APPROACHES TO NON-INTRUSIVE LOAD MONITORING (NILM) IN THE HOME

by

Stephen Makonin

BTech, British Columbia Institute of Technology, 2009

A DEPTH REPORT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2012

APPROVAL

Name:	Stephen Makonin
Degree:	Doctor of Philosophy
Title of Depth Report:	APPROACHES TO NON-INTRUSIVE LOAD MONITORING (NILM) IN THE HOME
Examining Committee:	Dr. Bob Hadley Chair
	Dr. Fred Popowich, Professor, Computing Science Simon Fraser University Senior Supervisor
	Dr. Veronica Dahl, Professor, Computing Science Simon Fraser University Supervisor
	Dr. Lyn Bartram, Associate Professor, Interactive Arts + Technology Simon Fraser University Supervisor
	Dr. Alan Mackworth, Professor, Computer Science The University of British Columbia Examiner
Date Approved:	October 24, 2012

Abstract

When designing and implementing an intelligent energy conservation system for the home, it is essential to have insight into the activities and actions of the occupants. In particular, it is important to understand what appliances are being used and when. In the computational sustainability research community this is known as load disaggregation or Non-Intrusive Load Monitoring (NILM). NILM is a foundational algorithm that can disaggregate a home's power usage into the individual appliances that are running, identify energy conservation opportunities. This depth report will focus on NILM algorithms, their use and evaluation. We will examine and evaluate the anatomy of NILM, looking at techniques using load monitoring, event detection, feature extraction, classification, and accuracy measurement.

Keywords: nonintrusive load monitoring; NILM; energy modelling; hidden Markov model; Viterbi algorithm; sustainability

Contents

List of Figures

Chapter 1

Introduction

Understanding how appliances within a home consume electricity (when, where, why, and how) can be challenging for the average homeowner or occupant. With an understanding of consumption patterns, home owners can conserve energy (whether the motive be financial, environmental, or both). The introduction of automation can aid in conservation activities through: *filtering* data by analysis, *informing* occupants of consumption, *presenting* occupants with conservation options, and *taking action* to conserve – a *conservation feedback loop* (see Figure 1.1).

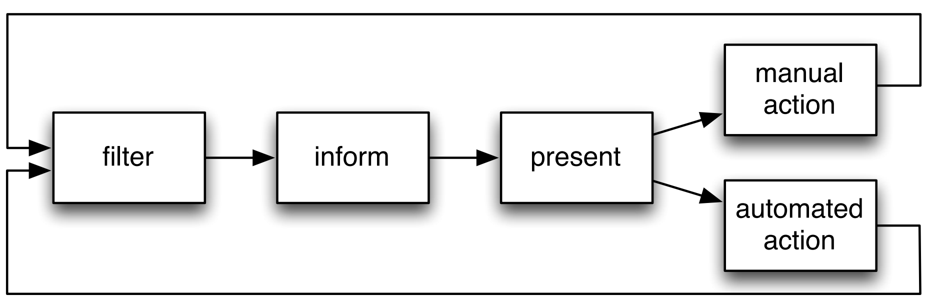


Figure 1.1: Our continuous *conservation feedback loop framework* for home energy conservation. As the occupants (and automated system) take actions, the effect of those actions should then become apparent when the occupants are next informed.

Currently the world is focused on reducing our consumption of electricity, but concern for other resources (such as natural gas, water, and grey water) are on the horizon. With a focus on electricity consumption we review the research on solving the hard problem of load disaggregation (a.k.a Non-Intrusive Load Monitoring or NILM). NILM algorithms can determine what appliances are running within a home from the power line analysis. Without a solution to NILM there is no accurate way to filter consumption data, inform about consumption/conservation issues, nor present conservation solutions to the homeowner that would allow them to understand their home's consumption and take action to conserve (*conservation feedback loop framework*).

1.1 Motivation

It is common knowledge that our increase in consumption is neither economically nor environmentally sustainable.

> *The essence of sustainable development is using our planet's resources as if we plan to stay. In the long term, economic sustainability depends on ecological sustainability. We must reassess and, where necessary, change our actions.*
> (Pew Oceans Commission (Arlington, Virginia), 2003)

Although the above quote comes from a publication on oceanic sustainability, the problems facing energy consumption are one in the same. There is now a growing consensus and realization that environmental sustainability and economical sustainability are inextricably linked.

Utilities (electricity providers) around the world are or will be bringing in *time-of-day usage tariffs*. These tariffs try to discourage power consumption during *peak hours* of usage and defer that consumption to *off-peak hours* (a.k.a. peak shaving). Peak shaving is a response to the rolling blackouts caused by demand exceeding supply[1]. In many cases this occurs during days of extreme hot weather when too many air conditioners are operating, cooling homes and businesses.

A blackout is an inconvenience to all. Residential homes consume about 34% of the total power consumption in the USA and are projected to increase to 39% by 2030 (Ehrhardt-Martinez et al., 2010). Homeowners and occupants can play a part in the conservation solution. But, before this can happen, homeowners need to be informed about how they consume. Many studies (Darby, 2006; Electric Power Research Institute, 2009; Granderson, 2009; Motegi et al., 2003; Parker et al., 2006) have shown that informed homeowners and occupants can and do reduce consumption (between 5% to 15%) when they are aware of their consumption behaviour. However, as we have stated before there can be a *conservation feedback loop*, the hard problem of

[1] A good example of this is the blackouts that occurred back in 2003 that affected much of Ontario and the Eastern United States. For more details see (Pereira, 2004).

NILM needs to be solved.

Without providing consumption details it is hard for occupants to know how energy was consumed and how different appliances contributed to the aggregate total. For example, say an occupant is notified every hour showing the aggregate amount of consumption that happened in the previous hour. The occupant is told that his consumption is 2.0 kWh for the previous hour (an increase of 0.2 kWh). Now the occupant is left to deduce from memory what appliances might have been on in the last hour. This can be an impossible task as the occupant has no way of relating 0.2 kWh to any one appliance or any number of appliances. The occupant might ask: *was the heating on? What does 0.2 kWh of heating feel like? Was it the lights that I forgot to turn off downstairs?* Details are needed to make consumption understandable to occupants as these details relate the consumption to the appliances being used. Details are needed to pinpoint consumption patterns and abnormalities. Consumption patterns and abnormalities, when found, can lead to identifying energy saving opportunities. NILM can provide the details needed for occupants to understand the aggregate total that was presented.

1.2 Defining NILM

Non-intrusive load monitoring (NILM), sometimes called non-intrusive appliance load monitoring (NALM or NIALM) or just load disaggregation, is an area of computational sustainability research that develops algorithms to disaggregate what appliances might be running from a metered/monitored power line. The non-intrusive part means using equipment that does not require house renovation or modification, or is visibly present. This area of research began in the late 1980s, early 1990s by researchers at the Electric Power Research Institute (EPRI) (Kitching et al., 1989) and Massachusetts Institute of Technology (MIT) (Hart, 1992).

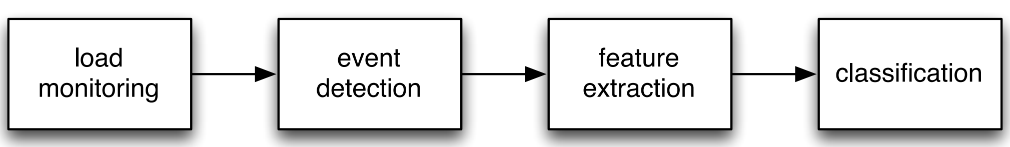


Figure 1.2: Block diagram of a typical NILM algorithm.

Figure 1.2 shows a block diagram of a typical NILM algorithm from Berges et al. (2010). Our report is organized in much the same manner, with each chapter delving into the details of each of the blocks in the diagram. Our approach in this report thus presents the reader with an efficient and effective *top-down view* of NILM algorithms. We round out our introduction to NILM with a subsection discussing why NILM is a hard problem.

The ability for NILM algorithms to perform load disaggregation accurately is a hard problem, but has the benefit of helping homeowners and occupants conserve energy by knowing how appliances within the home are used and how much energy they consume. We have identified eight various aspects as to why the performance accuracy of NILM algorithms is a hard problem.

P1: Different electrical characteristics

There is often a question of what measurements and electrical characteristics are needed for feature extraction and classification. Each measurement and electrical characteristic used adds dimensionality that can increase computational costs and complexity. We will discuss this in Section 2.1.

P2: Sampling rate

The lower the sampling rate the more errors in detection, due to event triggers (real power spikes) being missed. Higher sampling rates often require specialized measurement equipment which must either be purchased (which is costly) or be built (which is costly and inconvenient). See Section 2.2 for more details.

P3: Multiple, simultaneous load events

Multiple, simultaneous load events (switching on/off or changing state) can cause a miss-classification. For example, if we have two 100W lights and one 200W hand blender, then turning on two lights simultaneously could be mistakenly classified as the hand blender being used. Section 2.3 will examine this in more detail.

P4: Different appliance types

An increasing number of home appliances do not simply turn on/off. Appliances such as computers and other home electronics can have a standby mode. Kitchen ovens and clothes dryers are even more complex because as they operate they go through a number of states where heating elements and fans can be turned off and on in various combinations. This has further made the classification task difficult. We will look at this issue in Section 3.1.

P5: Noisy power signals

Electrical systems are inherently noisy. Causes of noise come from harmonic distortions, small fluctuations in appliance consumption, electronics that are constantly on, and appliances turning on/off with consumption levels too small to detect. Larger changes are better for event detection. Sections 3.2 and 3.3 touch on this.

P6: Dynamic and changing usage

Over a period of time, the appliances within a home can increase and decrease in number. They can also be replaced (e.g. an old dishwasher breaks down and replaced with new, more energy efficient model). Occupant-home interaction is very different when comparing the behaviour of one home to another. This can also be true when comparing a single home over a long period of time. So these dynamics must be accounted for in any NILM algorithm proposed. Chapter 4 brings light of this.

P7: Computational cost and complexity

For NILM algorithms to be practical they need to process online data and react in real-time to changes in the power being monitored. Some NILM algorithms, depending on the machine learning technique, can have a computational cost of $O(n^m)$. To reduce the computational cost, approximation algorithms are used resulting in higher classification inaccuracies. Although we discuss this concern throughout our paper, there is a concentrated discussion in Chapter 5.

P8: Accuracy measures

Some accuracy methods are questionable while others use established metrics (ROC, f-score). There is not yet a recognized standard of accuracy measurement for NILM algorithms. There are also issues related to the data being used to test NILM algorithms. For example, how real-world is the data. We will visit these issues in Chapter 6.

1.3 Summary

In the following chapters we will discuss in detail the four different components of a typical NILM algorithm (load monitoring, event detection, feature extraction, and classification) with a chapter devoted to each (Chapters 2–5, respectively). Chapter 6 will then discuss the problem of evaluation accuracy and the current methods used. We will continue with a chapter that further discusses broader issues at a higher level (Chapter 7) and close with conclusions and a look at future directions (Chapter 8).

Chapter 2

Load Monitoring

The first step in any NILM algorithm is to use a power meter[1] to measure the power on the line(s) we would like to monitor. When deciding power monitoring strategies, there are three important things to consider: the measurement types, the sample rates, and the sensing types. For measurement types (Section 2.1) we ask the question: *what do we want to measure?* For sample rates (Section 2.2) we ask the question: *how often do we want to measure it?* For sensing types (Section 2.3) we ask the question: *how many different points do we want to measure?* These three considerations are discussed in detail in the next three sections.

2.1 Measurement Types

There is often a question of what measurements and electrical characteristics are needed for feature extraction and classification. The most basic measurements are voltage and current. There are intermediate measurements derived from basic measurements: real power, energy, power factor, and reactive power. There are also advanced measurements which require signal analysis: harmonic distortion (background information see K. Lee et al., 2005; Wichakool et al., 2009) and electromagnetic interference (EMI). Electrical characteristics of these measurements can be used as well, such as startup and shutdown signatures, and aggregate values (minimum, maximum, average). The more advanced electrical characteristics (e.g. Energy Eigenvalues) require expensive and very specialized measurement equipment that can sample at high frequency (Liang et

[1] For more information on how power meters work see Ortiz et al. (2007) which also presents work on testing meter accuracy.

al., 2010a, 2010b). Additionally, appliance usage statistics have been added to the mix (e.g. on/off duration).

2.1.1 Survey of Measurements Used

As you might suspect, different researchers have been combining different combinations of measurements and electrical characteristics. Here is a short survey of what researchers have done. The next section continues the survey with a focus on research using real power measurement and smart meter data.

Hart (1992) used active power (P) and reactive power (Q) known as $\triangle P \triangle Q\ plot$ in addition to the appliance on/off duration. He found that the dishwasher had "ramping periods" (synonym to continuously variable) which would need a very complex FSM to model its full behaviour. He concluded that NILM is not suitable for "small appliances, continuously variable appliances, and appliances which are always on".

Norford and Leeb (1996) used active power and reactive power while looking at transient waves (see Section 4.2 for more details) for commercial buildings. They found that although they used reactive in their transient detector, "real power was sufficient to assess many of the major aspects". They noted that transient analysis is more advantageous for detecting equipment start-up and not equipment running at steady-state (see Section 4.1), but it comes at the cost of more computational power.

Laughman et al. (2003) extended the $\triangle P \triangle Q\ plot$ by adding harmonics as a third dimension. Adding harmonics allowed them to distinguish loads that overlapped on $\triangle P \triangle Q\ plots$. They discussed the fact that this type of transient analysis could lead to a form of equipment diagnostics, linking specific transients to specific equipment electrical and mechanical faults.

Fisera and Macek (2011) used only active power and reactive power in conjunction with the existing building management system (BMS). They found that their method had exponential computational cost as more equipment added. Computational cost was also burdened "with the increasing period of the control/status signals" provided by the BMS. They were, however, able to provide NILM results online in near real-time.

Chang et al. (2010) used voltage and current measurements. Where other researchers found that transient analysis was sufficient for NILM, they found it necessary to "combine transient and steady-state signatures ... to improve recognition accuracy and computational speed". They sampled voltage and current at a rate of 15kHz.

Figueiredo et al. (2011, 2012) used voltage, current, and power factor measurements. They concluded that simple NILM methods had high accuracy results when performing steady-state signature matching. However, they need to acquire more signature IDs to increase the accuracy and robustness.

Gupta et al. (2010) used EMI spectrum analysis to detect appliances, and found a number of interesting results. Appliances that are the same make and model had very similar signatures and that the signature is mostly independent of the home. But, the placement of the EMI sensor within a home affects the signal which is a “function of the line inductance” between the appliance and the sensor. This means that two identical appliances have different EMI signatures depending on were they where plugged in on the power line. They were able to detect near simultaneous appliance events at 102 milliseconds apart. Their solution required specialized equipment that needed to be trained on every appliance in the home.

Berenguer et al. (2008) used the electrical current startup signatures to detect appliances. They found that the short impulse, when the appliance was turned on, provided a unique signature for each different appliance. They observed that their system “produce[d] sensible information on the global activity of [a] person”.

S. Lee et al. (2010) used the different electrical characteristics of real power (e.g. raw, average, peak, etc.) measured from each appliance. They developed “activity-appliance models”. They felt these models “could help detect unattended appliances”.

Tsai and Lin (2011) used current electrical characteristic measurements (e.g. intensity, peak, average, etc.) on five appliances (fan, florescent light, radio, and microwave oven). They captured current waveforms as the appliances were being turned on at different voltage phase angles. This was done to create unique and repeatable profiles. They needed to reduce the sampling rate from evey 1 microsecond to every 500 microseconds to “reduce the computational burden and memory requirements of the system”.

2.1.2 Smart Meter Measurements

Newer research takes a simpler approach by only looking at real power measurements. The main reason for using only real power measurements is because this is the type of data recorded and communicated by smart meters. Another reason is that sensors that can be purchased and attached to the power meter outside the home are relatively inexpensive.

Baranski and Voss (2003, 2004a, 2004b) found the use of an optical sensor was an inexpensive option to read real power measurements off of a homes existing power meter. They used genetic algorithms that were able to create FMS of appliances that had high usage. The genetic algorithm could actively learn and discover new appliances that were either simple on/off or FMS (with $\leqslant 5$ states) without prior knowledge.

H. Kim et al. (2010) were motivated to use real power measurements because of the advent of smart meters. They found it difficult to disaggregate steady-states and needed to rely on appliance state changes. The accuracy of identifying appliances decreased as appliances were added, from one appliance (100%) to eight appliances (between 73–65%). They concluded that additional features (none mentioned) would need to be added to maintain higher accuracy as more appliances were added.

Berges et al. (2010) were also influenced by smart meters. They only presented results for one appliance (the refrigerator), no other appliances were monitored or tested. They found when the refrigerator went into defrost cycle there was significant error in detection. They identify the need for better signature capture and machine learning algorithms.

Zeifman (2012) identified a number of real world constraints for NILM algorithms, smart meters being one of them (see Section 7.1 for more details). He states that his solution meets all the real world constraints identified except for *various appliance types* where only simple on/off appliances could be detected. Even though he states that the *real-time capabilities* constrain was met, he only tested his system offline using MATLAB. Twenty-six days of data took two minutes of processing; therefore (he concluded) "real-time implementation is feasible".

Kolter et al. (2010); Kolter and Jaakkola (2012) were again influenced by smart meters. To be able to disaggregate with high accuracy they used complex unsupervised machine learning algorithms (see Section 5.4 for more details). Their NILM algorithm did have issues with distinguishing similar signatures.

Parson et al. (2012) used smart meters to gather real power measurements. They were able to tune a general appliance FSM to a specific appliance make and model across six different homes.

These researchers only considered real power because of the realization that *smart meters* are being installed on homes. Having access to power readings at no cost is a convenience that NILM cannot afford to overlook, but this would mean there is now a limit to the measurement types used by NILM algorithms (1 type). A reduction in measurement types used reduces the identification fidelity (Laughman et al., 2003).

2.2 Sample Rates

Concisely put, the higher the frequency in sampling the higher the accuracy of the NILM algorithm to detect and correctly classify an appliance being used. There are two categories of sampling rates NILM algorithms can use: high frequency (> 60Hz) and low frequency (⩽ 60Hz, but in the current research 1Hz or less). The sampling frequency needed is often determined by what measurements and electrical characteristics the NILM algorithm will use. For example, analyzing power signatures would require high frequency sampling.

High frequency sampling has to be greater than 60Hz–1 reading every 16.667 milliseconds (Norford & Leeb, 1996; Liang et al., 2010a). As appliances turn on/off they produce a unique signature that some NILM algorithms attempt to identify using pattern matching (Berenguer et al., 2008). In these cases power is monitored from rates of 10kHz (Berges et al., 2010), 15kHz (Chang et al., 2010), to rates of 100kHz (Patel et al., 2007). Power monitors that can sample at such high rates often require specialized measurement equipment that in many cases researchers either have custom built (e.g. Leeb et al., 1995; Baranski & Voss, 2003; Patel et al., 2007; Berenguer et al., 2008) or purchased (Liang et al., 2010a).

Low frequency sampling should be considered when focused on NILM for the home because of the installation of smart meters (Zeifman, 2012). Smart meters can communicate readings every 1–5 seconds (0.2–1Hz) (ZigBee Alliance, 2011). These rates are too infrequent for signature matching algorithms but more advanced machine learning algorithms can aid in uniquely identifying which appliances are being used (as discussed in Section 5).

2.3 Sensing Types

There are 2 strategies in which power can be monitored or sensed: single-point or multi-point. Single-point, as the name suggests, is to monitor power consumption for one point, usually the main power line that enters a home, or the point at which the utility places a meter on the side of a house. Multi-point is to monitor power consumption for more than one location. This can be done by metering one or more breakers within a house power panel (commonly known as branch circuit power metering or BCPM[2]) or by having specific appliances plugged into their own plug-level meter (e.g. Insteon iMeter Solo).

[2]An example of such a meter is the DENT PowerScout 18. More information at can be found at *http://www.dentinstruments.com/power_meter.html*

2.3.1 Single-Point Sensing

There are a number of reasons why single-point sensing is attractive: (1) relatively low cost, (2) less evasive installation, and (3) a generalized solution. Most, if not all, research is focused on single-point sensing. As smart meters are installed on homes around the world, the single sensor needed to achieve NILM is present. There is no need to install additional costly meters. However, trying to disaggregate appliances being used from a single monitoring point is a hard problem when it comes to accuracy. There are also issues with having ground truth to test the accuracy of NILM algorithms.

Early research relied on occupants manually recording their activity, in staged experiments (Froehlich et al., 2011). This creates a questionable ground truth because it does not represent a real world example which is more dynamic. Smart home researchers (Intille et al., 2006; Tapia et al., 2004) determined that real homes are important in evaluating the accuracy of intelligent systems. Research that requires the occupants to record their activities as a way to collect ground truth will have inaccurate ground truth (Intille et al., 2006). Multi-point sensing is now being used to establish a ground truth for accuracy testing (see Section 6 for additional details).

2.3.2 Multi-Point Sensing

Accurate ground truth is less of an issue with multi-point sensing; what is metered is accurate. In most cases large appliances with their own circuit breaker (e.g. furnace, clothes dryer, kitchen oven) are ideal for BCMP. Smaller appliances that can share the same circuit breaker (usually 15A circuits e.g. fridge, dishwasher, toaster, electronics) are well suited to have a plug-level meter. However, there is a consensus among researchers (Berges et al., 2010; H. Kim et al., 2010; Tsai & Lin, 2011; Froehlich et al., 2011; Zeifman, 2012) that the costs to purchase and install the required additional equipment make this option impractical for a final NILM algorithmic solution.

There is one final argument to make for using single-point sensing. If you cannot measure everything you still have a single-point sensing problem, albeit at a smaller scale. For example, if you can only afford to monitor the circuit breakers and the majority of breakers have more then two appliances serviced by then you still need to disaggregate the load from the circuit breaker. Countering, one can argue that using partial multi-point sensing can mitigate scalability problems by reducing the number of appliances that need to be disaggregated at one monitoring point, reducing the computational cost.

2.4 Summary and Discussion

It would be impractical (both in terms of cost and current state of the technology) to have a meter at every breaker and appliance. A more practical approach is to have some sort of metering on main appliances and an NILM algorithm to disaggregate the other smaller loads (e.g. lighting). But, is this the right focus?

At the consumer level, the future is smart meters. This means that the consumer has single-point metering for free. Typically the consumer will have access to real power (kW) and energy consumption (kWh) at a rate of 1 read every 1–5 seconds (ZigBee Alliance, 2011) from the smart meter. However, sampling rates which are too low (e.g. hourly) can result in low accuracy (as with Kolter et al., 2010, identifying the correct appliance only 59% of the time).

Measurements that involve harmonics, EMI, and transient waves require additional equipment with an associated additional cost for the consumer, and are thus unlikely to be adopted anytime soon. The combination of measurements may not be possible, as well. This is due to the data available from the smart meter. Acquiring other measures would require getting the consumer to purchase additional equipment to measure the signal characteristics (of power and current) which is a complex and time-consuming task (usually 2 days/appliance, see: Cavallo & Mapp, 2000) and is a barrier for consumer adoption. Consumers having to buy additional equipment and running that equipment can make energy conservation efforts unaffordable. The cost to run more equipment (which will consume more energy) can be greater than the energy saved, a point that seems to be missing in most papers.

Chapter 3

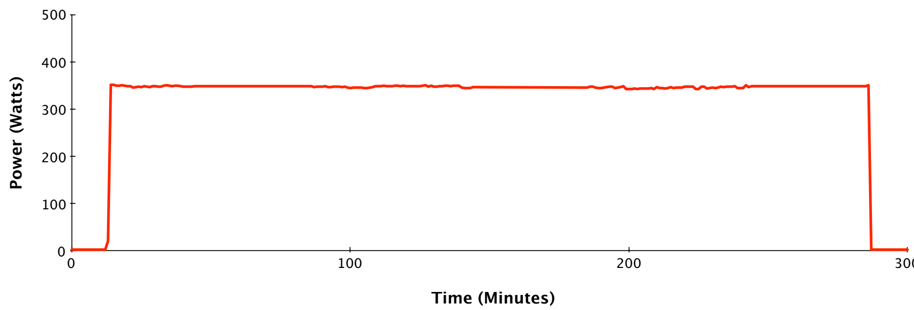
Event Detection

While the load on the main buswork of a home is being monitored, a way of detecting if an event has occurred needs to be developed. Event detection is complicated by the fact that homes have different appliance types (Section 3.1). So strategies, such as edge detection (Section 3.2), which check for rising and falling load levels, are used but can be problematic because edges may be too small to detect. More advanced approaches see edges as a probability distribution (Section 3.3). Without a reliable way to detect events, NILM algorithms cannot proceed with the tasks of feature extraction and classification. Events not detected contribute to inaccuracies.

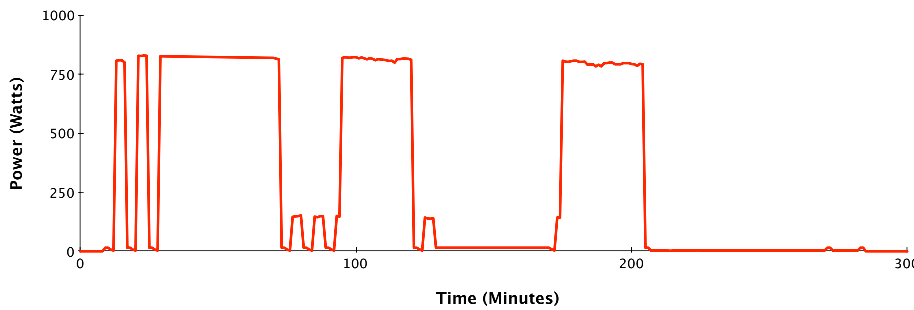
3.1 Appliance Types

Hart (1992) identified four appliance types: simple on/off, finite state, constantly on, and continuously variable. Simple on/off appliances are the easiest to detect and they are only ever on or off. In his more recent work, Zeifman (2012) developed an NILM algorithm that can only detect simple on/off appliances. Some examples of simple on/off appliances include: a light bulb, a toaster, and a kettle. Figure 3.1(a) shows an example of what the power readings would look like for lights being turned on then off.

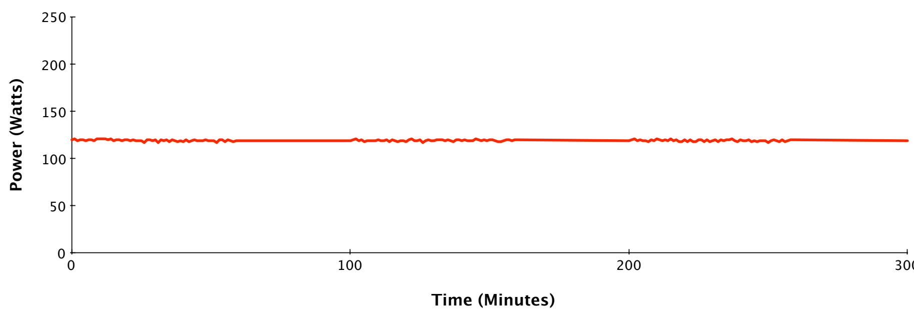
Finite state appliances are more complex and can be modelled using a finite state machine (Hart, 1992). Appliances like dishwashers, clothes dryers, and wall ovens do not just simply turn off and on. These appliances have a number of electrical components within them that may be turned on and off at different times. For instance, a dish washer has many wash cycles that start and stop, a water pump that drains the water between wash cycles, and a heating element that heats water (during wash cycles), dries the dishes, and can keep plates warm. This can be



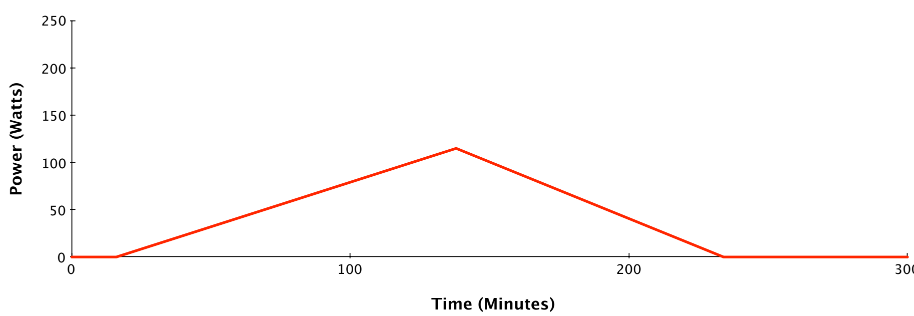
(a) Simple On/Off



(b) Finite State



(c) Constantly On



(d) Continuously Variable

Figure 3.1: Power monitored for 5 hours (readings taken once per minute) showing: (a) a set of six 65W light bulbs being turned on and then off, (b) a dishwasher washing a full load of dishes, (c) the HVAC fan which runs constantly for air circulation, and (d) a variable appliance such as a light dimmer being constantly modified for periods of time creating a ramp-effect.

further complicated with different temperature settings and other selectable options on some appliances. Figure 3.1(b) shows an example of what the power readings would look like for a dishwasher.

Constantly on appliances are particularly problematic as NILM algorithms detect events by rising and falling load levels. If an appliance is always on then no event is detected. Constantly on appliances do have small reading fluctuations but these fluctuations are too small for use in event detection. Figure 3.1(c) shows an example of what the power readings of a constantly on appliance would look like. Some examples of constantly on appliances are: smoke detectors, and alarm clocks. Notice that these appliances often consume only small amounts of power (usually $<$ 100W). This small amount of consumption is too small for detection and creates additive noise further impacting the accuracy of NILM algorithms (Chang et al., 2010).

Continuously variable appliances are problematic for NILM algorithms as well (Hart, 1992; Zeifman, 2012). The small increases and decreases in consumption (like a ramp) have the same event detection problems as constantly on appliances. A typical home does not contain many appliances that can be classified as continuously variable. A manual light dimmer would be considered a continuously variable appliance which we have demonstrated in Figure 3.1(d).

3.2 Edge Detection

Power monitored for the whole home is constantly changing (rising and falling, steps). These steps (if significant enough) can signal that an event has occurred. This is commonly known as edge detection (plotted in Figure 3.1) which some NILM algorithms use. Norford and Leeb (1996) used 3kW and 5kW steps to detect events in commercial buildings. The step sizes used are not suited for residential homes because the appliance consumption is far less (e.g. a heat pump, considered a high power consumer, turning on/off would generate steps of 1.7kW). In contrast Baranski and Voss (2004b) used 80W steps finding that most of the home appliances tested ranged between ±200W. Tsai and Lin (2011) based steps on an increase or decrease of current (Amps) with a difference of α, meaning $\Delta I_{intensity} \geqslant \alpha$ in the measured current signal (in this case $\alpha = 0.03$).

Edge detection can be extended from looking at one measurement to looking at a number of different measurements together, commonly referred to as clustering. Three-dimensional clusters were used by Laughman et al. (2003) using real power, reactive power, and harmonics. Some appliances can exhibit similar amounts of rising and falling but not for all types of measurements.

Using multiple measurements increases the accuracy of classifying the correct appliance. Other researchers have implemented complex multi-scalar edge detectors that can correctly identify events after some "tuning" (Leeb et al., 1995).

3.3 Probabilistic Approaches

There is benefit in replacing a simplistic, single-value edge detection algorithm with an algorithm that views edges as a probability distribution. Appliances that turn on/off or change running states do so creating different edge measurements which have a probability of matching to a specific appliance. Probability distributions have less complexity than the Leeb et al. (1995) multiscalar edge detector (Berges et al., 2010).

Berges et al. (2010) argues for a probabilistic approach and used the *generalized likelihood ratio* (GLR) work from Luo et al. (2002). Berges et al. (2010) modified the work by continuously computing the standard deviation instead of setting the parameter during initial training. They added a "voting scheme" that allowed each sample taken within the detection window to vote optimal change value to be selected. These modifications to the GLR improved the accuracy of the output.

Luo et al. (2002) published work on extending GLR. They monitored HVAC loads in commercial buildings at different frequencies (8Hz, 1Hz, 0.5Hz, etc). GLR used a ratio of probability distributions before and after a step change in load was detected. The natural log of this ratio was used to calculate a "decision statistic". In other words, instead of using edge detection they used a statistical algorithm that triggered on deviations from the mean power reading. After five hours of monitoring, they were able to detect 16 of 17 on/off events.

Kolter et al. (2010) used a probabilistic approach because edge detection is not suitable for very low sampling rates (hourly in this case). They used a *sparse coding* technique that focused on the task of disaggregation not classification and used hourly energy consumption amounts rather than real power readings. They then trained basis functions to detect appliance usage (see Section 4.3). They note that edge detection works best with high frequency sample rates.

3.4 Summary and Discussion

None of the current research literature examines the cost-benefit of monitoring and identifying loads with a focus on large appliances *vs* smaller appliances. This is an important issue for two reasons: (1) it determines the computational complexity of our event detection algorithm, and (2) utility companies have or are bringing in time-of-use charge rates. If we only need to detect large appliances then a simple, fast, efficient edge detection algorithm will work.

The cost of implementing complex and computationally intensive NILM algorithms may exceed the benefit and saving when trying to conserve consumption. For example, a simple NILM algorithm (that runs within the smart meter) that can detect the activity of large appliances (such as a clothes dryer) can ultimately save a large amount of money if the operation of the large appliance can be deferred until after peak hours (assuming time-of-use charge rates). In contrast, a more complex NILM algorithm may need the computational power of a PC (personal computer) to detect all appliance events within the home. But, does the detection of a light turning on or off have as much benefit? If we assume that the turning on of a light was a necessary action by an occupant and could not be deferred until later, then there may not be a benefit to detect all appliance events. This is especially the case when it comes to time-of-use charge rates. It is worth considering because using a PC to run NILM might cost more than the amount of energy savings realized by NILM.

Chapter 4

Feature Extraction

After an NILM algorithm has detected that an event has occurred it needs to compare various features of the monitored power signal. For example, these features could be different measurements (e.g. current, real power) that have resulted in a steady-state signature (Section 4.1), a transient signature (Section 4.2), or a derived basis function (Section 4.3). Labelled data (of the features selected) is used to train the NILM algorithm (Section 4.4). The right set of features need to be extracted and compared to the labelled data to allow the classification task to correctly identify appliances that have triggered events.

4.1 Steady-State Signatures

Steady-state signatures are monitored signals that do not change or change very little over time. A more formal definition would define a steady-state signal as a finite number of sinusoids or a fixed sum of sinusoids[1]. Figueiredo et al. (2011, 2012) define steady-state as “a difference between any two samples of a sequence [that] does not exceed a given tolerance value”. Using only real power steady-state signatures, they note that different appliances can have very similar signatures causing classification errors. To improve appliance distinction they base identification on “ratios between rectangular areas defined by the successive states values (Figueiredo et al., 2010). Steady-state signatures of low powered equipment (e.g. smoke alarm, cellphone charger, constantly on appliances) often go unclassified because there is no change in power signal (or no event detected) and are seen as signal noise.

[1] Definition retrieved from https://ccrma.stanford.edu/~jos/fp/Transient_Steady_State_Signals.html

4.2 Transient Signatures

Unlike steady-state signatures that change very little over time, transient signatures do. Each appliance is made up of different electrical components (e.g. resistors, capacitors, transistors, etc.) and it is the use of these components that can cause sudden changes in the power signal and create a unique transient signature (Laughman et al., 2003; Berenguer et al., 2008). Transient wave form analysis (background information see Chan et al., 2000; Tse et al., 2008; Leeb et al., 1995) can be used to capture signatures for different appliances and different monitored power lines. Once an event is detected the signal is captured which is then passed to a machine learning algorithm for pattern recognition (e.g. artificial neural network, support vector machine, nearest neighbour) that identified the type of appliance and its operation (e.g. turned on, cooling cycle, etc.) (e.g. Tsai & Lin, 2011; Berges et al., 2010). Capturing transient signatures requires high frequency sampling to obtain a high degree of signal uniqueness. Norford and Leeb (1996) stored transient signatures as a “precise time pattern of v-sections”. V-sections are the noted variations within the transient signature. Tsai and Lin (2011) calculated the maximum, average, and root mean square values of the transient signature as features used for subsequent classification using an artificial neural network and a nearest neighbour rule algorithm.

4.3 Basis Functions

Basis functions are a natural fit for NILM algorithms because we want to create mathematical descriptions of power readings (curves) over time. There are a number of standard basis functions available, with Fourier basis being one of the most commonly used in electrical engineering. Kolter et al. (2010) showed that “simplistic” edge detection is not sufficient; concluding that each appliance should have its own learnt basis for each of its states. They found that appliances like clothes dryers had a heavily peaked basis, refrigerators had a lower maximum magnitude basis, and lights had a “band pattern” basis (suggesting consistent times of use). Berges et al. (2009, 2010) used linear regression on both real and reactive power measures. The regression was performed using Fourier, polynomial, and radial basis functions (RBF). They found that Fourier regression coefficients provided best results when using a fixed time sample window for transient signatures.

4.4 Labelled Data

For NILM algorithms to be able to disaggregate and classify the events triggered by appliances there needs to be some prior knowledge of the home and the appliances within the home. The main task of the classifier is to label events that happen in the power signal based on selected features. For example, the clothes dyer was turned on *or* the clothes dryer is in the cooling off cycle. Labelling the unlabelled events/data helps in identifying how the home is using electricity. There are three main sources of labelled data that NILM algorithms currently use: signature corpora, finite state machines (FSM), and historical data. It is common for one NILM algorithm to use only one of these types.

Signature corpora are almost aways used in high frequency NILM algorithms where unlabelled steady-state and transient signatures (current readings) are compared to a set of existing labelled signatures (Tsai & Lin, 2011; Figueiredo et al., 2011). Signatures can be defined as a recorded power signal for a given amount of time for a given appliance. For instance, Fisera and Macek (2011) built the signature corpus during the training phase where control signals form the *building management system* (BMS). The BMS acted as a supervisor to help identify electrical events. Every time an event was identified the corresponding steady-state and transient signatures were stored in corpus and used to train the classifier.

For NILM algorithms that use low frequency sampling, signature corpora is not a suitable option; however, finite state machines (FSM) are (e.g. Hart, 1992; Norford & Leeb, 1996; Parson et al., 2012). A number of different appliance types (e.g. clothes dryer, dishwasher, oven, microwave, fridge, etc.) can be represented as an FSM; all but continuously variable appliances. The FSM has gone beyond a static set of data (unlike signature corpora). Parson et al. (2012) have published work that demonstrates how an FSM can be tuned while the NILM algorithm runs. In this case, a generic appliance FSM (e.g. clothes dryer) is trained to recognize a specific model of that appliance. Only the rough electrical characteristic values are entered for each generic appliance FSM.

Another source of data, for low frequency sampling NILM algorithms, is historical data. Historical data consists of periodic power readings that can be aggregated and analyzed in many different ways (e.g. histogram of appliance usage). Although such data is used primarily for testing datasets (detailed discussion in Section 6.1), historical data can be used in other ways. Baranski and Voss (2003, 2004a, 2004b) analyse the frequency of appliance usage. The historical usage of appliances that were used regularly were subjected to a genetic algorithm that generated an FSM for that appliance.

4.5 Summary and Discussion

One of the hardest parts of detection is detecting the steady-state signal. These are often unsmooth and (in the case of a smoke alarm) may never change–most often these signals are considered to be noise. Detection techniques need to be further investigated.

If we consider using smart meters for the load monitoring portion of our NILM algorithm, then capturing signatures and comparing them to a signature corpus is not a viable option. If we want an NILM algorithm that can be part of a “smart home” then there needs to be a solution that is mindful of homeowner’s and occupant’s time and patiences. Asking a homeowner or occupant to individually monitor each appliance so that the NILM algorithm can learn the appliance signature is impractical and a huge burden on the occupant. It can take up to 2 days of monitoring for each appliance to get an accurate enough reading for the NILM algorithm to learn the appliance’s power usage (Cavallo & Mapp, 2000). Asking an occupant to record the electrical characteristics (found on the sticker of each appliance) may be an acceptable solution to achieving the needed prior knowledge of the household environment.

Time of use has also been considered as a feature but this may only work for homes that have a “constant” routine (e.g. wash and dry clothes every Sunday morning). Kolter et al. (2010) found that lights had a “band pattern” basis which suggested that they where used at consistent times. However, we found that relying on “constant” routine (such as infant napping or nightly sleeping) may have consistent patterns in short-term (1–3 months) but longterm (8 months) continuity of the “constant” routine is broken (Makonin & Popowich, 2011, 2012).

Chapter 5

Classification

The final task in any NILM algorithm is to classify (with a high degree of accuracy) what appliance has triggered an event. Using the features extracted from the previous task and the labelled data, classification can now occur. Many researchers have used many different machine learning algorithms for the classification task. NILM algorithms that use supervised learning have implemented machine learning classifiers in a standard way. The most common supervised learning classifiers used are: Artificial Neural Networks (Section 5.1), Support Vector Machines (Section 5.2), and various Nearest Neighbour (Section 5.3) algorithms. NILM algorithms that use unsupervised learning classifiers have not been implemented in a standard way. Derivations have been presented and formalized in recent publications with a focus on using different implementations of factorial Hidden Markov Models (Section 5.4). Only the classifiers that are used most by researchers, or most recently used, or have the most promise are presented here.

5.1 Artificial Neural Networks (ANN)

An Artificial Neural Network (ANN) is a form of machine learning that is inspired from how neurons work in our brain; where individual neurons are connected to each other (this summary is based on Marsland, 2009). Each connection (between neurons) has a real value weight (w) where $0 \leqslant w \leqslant 1.0$ (typically). A neuron consists of an activation function (e.g. sigmoidal) that would fire if the sum (h) of the input connections (x) is greater than some predetermined threshold: $h = \sum_{i=1}^{I} w_i x_i$. A typical ANN consists of an input layer, 1 or more hidden layers, and an output layer. Each layer consists of a number of neurons that connect to the neurons in the layers immediately next to that layer. The number of neurons in each layer is arbitrary and is usually manually adjusted based on training results. Activations sent from the input layer neurons are sent to connected neurons in the hidden layer (e.g. input of “a” is encoded using 26 input neurons where the first neuron is activated and the others are not, representing “a”). This is repeated (forwards) for each hidden layer until the output layer is reached, where a result can be observed. This happens in training as well as after training. ANNs are typically trained using a backpropogation (BP) algorithm. The BP algorithm calculates the amount of error ($\delta = h' - h$, where h' is the expected h) on each output neuron then the weights of each input connection are modified accordingly. This is repeated (backwards) for all connections between each layer. Training sets are repeatedly (each iteration an epoch) used to train the connection weights of an ANN (through supervision). It takes an arbitrary amount of epochs before an ANN is to be considered trained–this is also a manually tuned process. There are issues with training this way, too much training and the ANN can become overfitted, too little training an an ANN may be stuck on a local maxima.

Even though ANNs are not considered to be theoretically sound like most modern machine learning algorithms, they are easy to use and there are a number of NILM researchers that use them. Tsai and Lin (2011) used a back-propagation ANN with mixed results. Although the classification accuracy was > 95%, there were a number of factors that were identified that discouraged their use. They needed to consider many factors when constructing the network and the optimal network was found through trial and error. Training is a necessity, both initial and continued. Each time appliances were added retraining was needed. Chang et al. (2010) provides a NILM algorithm using an ANN classifier to which they claimed 100% accuracy only if appliances are used one-at-a-time. However, when multiple appliances were used the training and testing accuracy diminished significantly ($\sim 59\%$ and $\sim 39\%$ respectively).

5.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) are a popular supervised training algorithm that use *optimal line separation* between classifications (this summary is based on Marsland, 2009). Optimal line separation is finding a line that separates classifications with the maximum margin where none of the training data falls within. Finding the maximum margin is needed because data used in test and implementation will not be the same as in training, in fact some of the non-training data may fall within the margin but still be classified correctly. SVMs generally perform significantly better at classification than other methods as long as the datasets used are not large (Marsland, 2009). Additionally, different kernel methods (basis functions) can be used within an SVM to create classification lines that are not linear. When comparing ANN *vs* SVM, SVM is preferred for two reasons: (1) SVMs do not suffer from converging on local maxima that are not the global maximum, and (2) the computational complexity is not tied to input dimensionality[1].

SVMs have been used in a number of NILM algorithms. Figueiredo et al. (2011, 2012) used pairwise SVM with a Linear Kernel and were able to disaggregate loads with high accuracy (~ 99%). They evaluated a Radial Basis Function Kernel (RBF) but found that the accuracy was ~ 90%. Pairwise (a.k.a. one-against-one) classifications were used to extend SVM from a binary classifier to a multi-classifier. An all-against-one classifier (also multi-class) was evaluated but results were less accurate (~ 88%). Kolter et al. (2010) used at SVM with "a variety of hand-engineered features" using low frequency sampling rate (hourly) with a classification accuracy of 59%.

5.3 Nearest Neighbour (*k*-NN)

Nearest Neighbour is a simple algorithm where classification is made when unlabelled data is found to be nearest (based on a distance function) to k labelled data neighbours. Data labelling is done using supervised training. For example, if $k = 1$ then the unlabelled data would be classified based on the single nearest data point in feature space. If $k = 5$ then the unlabelled data would be classified based on a majority vote of its neighbours (the classification with the highest count wins). The choice of k can cause classification problems. If k is too small then results are susceptible to noise and if k is too large accuracy diminishes (Marsland, 2009).

Nearest Neighbour has been used in a number of NILM algorithms. Figueiredo et al. (2011, 2012) used a 5-NN classifier to disaggregate with high accuracy (~ 99%). Gupta et al. (2010) claimed 100% accuracy using "KNN-based classifiers" but provided no details. Berges et al. (2009) used 1-NN with Fourier regression coefficients to classify unseen transient signatures with a 79% accuracy. They later improved their 1-NN using Euclidean distance measurements between signature vectors (~ 85% accurate, see: Berges et al., 2010).

Tsai and Lin (2011) used and prefered k-NNR (k-Nearest Neighbour Rule) for classification because of its simplicity of computation and implementation. Only one parameter (k) needs to be considered and is optimally found by an exhaustive search. No initial nor continued training was needed. Each time an appliance was added an exhaustive process was used to update

[1] As discussed in detail at http://www.svms.org/anns.html where a good discussion of SVM *vs* ANN is presented.

training data. The exhaustive search for the best k was considered the training process where the lowest k with the highest accuracy was used (which resulted in 7 neighbours with testing accuracies > 95%).

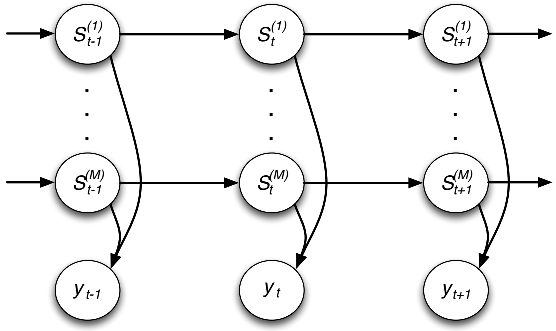
5.4 Factorial Hidden Markov Model (FHMM)

Hidden Markov Model (HMM) is a simple and efficient machine learning algorithm for modelling states over a length of time (Rabiner & Juang, 1986; Ghahramani, 2001; Marsland, 2009), which is well suited for NILM classification using only 1 appliance (an unlikely scenario). *Factorial HMM* (FHMM) is a derivative of HMM (Ghahramani & Jordan, 1997) which is well suited for NILM classification using multiple appliances. For this section let us assume the following nomenclature (adapted from Ghahramani & Jordan, 1997; Kolter & Jaakkola, 2012):

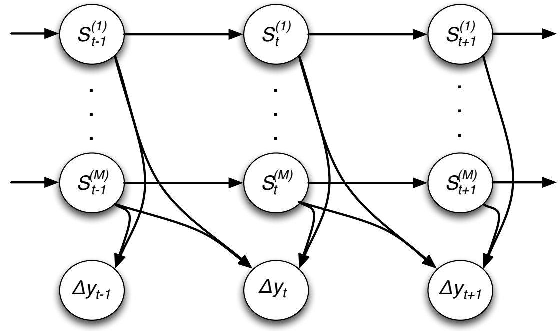
$$
\begin{array}{lcl}
T & = & \text{number of periods in the time series, } t = 1 \ldots T; \\
M & = & \text{number of appliances, } m = 1 \ldots M; \\
K^{(m)} & = & \text{number of load states for appliance } m, \{1 \ldots K\}; \\
S_t^{(m)} & = & \text{hidden load state of appliance } m \text{ at time } t; \\
x_t & = & \text{is the single-point load measurement at time } t\text{; and,} \\
y_t & = & \text{observed aggregate load state at time } t.
\end{array}
\tag{5.1}
$$

Each load state of an appliance is an a priori measurement (e.g. current, real power) that can be aggregated at time t using $y_t = \sum_{m=1}^{M} S_t^{(m)}$. The difference, $x_t - y_t$ would be the amount of load not accounted for in our FHMM model (this could be the result of error or noise). Each appliance would be considered a *Markov chain*. A Markov chain is a sequence of states where the probability of being in a state at time t depends only on the state at $t-1$. FHMM for a single observation to be determined from M Markov chains (in our case, appliances). An HMM using a single Markov chain would have K^M states (assuming that $K^{(m)} = K$ for all m) resulting in a $K^M \times K^M$ transition matrix, meaning exponential time and sampling complexity (Ghahramani & Jordan, 1997). Figure 5.1(a) is a graphical model representing FHMM using the nomenclature defined earlier.

FHMM has been used to model all states for all appliances (H. Kim et al., 2010; Kolter & Jaakkola, 2012; Parson et al., 2012). Essentially, each appliance has its own HMM that evolves in parallel to the others. FHMM avoids the exponential complexity of HMM, but at the cost of intractable exact inference so hidden state estimation is needed (e.g. Gibbs sampling). These



(a) Factorial HMM (FHMM)



(b) Difference FHMM

Figure 5.1: The (a) *factorial HMM* (FHMM) model and (b) *difference FHMM* mode adapted for NILM.

hidden states contain the load disaggregation information we need, we can only observe the resulting total load. In recent NILM literature a number of researchers have extended FHMM further to try solve the *hidden state estimation* problem of FHMM. Training an FHMM model using Baum-Welch (EM) and using Gibbs sampling provides for a straightforward supervised learning and hidden state estimation solution with an accuracy of ~ 65% (Kolter & Johnson, 2011). However, FHMM unsupervised solutions are much more complex.

H. Kim et al. (2010) used a combination of four FHMM variants to provide an unsupervised learning technique. FHMM was used in conjunction with a fractional hidden semi-Markov model (FHSMM), a conditional FHMM (CFHMM), and a conditional FHSMM (CFHSMM) and used Maximum Likelihood Estimation (MLE) to estimate the sequences of hidden states achieving accuracies of between 69%–98% (for 10 homes). Kolter and Jaakkola (2012) used a combination of two FHMM variants, additive FHMM (Figure 5.1(a)) and difference FHMM (Figure 5.1(b)). Additive FHMM was used for finding the aggregate observed load. Difference FHMM was used to find the difference in the load from $t-1$ and t. Using the difference, they performed state estimation using an Additive Fractional Approximate MAP algorithm (which they developed). They achieved an average accuracy of ~ 71% based on the classification of 7 appliances.

Parson et al. (2012) used the difference FHMM as well. Generic appliance FSMs and an extended Viterbi algorithm were used for hidden state estimation. The Viterbi algorithm was extended to ignore small joint probability observations and all sequences with joint probability were evaluated. It is hard to report on the accuracy as they reported *mean normalized error*. Zeifman (2012) proposed a Viterbi Algorithm with Sparse Transitions (VAST). They used a Markov

Chain that is optimally decoded by VAST using historical data and estimated distributions of appliance power and time features. They achieved accuracies with mixed results from one appliance at 41% to two appliances at 100% (9 simple on/off appliances were used).

5.5 Summary and Discussion

Machine learning algorithms such as ANN, SVM, k-NN and some forms of HMM that use supervised learning may not be the best approach to solving NILM. Asking occupants to "train their NILM before using" may not be the right approach–we believe that unsupervised learning algorithms are, as does Zeifman (2012). A machine learning technique such as k-NNR is memory intensive having large storage requirements and can be computationally intensive being susceptible to the *curse of dimensionality*. The curse of dimensionality basically states that the more features used (as inputs) the more sparse the dataset (Marsland, 2009) causing an increase in dataset size and increased training and classification times.

Researchers like Chang et al. (2010) fail to take into account *real world* situations when designing and evaluating experiments. Claims of 100% accuracy when only one appliance is being used at one time is really nothing significant and is a failure on the researcher's part as to understand what NILM research is about. With the exception of Zeifman (2012) all research (to this date) has been performed offline. Online performance is a key measure of algorithm viability that is all too often overlooked. Clearly, the task of classification needs more work.

Chapter 6

Evaluating Accuracy

A review of NILM algorithms and research has led us and others (H. Kim et al., 2010; Zeifman & Roth, 2011) to the conclusion that there is no consistent way to measure performance accuracy. However, this is slowly starting to change. Since 2011 a number of publicly available datasets have been released for researchers to use for testing (Section 6.1). Although some researchers still use the most basic forms of accuracy measure (Section 6.2) there has been discussion as to how to measure more accurately. The well known measure, f-score (Section 6.3) has been used. Recently, a modified version of f-score (Section 6.4) has been developed that is more suited to measuring NILM algorithm accuracy.

6.1 Datasets

The seriousness surrounding the accuracy measures has led to a number of high quality datasets being published and made publicly available. The MIT Reference Energy Disaggregation Data Set or REDD (Kolter & Johnson, 2011) supplies high and low frequency readings specifically for residential load disaggregation. The CMU Building-Level fUlly labeled Electricity Disaggregation dataset or BLUED (K. Anderson et al., 2012) contains high frequency readings of a single family home for one week specifically for residential load disaggregation. The UMASS Smart* Home Data Set (Barker et al., 2012) contains high and low frequency readings, but is not specifically designed for NILM evaluation. There are organizations that provide datasets for their customers.

Green Button has a number of sample dataset publicly available from their website[1]. The Plugwise dataset was used by Kolter et al. (2010) and Reinhardt et al. (2012), but is only available upon the submission of a request to the company.

6.2 Basic Accuracy

The most basic accuracy measure used by a majority of NILM researchers is defined as:

$$accuracy = \frac{correct\ matches}{total\ possible\ matches}\ . \tag{6.1}$$

Tsai and Lin (2011) used this accuracy measure by employing correct signals matched (no further details), and Chang et al. (2010) used recognition accuracy on both their training and the testing results. Both reported high accuracies of ⩾ 95% and 100%, respectively. These numbers can be misleading because they do not measure the classification's performance (Metz, 1978; Sokolova et al., 2006). For example, if a fridge is running only 10% of the time and an NILM algorithm (100% of that time) says the fridge was not running it would have a measured accuracy of 90%. H. Kim et al. (2010) points out that accuracy results are "very skewed because using an appliance is a relatively rare event .... appliances [that] are off will achieve high accuracy". Better accuracy performance measures need to be considered.

6.3 F-Score

F-score (a.k.a. f-measure or F_1score) is the harmonic mean of *precision* and *recall*:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}\ , \tag{6.2}$$

which has the more general form of

$$F_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}\ , \tag{6.3}$$

where eventually it is balanced when $\beta = 1$ (Sokolova et al., 2006). Precision is

$$precision = \frac{tp}{tp + fp}\ , \tag{6.4}$$

[1]These dataset can be downloaded from $http://www.greenbuttondata.org/greendevelop.aspx$.

and recall is

$$recall = \frac{tp}{tp + fn}, \tag{6.5}$$

where tp are true-positive results, fp are false-positive results, and fn are false-negative results. This form of accuracy measure is often found in information retrieval and text/document classification. Figueiredo et al. (2011, 2012) used f-score to measure their NILM algorithm based on 50 samples of data for each appliance. Berges et al. (2010) used f-score for both training and testing of their NILM algorithm comparing 5.5 days of consumption where the NILM algorithm predictions of an appliance was compared to the corresponding plug-level meter readings. H. Kim et al. (2010) argued that f-score measures binary classifier outcomes and power signals are not binary. This suggests that a better accuracy performance measure needs to be considered.

6.4 Modified F-Score

H. Kim et al. (2010) presented *modified f-score* which they argued was more suited for measuring the accuracy performance of NILM algorithms. They argue that we should not only measure the accuracy of the classification of the state of the appliance, additionally we should measure the accuracy of the predicted appliance consumption. To measure the accuracy of state classification we define the binary classification as such:

$$label = \begin{cases} positive, & \text{if } power > 0, \\ negative, & \text{otherwise}. \end{cases} \tag{6.6}$$

To measure the accuracy of the predicted appliance consumption, tp is divided into 2 measurements: accurate true-positives (atp), and inaccurate true-positives (itp)– predicted power consumption that is significantly different from the ground truth would be labelled as itp. The prediction is defined as follows:

$$prediction = \begin{cases} tn, & \text{if } \hat{c} = 0 \text{ and } c = 0, \\ fn, & \text{if } \hat{c} = 0 \text{ and } c > 0, \\ fp, & \text{if } \hat{c} > 0 \text{ and } c = 0, \\ atp, & \text{if } \hat{c} > 0 \text{ and } c > 0 \text{ and } \delta \leqslant \rho, \\ itp, & \text{if } \hat{c} > 0 \text{ and } c > 0 \text{ and } \delta > \rho, \end{cases} \tag{6.7}$$

where δ is the error ($\frac{\hat{c}-c}{c}$), ρ is the error threshold, c in the ground truth consumption, and $\hat{c}$ is the predicted consumption. Precision is redefined as

$$precision = \frac{atp}{atp + itp + fp}\,, \tag{6.8}$$

and recall is redefined as

$$recall = \frac{atp}{atp + itp + fn}\,. \tag{6.9}$$

Modified f-score is now the harmonic mean of the redefined precision and redefined recall.

Around the same time Zeifman and Roth (2011) showed that there are two types of errors that need accuracy measures: false detection (Type I) and missed detection (Type II). He argued that the *receiver operating characteristic* (ROC) curve (see Metz, 1978) depicts the "trade-off" between the *specificity* (Type I) and *sensitivity* (Type II) of those errors and this "trade-off" needs to be assessed. Specificity is

$$specificity = \frac{tn}{tn + fp}\,, \tag{6.10}$$

and sensitivity is the equivalent of recall ($sensitivity = recall$). ROC uses specificity and sensitivity to evaluate the performance of an algorithm on one type of classification (Sokolova et al., 2006). Later, Zeifman (2012) abandons the ROC argument in favour of the *modified f-score* measure.

6.5 Summary and Discussion

Zeifman and Roth (2011) discussed accuracy measures must be expanded to include detection accuracy, disaggregation accuracy, and overall accuracy– NILM algorithms are generally a complex conglomeration of a number of algorithms. Each algorithm that makes up NILM has a certain amount of error associated with it that is distinct and affects the down-chain algorithms and the overall accuracy and performance of the NILM algorithm. There are errors and inaccuracies in power monitoring which affect event detection. Errors and inaccuracies in event detection are compounded by load monitoring inaccuracies. Errors and inaccuracies in feature extraction are compounded by the previous two steps and compounded with errors in classification. By the time classification occurs the inaccuracies in load monitoring may contribute a significant amount to the inaccuracies in classification. Without examining each step (Figure 1.2) we cannot properly assess the impact of the algorithmic choices made at each step to determine how to improve/optimize the NILM algorithm.

Chapter 7

Analysis and Further Discussion

Looking at the broader picture, there are a number of broader issues that we need to look at and review. Implementing NILM in a residential environment can limit algorithm design (Section 7.1). Additionally, occupants must be engaged for an energy conservation system to work; providing information that is timely, convenient, and persuasive (Section 7.2). We also candidly reflect on what the situation is in the real world now (Section 7.3). These three sections will provide some grounding to all our previous discussions about the variety in the various parts of NILM algorithms.

7.1 Identifying Real World Constraints

Taking our focus on NILM algorithms in the home into account, there are a number of real world constraints that need to be met before NILM can be seen as a success for use within a home as part of a *conservation feedback loop framework*. Zeifman (2012) has identified six solution requirements that further restrict the solution space for applying NILM in a home. These are: feature selection, accuracy, no training, near real-time capabilities, scalability, and various appliance types.

Feature selection constrains suitable NILM algorithms to those that are measured and reported by smart meters (real power and energy readings at a rate of 0.2–1Hz). *Accuracy* constrains suitable NILM algorithms to a degree that is acceptable to occupants. Zeifman (2012) argues that "a minimum acceptable accuracy is 80-90%", but it may depend on the type of occupant. *No training* constrains suitable NILM algorithms to those that do "not involve significant occupant efforts" (Zeifman, 2012) to train and accurately disaggregate loads. *Near real-time capabilities* constrains suitable NILM algorithms to those that run online and respond to events as they happen; meaning that the algorithm must be robust and efficient. *Scalability* constrains suitable NILM algorithms to those that do not require additional processing time and/or hardware to account for the identification of new appliances. Zeifman (2012) states that a reasonable increase is between 10 to 20 appliances. *Various appliance types* constrains suitable NILM algorithms to those that can detect the four appliance types discussed in Section 3.1: simple on/off, finite state, constantly on, and continuously variable..

7.2 Occupant Engagement and HCI

As most Human-Computer Interaction (HCI) researchers will agree there are huge challenges in relaying feedback to occupants to influence consumption behaviour. Quintal et al. (2012) found that after a 3 month study of 9 households using a real-time eco-feedback device there was no significant amount of energy saving (40Wh) even though occupants made small adjustments in their consumption behaviour. They found that occupants "lost interest in the system after a while and even the small updates delivered over time were not enough to prevent this".

Bartram has done extensive work on occupant-home interaction by investigating and designing a framework for such devices (Rodgers & Bartram, 2010). Bartram has continued to look at issues affecting the ability of homeowners to conserve energy (Bartram & Woodbury, 2011; Bartram et al., 2011). Her research reiterates the difficulties for systems to communicate to the homeowner how their house is performing or even what to communicate. These difficulties most likely resulted in the occupants in Quintal et al. (2012) losing interest.

The study of current devices and the exploration of future designs have been a priority in trying to understand how to motivate and engage occupants (Froehlich et al., 2010; Horn et al., 2011; Pierce et al., 2008; Strengers, 2011; Rodgers & Bartram, 2011) and how occupants engage with a home (a.k.a. occupant intelligence; Cole & Brown, 2009; Leaman & Bordass, 2001). There is still a considerable amount of work to be done to provide energy consumption feedback either through eco-visualizations or eco-feedback devices in a timely, convenient, and persuasive manner.

7.2.1 Timely

Information presented in a timely manner is more than just meeting the *near real-time* constraint, it also includes presenting the information the occupant wants. Near real-time means having the eco-visualizations match visually the pattern of consumption that is happening. Lags in the consumption-to-visualization function can lead to a misunderstanding of what appliance is consuming at a given point in time. We found these lags to be a serious issue when creating persuasive eco-visualizations (Makonin et al., 2011).

Timely access to the right information is critical in motivating occupant engagement (which can also be classified as convenient). If an occupant must login to a computer and/or search through a program or menu to find the consumption feedback needed then that takes time. This may impact how often the occupant reviews consumption feedback, this is especially true when occupants perceive themselves to be busy (Kashani & Bartram, 2011). Devices that have these issues create a barrier to occupant engagement and create what DiSalvo et al. (2010) called *a wasteful rapid obsolescence cycle*, where occupants buy a device, try it, find it is too cumbersome, resulting in the device either being recycled, thrown into the garbage, or boxed up and forgotten. What must be done to mitigate these issues is to create feedback systems that understand different occupant preferences or personae (Clevenger & Haymaker, 2006; Noy et al., 2006; Goldstein et al., 2010).

Hutzler et al. (2000) published work on an ambient display called *Garden of Chances*, non-photorealistic art. They were able to convey environmental and weather conditions for a remote town in France. Different graphical elements of the canvas depict the season, whether it is raining, and/or it is windy. Subsequently, they performed a user study that evaluated the effectiveness of Garden of Chances *vs* text-based information (Hutzler & Gortais, 2004). They found that subjects could understand high-level information better and faster as compared to the text-based system. This suggests that ambient displays may need to be highly customizable based on an occupant's preferences so that the right consumption information would be seen when needed without much occupant interaction.

7.2.2 Convenient

Information presented in a convenient way means: (1) where the eco-feedback device is placed in a home, and (2) how home automation can help with energy conservation efforts without sacrificing occupant comfort. Rodgers and Bartram (2011) found that occupants considered the kitchen and the living/family room as two of the most convenient locations to place an eco-visualizations within the home. Recently, we confirmed this with a separate user that looked at lifestyle factors and placement (Makonin et al., 2012). Lifestyle factors are comprised of influences that emanate from the external environment, the household, and individual choices (Bin

& Dowlatabadi, 2005).

Equally important, in the area of smart homes, is involving a form of automation that can help conserve energy while still taking into account occupant comfort. However, there is often a conflicting goal of: (1) having a home running at optimal energy efficiency, and (2) allowing occupants to live in comfort and override the smart home system. Homes are particularly hard to manage (Eckl & MacWilliams, 2009) and they have challenged the usability of smart home systems (Brush et al., 2011) in terms of ease of system override and having the system provide information as to its operation. Chetty et al. (2008); Blevis (2007) have looked at increasing system usability by providing better information access and interaction design. Having a smart home system that can provide accurate information and can be easily overridden by the occupants decreases the apprehension that occupants feel (Woods, 1996; Velikov & Bartram, 2009) about smart home systems. This also increases the comfort and removes engagement barriers.

We have also explored these issues. Recently, we discussed case studies and design scenarios (Makonin & Popowich, 2011, 2012; Makonin et al., in press) that challenge the *running at optimal energy efficiency* way of thinking and shifts the focus to consider occupancy comfort. We believe there are a number of simple rules that can be enhanced with online learning algorithms that can achieve both energy conservation and comfort. Without taking comfort into account occupants may very well disable the automation.

7.2.3 Persuasive

Fogg (1999) introduced the idea of persuasive computing (or Captology) and described it as the science of a having computers influence our decision making. When used right, this can have many benefits. One potential area to explore persuasive computing is energy conservation and sustainability (Froehlich et al., 2010). Information presented in a persuasive way means attaching an emotional connection to an eco-visualization that may render differently depending on the amount of energy consumption occurring at that time.

He et al. (2010) proposed a *motivational framework* based on the Transtheoretical model of staged behaviour change. The motivational framework would help identify what types of persuasion would be needed to move the occupant from one stage to another. They identified five stages: precontemplation, contemplation, preparation, action, and maintenance/relapse/recycle. The *precontemplation* stage would try to inform those who did not perceive that there is a sustainability problem and encourage occupant action. The *contemplation* stage would try to further motivate doubtful occupants into identifying consumption behaviours and offering possible solutions. The *preparation* stage would help motivated occupants develop a short term plan with reasonable goals. The *action* stage would assist occupants taking action to identify problematic consumption behaviour and suggest appropriate actions. The *maintenance/relapse/recycle* stage would further motivate occupants from relapsing to problematic behaviours.

Torning and Oinas-Kukkonen (2009) have written extensively on persuasive system design issues creating a Persuasive System Design Model (PSD Model). They first identified three persuasion contexts: the intent, the event, and the strategy. The intent is the designer who is designing the persuasive system, (s)he has a specific goal in mind. The event is where the occupant uses the persuasive system and the persuasive system provides feedback through its designed features (e.g. identifying goals and measuring them). The strategy looks at the form and content of the message used by the designer to persuade and the use of a route (or a mode of argument) that could be direct (a persuasive argument) and/or indirect (providing facts). Next, they identified four design support principles: primary task, dialogue, system credibility, and social. Primary task support looks at target behaviours of the research. Dialogue support looks at the feedback that the persuasive system provides the occupant. System credibility support deals with the accuracy of the persuasive system and its believability. Social support looks at providing a social aspect to the persuasive system that would allow the occupant to compare conservation efforts, share advice, and work cooperatively within a multi-occupant home and within a community.

These issues also need to take into account design requirements for ambient displays. T. Kim et al. (2010) identifies ten such requirements. The information displayed must be focused and immediate. The display must not be a distraction. Subtle, ambient indicators help reduce distraction. What is displayed should be visually appealing and the representations should be iconic in nature. The information displayed must be accurate, data must not be contrived. Have the display convey reward for achievement to encourage and motivate developing an emotional attachment. Allow the display to provide feedback based on occupant preferences and personae. The display should remind an occupant of ways to conserve, especially if the system determines that an occupant may have forgotten. The display should also be connected to social media and websites that can drive community engagement.

7.3 Summary and Discussion

Our critique of the current research is motivated and grounded by *real world* situation. Because we are focused on NILM algorithms for the home, let us define what a home might look like between now and 5 years[1]. Firstly, we will see the installation of *smart meters* on residential homes globally. This is mainly due to energy conservation efforts by governments and initiatives by power utility companies. Smart meters are little more than digital meters with enhanced 2-way communication capabilities (commonly known as advanced metering infrastructure or AMI); they are not actually smart, nor do they contain any sort of artificial intelligence. These smart meters are often called *gateways into the home* as they communicate to the utility's central office over a WAN (wide area network) using protocols such as PLC (power line communication) or WiMAX (M. Anderson, 2010). They can also communicate within the home using a protocol such as ZigBee to other equipment in a HAN (home area network) (Farhangi, 2010).

In our time frame the most common equipment that the smart meter will communicate with over the HAN will be IHD (in-home displays). The IHD is used to report real power (kW) and energy (kWh) data from the smart meter at a low sampling rate ($\geqslant 1$ second) to the occupants of the house. A typical IHD is usually small and the information is displayed as numbers (much like a basic thermostat).

The idea of smart appliances has been discussed but the time frame for such devices is far on the horizon. This is mainly due to new appliances that would need to be bought creating a big financial burden on the homeowner. There has been no serious thought put toward retrofit solutions; a more eco-sustainable solution. As with smart appliances, the same can be said for DR/LC (demand response/load control) integration to perform automated peak shaving; a strategy to avoid grid brownouts when there is a sudden critical demand for power (Farhangi, 2010). This is evident in the US and Canada where air conditioners running on hot days cause power outages.

[1] Energy conservation needs to happen now. If we consider longterm projections then we face devaluing the urgency of the eco-sustainability problem which has a high cost (both financial and environmental). Some may even say that 5 years is too long.

Chapter 8

Conclusions

> *The future belongs to those who understand that doing more with less is compassionate, prosperous and enduring and thus more intelligent, even competitive.*
> (Paul Hawken)

The quote by Paul Hawken (above) is very relevant for today's environmental situation; in a way this is also true for NILM algorithm design. If we take into account the goals which we want to achieve (energy conservation in the home) we need to research and experiment with appropriate algorithms and data, with a goal of deployment and evaluation in ordinary homes. The definition of a successful NILM algorithm must be reassessed and realigned to the goal that we want to achieve.

When reflecting back at our general *conservation feedback loop framework* we see the NILM is only the beginning. There is still a great deal of work needed to be done to inform occupants how they are consuming energy and then recommending better ways of consumption to help save costs and the environment.

8.1 Future Directions

A number of researchers discuss the need to extend their NILM algorithms to include the classification and identification of complex appliances (Section 3.1), not just simple on/off appliances. Zeifman (2012) notes that no one NILM algorithmic solution can handle all four appliance types. There are many ways identified to achieve this and are discussed in the sections below.

Larger signature corpora

Figueiredo et al. (2011, 2012) plans to build a larger dataset of steady-state IDs. They concluded that having a larger dataset would increase the accuracy and robustness of their SVM and k-NN classifiers.

Data fusion techniques

Jiang et al. (2011) (in a small literature review comparing ANN *vs* SVM implementations) points out that no NILM algorithm can disaggregate all the different types of appliances discussed in Section 3.1. They recommend using data fusion techniques (e.g. Dampster Shafer (Li et al., 2010)) as a possible solution to this issue. These techniques can combine disparate solutions together. Zeifman and Roth (2011), in their literature survey, also agreed referencing the work of Basir and Yuan (2007). Zeifman (2012) reiterated that their NILM algorithm can be extended to use approaches found in Hart (1992) and Baranski and Voss (2003).

Increasing the time scale

Kolter and Jaakkola (2012) are investigating ways to lengthen the time scale used for their unsupervised learning procedure to classify appliances with complex states. They are also investigating a method that would iteratively examine the "unassigned" part of the power reading. This would allow them to "successively build models for more and more devices" in conjunction with joint inference and "hard EM" learning.

Additional and different data

Kolter and Jaakkola (2012) want to add features that include more sensor data (e.g. "vibrations from sensors"). They also want to include the monitoring of water and natural gas. Parson et al. (2012) is investigating adding appliance time-of-day usage using their current prior training method.

Bibliography

Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A., & Berges, M. (2012). BLUED: a fully labeled public dataset for Event-Based Non-Intrusive load monitoring research. In *Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD).* Beijing, China.

Anderson, M. (2010). Wimax for smart grids. *Spectrum, IEEE*, *47*(7), 14–14.

Baranski, M., & Voss, J. (2003). Nonintrusive appliance load monitoring based on an optical sensor. In *Power tech conference proceedings, 2003 ieee bologna* (Vol. 4, p. 8 pp. Vol.4).

Baranski, M., & Voss, J. (2004a). Detecting patterns of appliances from total load data using a dynamic programming approach. In *Data mining, 2004. icdm '04. fourth ieee international conference on* (p. 327 - 330).

Baranski, M., & Voss, J. (2004b). Genetic algorithm for pattern detection in nialm systems. In *Systems, man and cybernetics, 2004 ieee international conference on* (Vol. 4, p. 3462 - 3468 vol.4).

Barker, S., Mishra, A., Irwin, D., Cecchet, E., Shenoy, P., & Albrecht, J. (2012). Smart*: An open data set and tools for enabling research in sustainable homes. In *2012 workshop on data mining applications in sustainability (sustkdd 2012).*

Bartram, L., Rodgers, J., & Woodbury, R. (2011). Smart homes or smart occupants? supporting aware living in the home. *Human-Computer Interaction–INTERACT 2011*, 52–64.

Bartram, L., & Woodbury, R. (2011). Smart homes or smart occupants? reframing computational design models for the green home. In *2011 aaai spring symposium series.*

Basir, O., & Yuan, X. (2007). Engine fault diagnosis based on multi-sensor information fusion using dempster-shafer evidence theory. *Information Fusion*, *8*(4), 379–386.

Berenguer, M., Giordani, M., Giraud-By, F., & Noury, N. (2008). Automatic detection of activities of daily living from detecting and classifying electrical events on the residential power line. In *e-health networking, applications and services, 2008. healthcom 2008. 10th international conference on* (pp. 29–32).

Berges, M. E., Goldman, E., Matthews, H., & Soibelman, L. (2009). Learning systems for electric consumption of buildings. In *Asci international workshop on computing in civil engineering.*

Berges, M. E., Goldman, E., Matthews, H. S., & Soibelman, L. (2010). Enhancing electricity audits in residential buildings with nonintrusive load monitoring. *Journal of Industrial Ecology*, *14*(5), 844–858.

Bin, S., & Dowlatabadi, H. (2005). Consumer lifestyle approach to US energy use and the related CO2 emissions. *Energy Policy*, *33*(2), 197–208.

Blevis, E. (2007). Sustainable interaction design: invention & disposal, renewal & reuse. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 503–512).

Brush, A., Lee, B., Mahajan, R., Agarwal, S., Saroiu, S., & Dixon, C. (2011). Home automation in the wild: Challenges and opportunities. In *Acm chi*.

Cavallo, J., & Mapp, J. (2000). Monitoring refrigerator energy usage. *Home Energy Magazine*, *17*, 32–36.

Chan, W., So, A., & Lai, L. (2000). Harmonics load signature recognition by wavelets transforms. In *Electric utility deregulation and restructuring and power technologies, 2000. proceedings. drpt 2000. international conference on* (pp. 666–671).

Chang, H.-H., Lin, C.-L., & Lee, J.-K. (2010). Load identification in nonintrusive load monitoring using steady-state and turn-on transient energy algorithms. In *Computer supported cooperative work in design (cscwd), 2010 14th international conference on* (p. 27 -32).

Chetty, M., Tran, D., & Grinter, R. (2008). Getting to green: understanding resource consumption in the home. In *Proceedings of the 10th international conference on ubiquitous computing* (pp. 242–251).

Clevenger, C. M., & Haymaker, J. (2006). The impact of the building occupant on energy simulations. *Joint International Conference on Computing and Decision Making in Civil and Building Engineering, Montreal, Canada*, 1–10.

Cole, R., & Brown, Z. (2009). Reconciling human and automated intelligence in the provision of occupant comfort. *Intelligent Buildings International*, *1*(1), 39–55.

Darby, S. (2006). The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, *486*.

DiSalvo, C., Sengers, P., & Brynjarsdóttir, H. (2010). Mapping the landscape of sustainable hci. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 1975–1984).

Eckl, R., & MacWilliams, A. (2009). Smart home challenges and approaches to solve them: A practical industrial perspective. *Intelligent Interactive Assistance and Mobile Multimedia Computing*, 119–130.

Ehrhardt-Martinez, K., Donnelly, K., & Laitner, S. (2010). Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities. *American Council for an Energy-Efficient Economy*.

Electric Power Research Institute. (2009). Residential electricity use feedback: A research synthesis and economic framework. *Retrieved October*, *26*, 1–126.

Farhangi, H. (2010). The path of the smart grid. *Power and Energy Magazine, IEEE*, *8*(1), 18–28.

Figueiredo, M., de Almeida, A., & Ribeiro, B. (2011). An experimental study on electrical signature identification of non-intrusive load monitoring (nilm) systems. *Adaptive and Natural Computing Algorithms*, 31–40.

Figueiredo, M., de Almeida, A., & Ribeiro, B. (2012). Home electrical signal disaggregation for non-intrusive load monitoring (nilm) systems. *Neurocomputing*(0), -. (in press)

Figueiredo, M., de Almeida, A., Ribeiro, B., & Martins, A. (2010). Extracting features from an electrical signal of a non-intrusive load monitoring system. *Intelligent Data Engineering and Automated Learning–IDEAL 2010*, 210–217.

Fisera, R., & Macek, K. (2011). Virtual sub-metering via combined classifiers. In *Intelligent data acquisition and advanced computing systems (idaacs), 2011 ieee 6th international conference on* (Vol. 1, p. 126 -131).

Fogg, B. (1999). Persuasive technologies. *Communications of the ACM*, *42*(5), 27–29.

Froehlich, J., Findlater, L., & Landay, J. (2010). The design of eco-feedback technology. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 1999–2008).

Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M., & Patel, S. (2011). Disaggregated end-use energy sensing for the smart grid. *Pervasive Computing, IEEE*, *10*(1), 28 -39.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *IJPRAI*, *15*(1), 9–42.

Ghahramani, Z., & Jordan, M. (1997). Factorial hidden markov models. *Machine learning*, *29*(2), 245–273.

Goldstein, R., Tessier, A., & Khan, A. (2010). Customizing the behavior of interacting occupants using personas.

Granderson, J. (2009). Preliminary findings from an analysis of building energy information system technologies. *Lawrence Berkeley National Laboratory*.

Gupta, S., Reynolds, M., & Patel, S. (2010). Electrisense: single-point sensing using emi for electrical event detection and classification in the home. In *Proceedings of the 12th acm international conference on ubiquitous computing* (pp. 139–148).

Hart, G. (1992). Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, *80*(12), 1870–1891.

He, H., Greenberg, S., & Huang, E. (2010). One size does not fit all: applying the transtheoretical model to energy feedback technology design. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 927–936).

Horn, M., Davis, P., Hubbard, A., Keifert, D., Leong, Z., & Olson, I. (2011). Learning sustainability: families, learning, and next-generation eco-feedback technology. In *Proceedings of the 10th international conference on interaction design and children* (pp. 161–164).

Hutzler, G., & Gortais, B. (2004). From computer art to ambient displays. *Machine Graphics and Vision*, *13*(1/2), 181–191.

Hutzler, G., Gortais, B., & Drogoul, A. (2000). The garden of chances: a visual ecosystem. *Leonardo*, *33*(2), 101–106.

Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., & Rockinson, R. (2006). Using a live-in laboratory for ubiquitous computing research. *Pervasive Computing*, *2006*(4), 349–365.

Jiang, L., Li, J., Luo, S., Jin, J., & West, S. (2011). Literature review of power disaggregation. In *Modelling, identification and control (icmic), proceedings of 2011 international conference on* (p. 38 -42).

Kashani, M. H., & Bartram, L. (2011). *Lifestyle factors and energy conservation: A report prepared for bc hydro power smart* (Tech. Rep.). Simon Fraser University.

Kim, H., Marwah, M., Arlitt, M., Lyon, G., & Han, J. (2010). Unsupervised disaggregation of low frequency power measurements. In *11th international conference on data mining* (pp. 747–758).

Kim, T., Hong, H., & Magerko, B. (2010). Design requirements for ambient display that supports sustainable lifestyle. In *Proceedings of the 8th acm conference on designing interactive systems* (pp. 103–112).

Kitching, H., Abbott, R., & Hadden, S. (1989). *Requirements for an advanced utility load monitoring system* (Tech. Rep.). Electric Power Research Inst., Palo Alto, CA (USA); New England Power Service Co., Westborough, MA (USA); Plexus Research, Inc., Acton, MA (USA).

Kolter, J., Batra, S., & Ng, A. (2010). Energy disaggregation via discriminative sparse coding. In *Proc. neural information processing systems.*

Kolter, J., & Jaakkola, T. (2012). Approximate inference in additive factorial hmms with application to energy disaggregation. *Journal of Machine Learning Research - Proceedings Track*, *22*, 1472-1482.

Kolter, J., & Johnson, M. (2011). Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (sigkdd), san diego, ca.*

Laughman, C., Lee, K., Cox, R., Shaw, S., Leeb, S., Norford, L., & Armstrong, P. (2003). Power signature analysis. *Power and Energy Magazine, IEEE*, *1*(2), 56 - 63.

Leaman, A., & Bordass, B. (2001). Assessing building performance in use 4: the probe occupant surveys and their implications. *Building Research & Information*, *29*(2), 129–143.

Lee, K., Leeb, S., Norford, L., Armstrong, P., Holloway, J., & Shaw, S. (2005). Estimation of variable-speed-drive power consumption from harmonic content. *Energy Conversion, IEEE Transactions on*, *20*(3), 566 - 574.

Lee, S., Lin, G., Jih, W., & Hsu, J. (2010). Appliance recognition and unattended appliance detection for energy conservation. In *Workshops at the twenty-fourth aaai conference on artificial intelligence.*

Leeb, S., Shaw, S., & Kirtley, J., J.L. (1995). Transient event detection in spectral envelope estimates for nonintrusive load monitoring. *Power Delivery, IEEE Transactions on*, *10*(3), 1200 -1210.

Li, J., Luo, S., & Jin, J. (2010). Sensor data fusion for accurate cloud presence prediction using dempster-shafer evidence theory. *Sensors*, *10*(10), 9384–9396.

Liang, J., Ng, S., Kendall, G., & Cheng, J. (2010a). Load signature study part i: Basic concept, structure, and methodology. *Power Delivery, IEEE Transactions on*, *25*(2), 551 -560.

Liang, J., Ng, S., Kendall, G., & Cheng, J. (2010b). Load signature study part ii: Disaggregation framework, simulation, and applications. *Power Delivery, IEEE Transactions on*, *25*(2), 561 -569.

Luo, D., Norford, L., Leeb, S., & Shaw, S. (2002). Monitoring hvac equipment electrical loads from a centralized location- methods and field test results. *ASHRAE Transactions*, *108*(1), 841–857.

Makonin, S., Bartram, L., & Popowich, F. (in press). Redefining the "smart" in smart home: Case studies of ambient intelligence. *IEEE Pervasive Computing*.

Makonin, S., Kashani, M., & Bartram, L. (2012). The Affect of Lifestyle Factors on Eco-Visualization Design. In *Computer graphics international (cgi)* (pp. 1–10).

Makonin, S., Pasquier, P., & Bartram, L. (2011). Elements of Consumption: an abstract visualization of household consumption. *Smart Graphics*, 194–198.

Makonin, S., & Popowich, F. (2011). An intelligent agent for determining home occupancy using power monitors and light sensors. *Toward Useful Services for Elderly and People with Disabilities*, 236–240.

Makonin, S., & Popowich, F. (2012). Home Occupancy Agent: Occupancy and Sleep Detection. *GSTF Journal on Computing*, *2*(1), 182–186.

Marsland, S. (2009). *Machine learning: an algorithmic perspective*. Chapman & Hall/CRC.

Metz, C. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine* (Vol. 8, pp. 283–298).

Motegi, N., Piette, M., Kinney, S., & Dewey, J. (2003). Case studies of energy information systems and related technology: operational practices, costs, and benefits. *Lawrence Berkeley National Laboratory*.

Norford, L. K., & Leeb, S. B. (1996). Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms. *Energy and Buildings*, *24*(1), 51 - 64.

Noy, P., Liu, K., Clements-Croome, D., & Qiao, B. (2006). Design issues in personalizing intelligent buildings. In *Intelligent environments, 2006. ie 06. 2nd iet international conference on* (Vol. 1, pp. 143–149).

Ortiz, A., Lehtonen, M., Mañana, M., Renedo, C., Muranen, S., & Eguíluz, L. (2007). Evaluation of energy meters' accuracy based on a power quality test platform. *Electric Power Components and Systems*, *35*(2), 221–237.

Parker, D., Hoak, D., Meier, A., & Brown, R. (2006). How much energy are we using? potential of residential energy demand feedback devices. *Florida Solar Energy Center, American Council for an Energy Efficient Economy. Ansilomar, Ca.(Sep 28, 2010)*.

Parson, O., Ghosh, S., Weal, M., & Rogers, A. (2012). Non-intrusive load monitoring using prior models of general appliance types. In *Twenty-sixth conference on artificial intelligence (aaai-12)*.

Patel, S., Robertson, T., Kientz, J., Reynolds, M., & Abowd, G. (2007). At the flick of a switch: Detecting and classifying unique electrical events on the residential power line. In *Proceedings of the 9th international conference on ubiquitous computing* (pp. 271–288).

Pereira, L. (2004). Cascade to black [system blackouts]. *Power and Energy Magazine, IEEE*, *2*(3), 54–57.

Pew Oceans Commission (Arlington, Virginia). (2003). *America's living oceans: Charting a course for sea change: A report to the nation: Recommendations for a new ocean policy*. Pew Oceans Commission.

Pierce, J., Odom, W., & Blevis, E. (2008). Energy aware dwelling: a critical survey of interaction design for eco-visualizations. In *Proceedings of the 20th australasian conference on computer-human interaction: Designing for habitus and habitat* (pp. 1–8). New York, NY, USA: ACM.

Quintal, F., Pereira, L., & Nunes, N. (2012). A long-term study of energy eco-feedback using non-intrusive load monitoring. *Persuasive Technology*, 49–52.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, *3*(1), 4 -16.

Reinhardt, A., Baumann, P., Burgstahler, D., Hollick, M., Chonov, H., Werner, M., & Steinmetz, R. (2012). On the accuracy of appliance identification based on distributed load metering data. In *2nd ifip conference on sustainable internet and ict for sustainability (sustainit)* (pp. 1–9).

Rodgers, J., & Bartram, L. (2010). Visualizing residential resource use: A framework for design. In *Infovis.*

Rodgers, J., & Bartram, L. (2011). Exploring ambient and artistic visualization for residential energy use feedback. *Visualization and Computer Graphics, IEEE Transactions on*, *17*(12), 2489 -2497.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In A. Sattar & B.-h. Kang (Eds.), *Ai 2006: Advances in artificial intelligence* (Vol. 4304, p. 1015-1021). Springer Berlin / Heidelberg.

Strengers, Y. A. (2011). Designing eco-feedback systems for everyday life. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 2135–2144). New York, NY, USA: ACM.

Tapia, E., Intille, S., & Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, 158–175.

Torning, K., & Oinas-Kukkonen, H. (2009). Persuasive system design: state of the art and future directions. In *Proceedings of the 4th international conference on persuasive technology* (p. 30).

Tsai, M., & Lin, Y. (2011). Modern development of an adaptive non-intrusive appliance load monitoring system in electricity energy conservation. *Applied Energy*.

Tse, N., Zhou, L., & Lai, L. (2008). Wavelet-based algorithm for nonstationary power system, waveform analysis. In *Wavelet analysis and pattern recognition, 2008. icwapr '08. international conference on* (Vol. 2, p. 729 -735).

Velikov, K., & Bartram, L. (2009). North House: Developing intelligent building technology and user interface in energy independent domestic environments. In *The 26th conference on passive and low energy architecture (plea2009).* Quebec City, Canada, 22-24 June.

Wichakool, W., Avestruz, A.-T., Cox, R., & Leeb, S. (2009). Modeling and estimating current harmonics of variable electronic loads. *Power Electronics, IEEE Transactions on*, *24*(12), 2803 -2811.

Woods, D. (1996). Decomposing automation: Apparent simplicity, real complexity. *Automation and human performance: Theory and applications*, 3–17.

Zeifman, M. (2012). Disaggregation of home energy display data using probabilistic approach. *Consumer Electronics, IEEE Transactions on*, *58*(1), 23 -31.

Zeifman, M., & Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *Consumer Electronics, IEEE Transactions on*, *57*(1), 76–84.

ZigBee Alliance. (2011). Zigbee smart energy profile specification. *Zigbee Doc. 075356r16ZB, rev 16*.