Rota and M. Thonnat, "Activity Recognition from Video Sequences Using Declarative Models," in *Proceedings of the 14th European Conference on Artificial Intelligence*, pp. 673-680, Aug. 2000.

[35] J. Chen, J. Zhang, A. H. Kam, and L. Shue, "Bathroom Activity Monitoring Based on Sound," in *PERVASIVE 2005, LNCS 3468*, pp. 47-61, 2005.

[36] M. Stager, P. Lukowicz, and G. Troster, "Power and Accuracy Trade-off in Sound-based Context Recognition Systems," in *Elsevier's Pervasive and Mobile Computing*, pp. 300-327, 2007.

[37] M. Stager, "Low-Power Sound-Based User Activity Recognition," *ETH Ph.D Thesis*, 2006.

[38] N. B. Bharatula, M. Stager, P. Lukowics, and G. Troster, "Empirical Study of Design Choices in Multi-Sensor Context Recognition Systems," in *the 2nd International Forum on Applied Wearable Computing (IFAWC)*, pp. 79–93, 2005.

[39] R. Munkong and B.H. Juang, "Auditory Perception and Cognition," in *IEEE Signal Processing Magazine*, Vol. 25, No. 3, pp. 98-117, May 2008.

[40] M. Cowling and R. Sitte, "Comparison of Techniques for Environmental Sound Recognition," in *Pattern Recognition Letters 24*, pp. 2895–2907, 2003.

[41] R. S. Goldhor, "Recognition of Environmental Sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.149–152, Apr. 1993.

[42] N. Kern, B. Schiele, and A. Schmidt, "Recognizing Context for Annotating a Live Life Recording," in *Personal and Ubiquitous Computing*, Vol. 11, No. 4, pp. 251-263, 2007.

[43] S. Chu, S. Narayanan, and C,-C. J. Kuo, "Environmental Sound Recognition with Time-Frequency Audio Features," in *IEEE Transactions on Audio, Speech and Language Processing*, Vol.17, No. 6, pp. 1142-1158, Aug. 2009.

[44] V. Peltonen, J. Tuomi, A. Klapuri, and J. Huopaniemi, and T. Sorsa, "Computational Auditory Scene Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 1941-1944, May 2002.

[45] A. J. Eronen, V. Peltonen, et al., "Audio-Based Context Recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 321-329, Jan. 2006.

[46] B. Gold and N. Morgan, "Speech and Audio Signal Processing," *John Wiley & Sons*, 2000.

[47] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," *Prentice-Hall, Englewood Cliff*, New Jersey, 1993.

[48] L. Ma, B. Milner, and D. Smith, "Acoustic Environment Classification," in *ACM Transactions on Speech and Language Processing*, Vol. 3, No. 2, pp. 1-22, Jul. 2006.

[49] L. Lu and A. Hanjalic, "Text-like Segmentation of General Audio for Content-based Retrieval," in *IEEE Transactions on Multimedia*, Vol. 11, No. 4, pp. 658-669, Jun. 2009.

[50] P. Lukowicz, J. Ward, H. Junker, M. Stager, G. Troster, A. Atrash and T. Starner, "Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers," in *LNCS*, Vol. 3001, pp. 18-32, 2004.

[51] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner, "Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers," in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 10, 2006.

[52] V. Berisha, H. Kwon, and A. Spanias, "Real-Time Collaborative Monitoring in Wireless Sensor Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 1120-1123, 2006.

[53] G. Wichern, H. Kwon, A. Spanias, A. Fink, and H. Thornburg, "Continuous Observation and Archival of Acoustic Scenes Using Wireless Sensor Networks," in

Proceeding of the 16th International Conference on Digital Signal Processing (DSP 2009), pp. 428-433, 2009.

[54] H. Kwon, H. Krishnamoorthi, V. Berisha, and A. Spanias, "A Sensor Network for Real-Time Acoustic Scene Analysis," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 169-172, 2009.

[55] K. Hanaoka, A. Takagi, and T. Nakajima, "A Software Infrastructure for Wearable Sensor Networks," in *IEEE Proceeding of the 12th International Conference on Embedded and Real-Time Computing*, pp. 27-53, 2006.

[56] T. Yamabe, A. Tekagi, and T. Nakajima, "Citron: A Context Information Acquisition Framework for Personal Devices," in *IEEE Proceedings of International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pp. 489-495, 2005.

[57] R. Dong, D. Hermann, E. Cornu, and E. Chau, "Low-Power Implementation of an HMM-based Sound Environment Classification Algorithm for Hearing Aid Application," in *Proceeding of 15th European Signal Processing Conference (EUSIPCO)*, 2007.

[58] M. Stager, P. Lukowicz, and G. Troster, "Implementation and Evaluation of a Low-power Sound-based User Activity Recognition System," in *Proceeding of the 8th IEEE International Symposium on Wearable Computers (ISWC)*, pp. 138–141, 2004.

[59] A. Kulakov, G. Stojanov, and D. Davcev, "Sound and Video Processing in Wireless Sensor Networks," in *IEEE Industrial Electronics, IECON*, pp. 3550-3555, 2006.

[60] S. Phadke, R. Limaye, S. Verma, and K. Subramanian, "On Design and Implementation of an Embedded Automatic Speech Recognition System," in *Proceeding of the 17th International Conference on VLSI Design (VLSID)*, 2004.

[61] A. Bonfiglio and D. R. Rossi, "Wearable Monitoring Systems," *Springer*, 2011.

[62] R. Veitch, L. M. Aubert, R. Woods, and S. Fischaber, "FPGA Implementation of a Pipelined Gaussian Calculation for HMM-Based Large Vocabulary Speech Recognition," in *International Journal of Reconfigurable Computing*, Vol. 2011.

[63] K. Yano and H. Kuriyama, "Human×Sensor: How Sensor Information Will Change Human, Organization, and Society," in *Hitachi Review*, Vol.89, No. 07, pp. 572-573, 2007. (in Japanese)

[64] H8S_ http://www.renesas.com/fmwk.jsp?cnt=h8s2218_h8s2212_root.jsp&fp=/products/mpumcu/h8s_family/h8s2200_series/h8s2218_h8s2212_group/. (Accessed Jun. 2012)

[65] K. Roy and M. C. Johnson, "Software Design for Low Power," in *Nato Advanced Study Institutes Series on Low Power Design in Deep Sub-micron Electronics*, pp. 433–460, 1997.

[66] V. Tiwari, S. Malik, and A. Wolfe, "Power Analysis of Embedded Software: A First Step Towards Software Power Minimization," in *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 2, No. 4, pp. 437–445, 1994.

[67] V. Tiwari, S. Malik, A. Wolfe, and M. T-C. Lee, "Instruction Level Power Analysis and Optimization of Software," in *Journal of VLSI Signal Processing*, Vol. 1, No. 18, pp. 1–18, 1997.

[68] www.ti.com (TI_DSP_TMS320C54x data sheet. Accessed Jun. 2012)

[69] www.xinlinx.com (Xinlinx_FPGA_Spartan-6 data sheet. Accessed Jun. 2012)

[70] A. Gatherer and E. Auslander, "The Application of Programmable DSPs in Mobile Communications," *John Wiley & Sons*, 2002.

[71] S. J. Morris Bamberg, A. Y. Benbasat, D. M. Scarborough, D. E. Krebs, and J. A. Paradiso, “Gait Analysis Using a Shoe-Integrated Wireless Sensor System,” in *IEEE Transactions on Information Technology in BioMedicine*, Vol. 12, No. 4, Jul. 2008.

[72] L. Doherty, B. A. Warneke, B. E. Boser, and K. S. J. Pister, "Energy and Performance Considerations for Smart Dust," in *International Journal of Parallel and Distributed Systems and Networks*, Vol. 4, No. 3, pp. 121-133, 2001.

[73] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, “Feature Analysis and Selection for Acoustic Event Detection”, in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 17-20, 2008.

[74] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, “Acoustic Fall Detection Using Gaussian Mixture Models and GMM Supervectors,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 69-72, 2009.

[75] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. S. Huang, “Real-world Acoustic Event Detection”, in *Pattern Recognition Letters*, pp. 1543-1551, 2010.

[76] Y. Zhan, S. Miura, J. Nishimura, and T. Kuroda, “Human Activity Recognition from Environmental Background Sounds for Wireless Sensor Networks,” in *IEEE International Conference on Networking, Sensing and Control(ICNSC)*, pp.307-312, Apr. 2007.

[77] Y. Zhan, J. Nishimura, and T. Kuroda, “Human Activity Recognition from Environmental Background Sounds for Wireless Sensor Networks,” in *IEEJ Transaction EIS*, Vol.130, No. 4, pp.565-572, 2010.

[78] Y. Zhan, T. Kuroda, “Wearable Sensor-Based Human Activity Recognition from Environmental Background Sounds,” in *Springer Journal of Ambient Intelligence and Humanized Computing (JAIHC)* (accepted and will appear on June 2012).

[79] D. Hwang, C. Mittelsteadt, and I. Verbauwhede, “Low Power Showdown: Comparison of Five DSP Platforms Implementation an LPC Speech Codec”, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1125-1128, 2001.

[80] S. Ravindran, D. Anderson, and M. Slaney, “Low-power Audio Classification for Ubiquitous Sensor Networks”, in *IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP)*, pp. 337-340, 2004.

[81] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[82] J. Nishimura and T. Kuroda, "Versatile Recognition Using Haar-Like Feature and Cascaded Classifier," in *IEEE Sensor Journal*, Vol. 10, No. 5, pp. 942-951, 2010.

[83] F. Zhen, G. Zhang, and Z. Song, “Comparisons of Different Implementations of MFCC,” in *J. Computer Science & Technology*, Vol.16, No. 6, pp.582-589, 2001.

[84] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 357-366, 1980.

[85] http://en.wikipedia.org/wiki/Mel_scale. (Accessed Jun. 2012)

[86] S.S. Stevens and J. Volkmann, “A Scale for the Measurement of the Psychological Magnitude Pitch,” in *J. A. S. A.*, pp. 185-190, Jan. 1937.

[87] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for Vector Quantizer Design," in *IEEE Transactions on Communications*, Vol. 28, No.1, pp. 84-95, 1980.

[88] http://www.data-compression.com/vq.shtml. (Accessed Jun. 2012)

[89] http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition/speaker_recognition.html. (Accessed Jun. 2012)

[90] N. J.-C. Wang, W. H. Tsai, and Lin-Shan Lee, "Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification," in *EUROSPEECH*, pp. 1385-1388, 2001.

[91] H. Toshio, Y. Nishida, H. Aizawa, S. Murakami, and H. Mizoguchi, "Sensor Network for Supporting Elderly Care Home," in *Sensor*, Vol. 2, pp. 575-578, 2004.

[92] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *Proc. IEEE Computer Vision Pattern Recognition*, pp. 511–518, 2001.

[93] J. Nishimura and T. Kuroda, "Low Cost Speech Detection Using Haar-like Filtering for Sensornet," in *9th International Conference on Signal Processing*, Vol. 3, pp.2608-2611, 2008.

[94] Y. Hanai, J. Nishimura, and T. Kuroda, "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, 2009

[95] J. Nishimura and T. Kuroda, "Haar-like Filtering Based Speech Detection Using Integral Signal for Sensornet," in *International Conference on Sensing Technology*, pp.52-56, Dec. 2008.

[96] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification, Second Edition," *John Wiley & Sons*, 2001.

[97] http://en.wikipedia.org/wiki/K-means_clustering. (Accessed Jun. 2012)

[98] L R. Welch, "Hidden Markov Models and the Baum-Welch Algorithm," in *IEEE Information Theory Society Newsletter*, Vol. 53, No. 4, Dec. 2003.

[99] D. Zhang, B. Guo, and Z. Yu, "The Emergence of Social and Community Intelligence," in *IEEE Computer*, Vol. 44, pp 21-28, 2011.

[100] www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html. (HMM tutorial on the Internet. Accessed Jun. 2012).

[101] N. Liu and B. C. Lovell, "Gesture Classification Using Hidden Markov Models and Viterbi Path Counting ," in *Proceeding of Digital Image Computing: Techniques and Applications*, pp. 273-282, 2003.

[102] R. I. A. Davis and B. C. Lovell, "Comparing and Evaluating HMM Ensemble Training Algorithms Using Train and Test and Condition Number Criteria," in *Journal of Pattern Analysis and Applications*, pp. 327-336, 2000.

[103] www.coa.edu/greatduckisland.htm. (Accessed Jun. 2012)

[104] J. Nishimura, "A Study on Versatile Recognition using Haar-like Features," *Keio University Ph.D Thesis*, 2012.

[105] T. Nishimura, S. Nakamura, K. Miki, and K. Shikano, "Environmental Sound Source Identification Based on Hidden Markov Model for Robust Speech Recognition," in *EUROSPEECH*, pp. 2157-2160, 2003.

[106] F. Su, L. Yang, and T. Lu, "Environmental Sound Classification for Scene Recognition using Local Discriminant Bases and HMM," in *19th ACM international conference on Multimedia*, pp. 1389-1392, 2011.

[107] C. Couvreur, V. Fontaine, P. Gaunard, and C. G. Mubikangiey, "Automatic Classification of Environmental Noise Events by Hidden Markov Models," in *Applied Acoustics*, Vol. 54, No. 3, pp. 187-206, 1998.

[108] A. Rabaoui, Z. Lachiri, and N. Ellouze, "Using HMM-based Classifier Adapted to Background Noises with Improved Sounds Features for Audio Surveillance Application," in *International Journal of Information and Communication Engineering*, Vol. 5, No. 1, pp. 46-55, 2009.

[109] L. Wang, T. Gu, X. Tao, and J. Lu, "Sensor-based Human Activity Recognition in a Multi-user Scenario", in *AMBIENT INTELLIGENCE, LNCS 5859*, pp. 78–87, 2009.

List of Abbreviations

ADC	analog to digital converter
AR	*A*ccuracy *R*ate
A_u	*A*ll input sound *u*nits
CPU	central processing unit
C_u	*C*orrectly recognized *u*nits
DCT	discrete cosine transform
DSP	digital signal processor
DTW	dynamic time wrapping
HMM	hidden Markov mode
FFT	fast Fourier transformation
FPGA	field-programmable gate array
fps	frame per second
GMM	Gaussian mixture model
IC	integrated circuit
IR	infrared ray
IS	integral signal
LDA	linear discriminant analysis
LBG	Linde-Buzo-Gray algorithm
LPCC	linear prediction cepstral coefficients
MCU	micro control unit
MEMS	micro-electro-mechanical system

MFCC	Mel-frequency cepstral coefficients
MP	matching pursuit
PDA	personal digital assistant
RAM	random access memory
RF	radio frequency
ROM	read only memory
SVM	support vector machine
VLSI	very-large-scale integration
VQ	vector quantization
WSNs	wireless sensor networks