User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings

Mario Berges, A.M.ASCE[1]; Ethan Goldman[2]; H. Scott Matthews, A.M.ASCE[3]; Lucio Soibelman, M.ASCE[4]; and Kyle Anderson[5]

Abstract: This paper presents a nonintrusive electricity load-monitoring approach that provides feedback on the energy consumption and operational schedule of electrical appliances in a residential building. This approach utilizes simple algorithms for detecting and classifying electrical events on the basis of voltage and current measurements obtained at the main circuit panel of the home. To address the necessary training and calibration, this approach is designed around the end-user and relies on user input to continuously improve its performance. The algorithms and the user interaction processes are described in detail. Three data sets were collected with a prototype system (from a power strip in a laboratory, a house, and an apartment unit) to test the performance of the algorithms. The event detector achieved true positive and false positive rates of 94 and 0.26%, respectively. When combined with the classification task, the overall accuracy (correctly detected and classified events) was 82%. The advantages and limitations of this work are discussed, and possible future research is presented.
DOI: 10.1061/(ASCE)CP.1943-5487.0000108.

CE Database subject headings: Energy consumption; Monitoring; Signal processing; Residential buildings.

Author keywords: Nonintrusive load monitoring; Energy management; Signal processing; Demand-side management; Electrical load signatures; Power systems.

Introduction

Global climate change, record high oil prices, and the decreasing availability of fossil fuels are forcing all of us to rethink the way we deal with our energy needs. Out of the 99 quadrillion BTU (quads) of total annual primary energy consumption in the United States in 2008, 40% was used to generate electricity. Most of this energy, 68%, is lost during the processes of generation and distribution. Of the portion that reaches the end-user, 36% is consumed by commercial buildings, 37% by the residential sector, and the remaining 27% is used, primarily, in manufacturing (Lawrence Livermore National Laboratory 2009). Reducing the consumption in residential and/or commercial facilities would thus have a significant effect on the total energy savings for the country.

Despite improvements in efficiency, such as green buildings, this consumption is expected to continue increasing, largely because of population growth and an increasing density of electricity-using devices in modern buildings. Policies such as new building codes, utility energy-efficiency programs, and appliance standards have had some success in reducing these growing demands. When combined with demand-side management programs, these approaches have the potential to reduce electrical usage by as much as 30% over the next decade (Committee on America's Energy Future; National Academy of Sciences; National Academy of Engineering; National Research Council 2009). Technological solutions for simplifying and automating demand-side management strategies are also a crucial part of the solution.

To effectively identify opportunities for consumption reduction, measurement and feedback on current energy usage is necessary. Utilities provide their customers with monthly reports of their consumption, but automated meter reading (AMR) systems are starting to provide more frequent updates. For example, Duquesne Light, one of the main electricity suppliers for the city of Pittsburgh, provides daily averages through a web interface.

Monthly utility bills are inadequate for planning conservation programs or even for assessing the impact of such programs once implemented. Furthermore, as the construction industry slowly transitions toward green buildings, and the number of Leadership in Energy and Environmental Design (LEED)–certified facilities continues to increase, the question of how to evaluate as-built energy performance becomes more important because the criteria were devised primarily for the design and construction phases.

The behavioral impacts of providing users with real-time energy use feedback, even at the aggregate level (e.g., overall consumption of the building), has been shown through limited studies to produce savings of up to 10–15% (Parker et al. 2006; Fischer 2008; Darby 2006). However, there exists evidence that this effect may not be long-lasting (Peschiera et al. 2010), and further studies need to be conducted to properly assess this. Despite this, we believe that

[1]Assistant Professor, Civil and Environmental Engineering Dept., Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213 (corresponding author). E-mail: mberges@andrew.cmu.edu

[2]Measurement and Verification Specialist, Vermont Energy Investment Corporation, 128 Lakeside Ave., Suite 401, Burlington, VT 05401. E-mail: egoldman@veic.org

[3]Professor, Civil and Environmental Engineering Dept., Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: hsm@andrew.cmu.edu

[4]Professor, Civil and Environmental Engineering Dept., Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: lucio@andrew.cmu.edu

[5]Ph.D. Candidate, Electrical and Computer Engineering Dept., Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: kda@andrew.cmu.edu

Note. This manuscript was submitted on June 30, 2010; approved on February 1, 2011; published online on May 6, 2011. Discussion period open until April 1, 2012; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, Vol. 25, No. 6, November 1, 2011. ©ASCE, ISSN 0887-3801/2011/6-471–480/$25.00.

larger and longer-term savings are potentially achievable if more detailed data were available not only to the user but also to automated building control systems or to electricity suppliers (allowing them to reward peak-shifting loads or subsidize equipment upgrades, for example).

A system fed with disaggregated data, detailing the energy used by each appliance in a building, could illustrate the impact of an air conditioning unit on a user's total energy expenditures and suggest possible changes along with the expected savings. This type of information would benefit numerous applications: building control and automation, load scheduling and optimization, peak shifting, and more. All these scenarios require more detailed consumption data than what is currently available from monthly bills alone. The industry is only now starting to realize the potential of this information.

In principle, one way to obtain detailed appliance usage information would be through extensive hardware submetering. In such a system, each electrical load in a building would be individually metered (or at least metered by circuit). However, such an approach would have a high price because of the hardware and installation costs.

Additionally, there exists a growing number of commercially available power meters that provide traditional power metrics in real time at varying degrees of detail: the building's main electrical feed, circuit panels, individual circuits, or individual outlets. These latter solutions range from hundreds to tens of thousands of dollars, their cost increasing with the level of detail obtained (Berges et al. 2010c).

Context

Given the relationship that exists between the granularity of electricity consumption information and the cost to obtain it, any proposed solution will necessarily consist of some combination of technology, granularity, and uncertainty. In the average residence in the United States, just 12 types of appliances account for 80% of electricity consumption (Energy Information Administration 2001). Therefore, if we were to focus installation on those that are shown to be the biggest energy consumers instead of using the most detailed measurement tools (i.e., plug-level power meters) on all appliances throughout the building, we could then start to make intelligent compromises about the cost and uncertainty.

Another possible solution that has been explored in the past, and is the main topic of this paper, is to install power meters at higher aggregation points in the building's power distribution system (e.g., the main feed for a residential unit) and extract detailed information by carefully processing it, taking advantage of the fact that different electrical loads may have a characteristic way of drawing current. By sampling voltage and current at this level, at a high enough rate, specialized signal processing and machine learning algorithms can be used to classify distinctive characteristics of the measured signals that are associated with the operation of individual appliances. Because it relies mostly on software, this disaggregation solution, known as nonintrusive load monitoring (NILM), has a lower cost than hardware-based submetering and a high level of detail.

Problem Description

NILM has been a topic of research for more than 20 years. In 1997 the Electric Power Research Institute (EPRI) in California published a technical report (EPRI 1997a) on the market feasibility of the NILM technology of the time from the electric utility's perspective. At that time, the report concluded that costs were the largest obstacle for the technology to achieve mass-market penetration. The accompanying technical assessment report (EPRI 1997b) discusses some methods to resolve the issues. However, despite the fact that this was more than a decade ago and hardware costs have fallen, there exists still no important penetration of this technology into the market. In 2003, a report prepared for the California Energy Commission's Public Interest Energy Research (PIER) program concluded that "there are issues regarding identification of multiple units of devices that are of the same make and model within a facility or on a branch circuit" (Smith et al. 2003). It also stated that the commercial value of NILM would still need to be determined. One of the sections of this report (Lee 2003) detailed all the improvements made to the technology up to that point and supported the technical recommendations for future improvements. Among the latter, they mentioned the need to automate the training process "to the extent possible."

This paper presents a user-centered NILM approach that introduces the homeowners or facility managers into an online, continuous training process. We discuss some initial results and relevant findings that we have obtained from our implementation of the algorithms in a prototype system deployed in a laboratory setup and in two occupied single-family apartment units. We also discuss the problems with obtaining ground-truth data along with our solution for this and some comments about the issues that arise when deploying these systems in the real world. The paper concludes with a discussion of the experimental findings and future research questions.

Background and Literature Review

Before exploring the literature and deciding to focus on NILM, we investigated the available commercial solutions on the market. We found a wide range of products (Berges et al. 2010a, c; Matthews et al. 2008), but very few addressed our specific problem. A variety of plug-through power meters are available, designed to monitor the consumption of individual appliances. However, within this group, only a small portion provide a communication link that enables the users to concurrently monitor all the meters in the building from a central location. On the other end of the spectrum, numerous power meters were designed to measure the building's total consumption at the main feed. They use a variety of different methods to obtain their readings; some rely on the utility's power meter, others attach current transducers to the main electrical lines. However, these solutions had too low a reporting rate for our purposes, providing an updated power measurement no more than once per second.

Some intermediate solutions measure individual electrical circuits in the building. We have installed one such system for two electrical panels in a building at Carnegie Mellon University. Although the acquired data had a higher level of detail than what would be obtained from measuring the total building consumption, the hardware and labor costs far exceed the value of this information. We needed to trace all circuits to confirm the loads that were served, and even after doing so, we did not completely eliminate uncertainty. One of the reasons for this is that some circuits feed a variety of different, sometimes nonfixed loads (Berges et al. 2010a).

Commercial products meeting our needs were virtually nonexistent. We did find, however, a company that manufactures power meters that implement NILM algorithms. This company, Enetics, leverages the results of research by George Hart (Hart 1992), the pioneer of NILM. However, the products that they offer are geared toward utilities, not consumers, and their prices are still cost-prohibitive for a residential buyer. Other companies like Intel (Engadget 2010) and Belkin (MIT Technology Review 2010) have recently announced a new line of products using NILM techniques, but the effectiveness of these products remains to be tested. The details of the approaches used by these companies are not publicly available.

NILM derives its name from the fact that, from the perspective of the electric utility company, the technique is able to monitor individual loads in a building without intruding (e.g., placing sensors or other devices) into the customer's property. The term is commonly used, however, to refer to single-point-of-measurement techniques, regardless of whether the measurements are taken inside (i.e., the circuit panel) or outside the property. The algorithms, in high-level terms, can be described in five steps:

- Data acquisition: Signals indicative of the overall voltage and/or current of the building are acquired;
- Event detection: Changes in certain metrics obtained from these signals are detected and flagged as events for later identification; it is assumed that these changes are caused by the operation of individual appliances (e.g., water heater turning on);
- Feature extraction: Certain properties of the samples surrounding the detected event are used to describe it (e.g., the size of the step change of the total power);
- Classification: Flagged events are processed by using a statistical method, usually a supervised learning algorithm, which assigns a label to them (e.g., television turning off) on the basis of a trained model or set of labeled instances; and
- Energy computation: By keeping track of these transitions and their associated power level, the specific consumption of each detected appliance can be estimated by using different techniques.

George Hart (Hart 1989, 1992) was one of the first researchers to publish in the area. His early publications describe a method for utilizing normalized real and reactive power (P and Q, respectively) measurements from the main electrical feed of a residential building. His technique relied on steady-state power metrics (i.e., disregarding any transient, nonstable state) to describe in a distinct way the power draw of most home appliances of the time. In other words, when an individual appliance changed its state from off to on, the change in the total real and reactive power of the house would be almost unique to the mentioned appliance. Hart referred to these changes as the appliance's signature and described methods for correcting possible overlaps in this signature space by making use of appliance-state transition models (e.g., an appliance cannot go from off to on and then again to on).

Norford and Leeb improved on Hart's technique by analyzing the start-up transients of appliances (Norford and Leeb 1996) and introducing better algorithms for detecting when state transitions have occurred (Luo et al. 2002). In Laughman et al. (2003), investigators describe how the use of current harmonics can improve the process even further, allowing for the detection and classification of certain continuously variable loads. Moreover, Wichakool et al. (2007) presents further improvements to the solution for the problem of variable power electronics by using a spectral estimation method and a switching function technique. A summary and presentation of the latest achievements in this line of work can be found in Shaw et al. (2008).

Other research has focused on utilizing the technique for monitoring the health of large appliances, by carefully analyzing any changes to their start-up transient and associated signature (Paris 2006; Cox et al. 2006; Norford and Leeb 1996). Efforts have also been made toward eliminating the need to collect current readings by inferring these from pure voltage measurements (Cox et al. 2006), whereas others have focused on methods that do not require an appliance to change from one state to the other but rather detect the presence of an appliance while it is used (Srinivasan et al. 2006).

There exists also a growing number of research projects that have explored different classification algorithms and feature extraction methods. Neural networks have been used by Prudenzi (2002), and more recently by Chang et al. (2008). Genetic algorithms and clustering approaches were applied by Baranski and Voss (2004) to data acquired from utility meters by using an optical sensor. A rule-based system was developed by Farinaccio and Zmeureanu (1999) to solve the disaggregation problem. An attempt to create a general taxonomy for appliance signatures is presented in Lam et al. (2007), in which by using clustering techniques and a novel feature set, the researchers found common traits in the signatures of same-type appliances present in modern residential buildings.

As previously discussed, despite almost two decades of research in the area, techniques for nonintrusively disaggregating the total electrical load of buildings remain in the hands of researchers and have not yet been adopted by society, in general. We argue that, to achieve wide adoption, the solutions need to be simple, easy to install, inexpensive, and be able to return the investment in a reasonable time.

The problem of appliance disaggregation sought by proponents of NILM needs to be reframed from a systems integration perspective, and the focus should not be limited to the hardware and software requirements for analyzing voltage and current waveforms. We propose to engage the user in the regular operation of NILM systems to continuously train the algorithms providing corrections and labels for new signatures. Additionally, although not explored in this paper, we also believe that integrating other sensor data, other than power measurements, such as audio levels, light intensity, and occupancy, would help overcome many of the obstacles that the technology faces today.

The following section explains our approach for using aggregate power readings to estimate the operational schedule and energy consumption of appliances in the building.

System Overview

Ultimately, our goal is to track each appliance in the house by using as few measurement points as possible to reduce installation costs. Fig. 1 shows the flowchart of how our proposed NILM system operates; a brief overview is given here and individual steps are subsequently explained in depth. Following the arrows and starting from the data source, the first step is to compute different power metrics from the raw data coming from the different sensors (voltage and current) to generate a useful and coherent data stream. After this, the data are passed on to a load-disaggregation process that has three steps. It begins with the event detector in which regions in the signals that reflect appliance-state transitions are detected. Following that, features are extracted from a set of samples surrounding the detected event to characterize it. Lastly, the load-disaggregation process finishes with a classifier that uses machine learning techniques to automatically classify the detected events as described by the features extracted in the previous stage. A signature database provides the training data for the classifier. Unrecognized signatures, either because they are novel to the system or because of poor training, can be labeled by an external agent (human user, submetering,) and inserted into the signature database for future reference.

Continuing to follow the diagram in Fig. 1, once the event has been classified, the output, an appliance-state transition (e.g., refrigerator went from off to on), is fed simultaneously to an appliance-state model and the interested parties. The appliance-state model keeps track of the operation of appliances and also communicates with the load disaggregator to prevent misclassifications

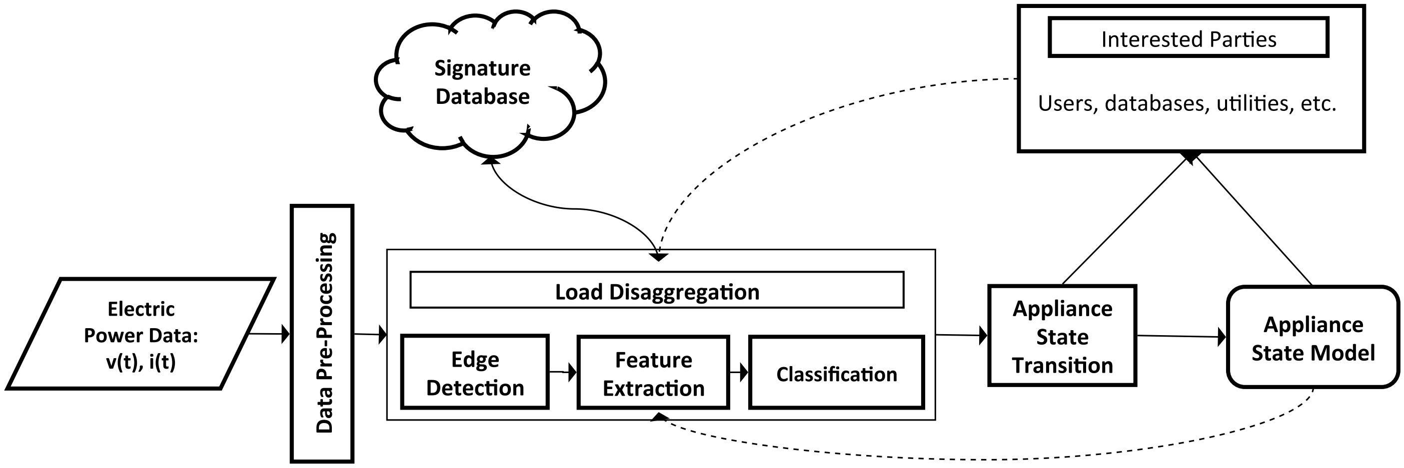

Fig. 1. Block diagram illustrating the different steps involved in our basic NILM implementation

(e.g., a single appliance cannot go from off to on two consecutive times). Both the appliance-state model (historical) and the last recognized appliance-state transition (current) are made available to the interested parties: human users who occupy, manage, or operate the building, building control systems, electric utilities.

The following subsections describe each of the steps in more detail. Given that we are building on top of previous knowledge, many of the steps (preprocessing and load disaggregation) represent only variations on techniques already found in the literature. The main contribution in this section will be related to supporting an online and distributed training process.

Hardware Setup and Data Acquisition

Residential buildings in the United States generally utilize a single phase (sometimes referred to as split-phase) distribution system, in which two alternating current (AC) voltage sources, phase offset by half a cycle, are used to feed two separate sets of circuits in the building. These sources, often referred to as Phase A and B, have a fundamental frequency of 60 Hz in the United States. It is at this level of aggregation, the main sources feeding an electrical panel in a building, in which we carry out our analysis and in which we believe that NILM-like algorithms can provide the best balance between cost and the value of the resulting information.

We install voltage and current transformers at both Phase A and Phase B inside the main circuit panel of a residence (the installation could alternatively be applied outside of the house). These analog signals are fed into a data-acquisition card connected to a computer that digitizes and samples them. Sometimes we only measured the voltage on one of the phases and assumed that the other phase could be represented by shifting the measured voltage by a half-cycle. To accommodate the processing steps in disaggregating the loads, it is necessary to choose a sampling rate that is sufficiently high (e.g., greater than 1 kHz) to capture harmonic content in both signals and nonlinear current waveforms that may occur in modern buildings because of the increasing presence of power supplies for electronics (which include rectifiers) and fluorescent lamps.

Power Calculations

At this point we have access to sampled versions of the raw voltage $v(t)$ and current $i(t)$ waveforms. From these signals, we calculate various power metrics that are used in the event detection and classification blocks in Fig. 1.

Both signals, especially the current, contain valuable information higher in the frequency spectrum. The low-order, odd-numbered current harmonics (i.e., 180 and 300 Hz) are especially important (Lee 2003). Because of nonlinear circuit elements, certain appliances increase the amplitude of those frequency components during operation. For these reasons, we compute power metrics from the raw voltage and current signals that preserve this information. These metrics are known in the literature as spectral envelope coefficients and are subsequently defined (Shaw et al. 2008; Leeb et al. 1995).

$$P_k(t) = \frac{2}{T}\int_{t-T}^{T} i(\tau)\cos(k\omega\tau)d\tau \tag{1}$$

$$Q_k(t) = \frac{2}{T}\int_{t-T}^{T} i(\tau)\sin(k\omega\tau)d\tau \tag{2}$$

in which k = harmonic index; and T can be one or more periods of the fundamental frequency of the voltage signal; P and Q = analogous to the conventional definitions of real power and reactive power, respectively, when $k = 1$.

Eqs. (3) and (4) are defined for continuous time signals, but we are working with sampled versions of these continuous voltage and current signals. In this case, we refer to the method presented by Lee in Chapter 2 of his work (Lee 2003) for computing the spectral envelope coefficients, also referred to in the literature as current harmonic powers. Lee gives an efficient computation of these quantities by performing a fast Fourier transform (FFT) on the sampled current signal.

Event Detection

Once the power has been computed from the current and voltage signals, the load-disaggregation process begins with event detection. To be able to classify state transitions, we must first be able to recognize when they occur. The purpose of the event detector is to extract the time instant that an appliance changes states. When an appliance changes its state (e.g., turns on or off), its change in power consumption is seen in the aggregate power signal. In the simplest terms, the event detector is looking for times when significant changes in power consumption occur.

A naïve approach to event detection on the basis of power consumption change would be to compare adjacent samples of the power signal and flag an event when the power change deviates beyond a fixed threshold. For example, we have used changes in real power as indicators of the state transitions. For instance, a change of 5 W between one sample and the following one can be an indication of a state transition for one or more appliances, signal noise, or part of the normal operation of a given appliance without it being a state transition. For these reasons, a probabilistic model was used to detect such changes. More specifically, we used a modified version of the generalized likelihood ratio (GLR) presented in Luo et al. (2002).

The main differences in this version of the GLR algorithm are that (1) to reduce the number of parameters that need to be set, instead of assuming fixed values for the standard deviation, we continuously compute this metric from the samples; and (2) we implement a voting scheme on top of the output of the maximization of the detection statistic.

In thid algorithm, two sliding windows are utilized for detection; both are of fixed length. The larger window is used for the voting procedure, and it is referred to as the event-detection window w^e. The smaller window, which slides inside the larger one, is used in calculating the likelihood ratio l_n for each point and is called the likelihood ratio window w^l. The test statistic s_n used for determining which point receives a vote is the cumulative sum of likelihood ratios from the point in question to the last point of the event-detection window.

With these quantities established, the algorithm works as follows. For each point in the event-detection window, both l_n and in turn s_n are calculated. If, in the larger event-detection window, the point's test statistic is the maximum among all points, it receives a vote. The event-detection window then slides one sample, and the point with the highest test statistic in the new window receives a vote. This is how a single point can receive multiple votes, namely, if it has the highest test statistic for multiple increments of the event-detection window. Every point that receives a number of votes greater than a predetermined threshold ($V_{\min}$) is labeled as an event, and features from samples surrounding it are collected and fed to the classification algorithms (the topic of the next section).

The likelihood ratio calculated at every point is

$$l_n = \ln \frac{P(P_1[n]|\mu_{\text{after}}, \sigma_{\text{after}})}{P(P_1[n]|\mu_{\text{before}}, \sigma_{\text{before}})} \tag{3}$$

The test statistic s_n is then calculated by using the likelihood ratios from the point in question to the last point in the event-detection window, w^e

$$S_n = \sum_{j=n}^{\text{last}(w^e)} l_j \tag{4}$$

Votes are assigned as follows:

$$\text{vote}_{\text{index}} = \arg \max_{n \epsilon w^e} s_n \tag{5}$$

in which the probability distributions are assumed to be Gaussian; $\mu_{\text{after}}, \sigma_{\text{after}}$ = sample mean and variance over $[n+1, n+w_a^l+1]$; $\mu_{\text{before}}, \sigma_{\text{before}}$ = sample mean and variance over $[n-w_b^l-1, n-1]$; and w_b^l, w_a^l = number of points before and after the current point in the test statistic window. Thus, the length of the test window is $w_b^l + w_a^l + 1$.

This event-detection scheme works well for single transients that fit within the event-detection window, but it fails on longer transients and cannot distinguish overlapping transients.

Feature Extraction

Now that we have a way of detecting the events, we can attempt to automate the recognition of the appliance-state transition that most likely caused it. This requires that we first provide a way to describe or encode the relevant features about the event in question and then that we find a function that can map these features into a discrete set of labels representing all the possible appliance-state transitions of interest.

For our purposes, an appliance signature is a feature vector f containing F features extracted from the power signal in the immediate vicinity of the detected event. Examples of these are the change in real power, the change in reactive power, and the shape of the transient profile in any of the P_k or Q_k coefficients. We characterize the shapes of the transient profiles by doing a linear regression with Fourier basis functions. Features can also be extracted from other data streams (e.g., environmental sensors), as explained in Berges et al. (2010b). If the feature selection is done appropriately, then f provides a fingerprint for the event in question and can be used in machine learning classification algorithms.

The immediate vicinity of the event used to select the features is a fixed-length feature-selection window of length γ, and it is different from the event-detection window in the previous section. The feature-extraction window consists of β samples before the event, the sample at the index in which the event was detected, and α samples after the event; the total length is $\gamma = \beta + \alpha + 1$. With proper tuning of the β and α parameters, it is possible to capture most of the relevant information about the state transition in question. Cross-validation is typically used to select these parameters. Fig. 2 shows the turn-on and turn-off transients of a refrigerator using a window of $\gamma = 100$ samples, or approximately 1.5 s at 60 Hz.

The feature vector f for each of these transients (turn-on and turn-off) is of the form $f = [P_1^a, P_1^b, Q_1^a, Q_1^b]$, in which P and Q are the regression coefficients used to characterize the transient shapes in real and reactive power; the superindices indicate the source of the measurements (Phase A or B). In practice, regression coefficients from the transient in other harmonics P_k, Q_k may also be included, but they are omitted in this explanation for simplicity. Including higher harmonic content may increase the separation of the transients in the feature space, and depending on the type and number of appliances of interest, this may or may not translate into better classification results. Finally, the model order for the regression can be determined through cross-validation.

Classification and Training

Each detected event generates a transient profile, the feature vector f, like the one shown previously. When a human provides a label, the profile is added to a signature library and can then be used to classify future events of the same type automatically. The classification is performed by using standard machine learning techniques. In our experiments, a 1-nearest neighbor classifier worked best, as will be shown in the experiments section. When the system sees a new transient described by feature vector f', it is classified by finding the closest feature vector in the signature library in the feature space. That is

$$\text{class} = \arg \min_i \|f_i - f'\| \tag{6}$$

By using this general approach for our problem, it was necessary to define a way for the system to receive labeled examples (the training phase) and then select an appropriate algorithm for the classification. A number of different algorithms are discussed in a subsequent section, but the training process is described in more detail here.

The system, right after the initial deployment, is untrained; it has not been presented with any labeled examples. The first time the event detector finds an event, it immediately allows the user to provide a label for it. For example, when first deployed, the user turns on his television, and the system detects the change but is unable to classify it. The user can train the algorithm by adding the appliance to the system's database and selecting the appropriate label for the

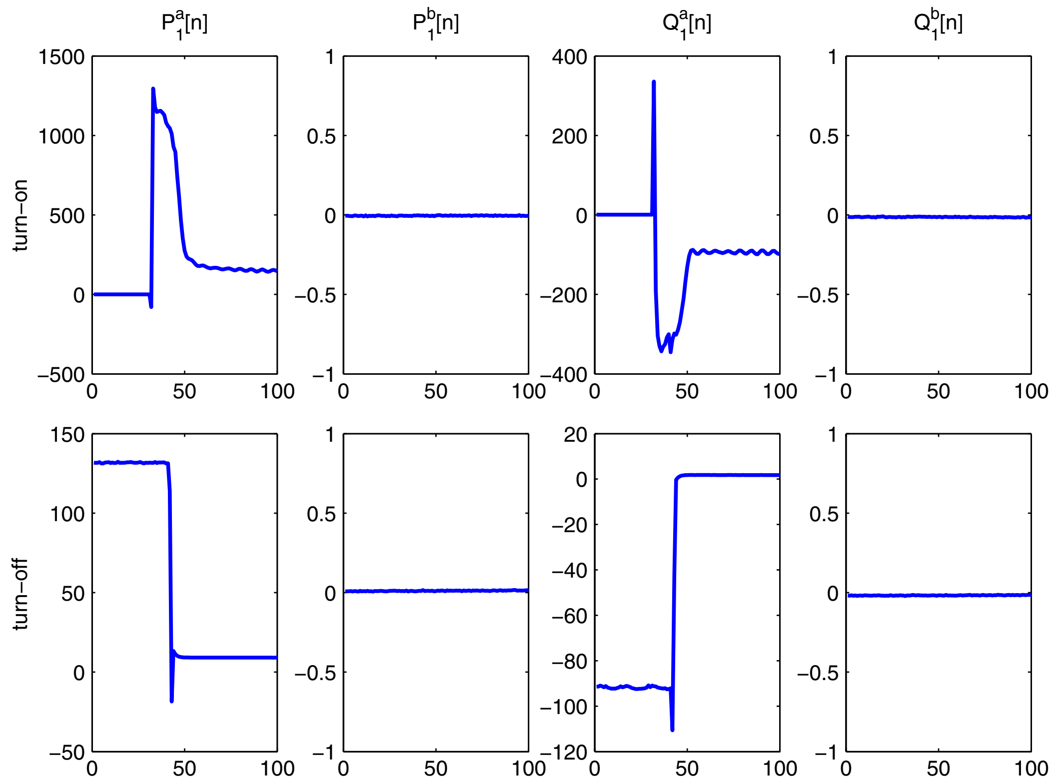

Fig. 2. Example turn-on and turn-off transients for a refrigerator; the *x*-axis represents sample index and the *y*-axis represents the value of the P and Q coefficients; the samples are separated by $1/60$ s

transient that was just detected (e.g., television from off to on) through an interface like the one shown in Fig. 3. At this point, the system saves a copy of the feature vector along with the label. The same process is repeated every time that the system has a low confidence on its prediction.

This interactive process allows the system to adapt to the changing environment (e.g., new or replacement appliances and degrading signatures). It also removes the need for having to pretrain the system or devising signatures that can generalize across appliance classes.

Description of Data Sets and Experiment Method

Three data sets were used in this paper for this exploration of the performance of the algorithms. They were obtained by monitoring different power-distribution systems using variations of the prototype system presented in Fig. 3. These were (1) a power strip in a laboratory with eight different appliances connected to it; (2) a single-family house with 17 appliances studied; and (3) a single-family apartment unit with 34 appliances under study. Table 1 shows more details about these data sets.

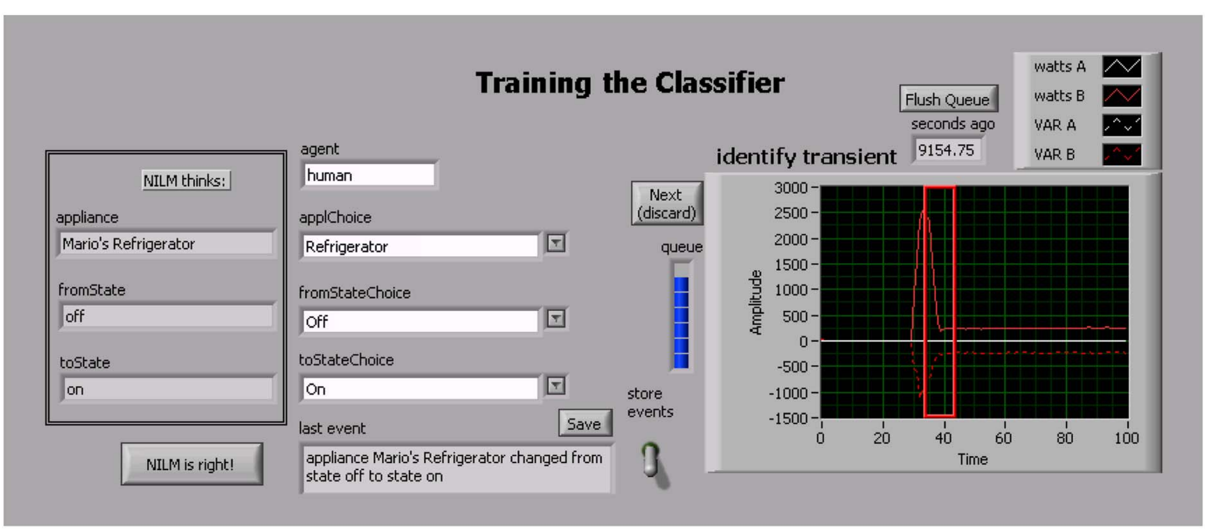

Fig. 3. Example graphical user interface (GUI) for training the classifier

Table 1. Description of Data Sets

Data set	Number of appliances	Number of state transitions	Number of events	Example appliances	Power metrics
A: Power strip	8	34	483	Radio, microwave, toaster	20 Hz real and reactive power
B: House	17	44	281	Stove, oven, kettle	20 Hz real and reactive power
C: Apartment	14	27	62	Refrigerator, lamp, television	60 Hz spectral envelope coefficients up to the 10th harmonic

To evaluate the performance of the algorithms, it was necessary to obtain the power consumption of the devices as separately monitored for the duration of the study and the times at which each appliance-state transitions occurred. However, obtaining this information (the ground truth) for this type of problem is difficult.

Nonetheless, these data sets were obtained under relatively controlled conditions by using one of two methods. In the first, one person would switch each appliance on or off (or change its state) while another person at approximately the same time would push a button on the graphical user interface of the prototype to record the time and description of the event. The second option relied on the event detector previously described to detect the state changes, which the user would then label accordingly. This last approach turned out to be more accurate and efficient, given that the time stamps for the events were more precise and one person was sufficient to complete the task.

We realize that these two approaches are highly impractical for collecting long periods of fully labeled data. Besides the obvious problem of requiring human presence during the duration of the data collection process, a less evident issue is the difficulty of labeling internal states of appliances (e.g., an electric stove cycles between on and off to maintain the temperature selected by the user) and maintaining control of some cycling loads present in the building (e.g., refrigerators and water heaters) to prevent the overlap of events or mislabeling.

Because of the limited availability of data and the short duration of the data sets (approximately 2 h on average), we carefully designed a strategy to make the best use of it during the process of exploring the parameter space for the different algorithms involved in our prototype. The three data sets were used in a series of experiments to evaluate the performance of the event detector and the classification algorithms.

To fine-tune the parameters for the event detector, we made use of Data Set A, and to evaluate the performance with the chosen parameters, we used Data Set B. For the classification task, first we collected all the transients from each data set to build separate signature libraries. These libraries were further split into training/test sets and validation sets. With the training sets, 10-fold cross-validation was performed to find appropriate settings for the algorithms, whereas the validation sets were reserved for providing unbiased estimates of the true error. Data Sets A and B were used to find the appropriate basis functions and the order of the models. Data Set C was used in the end to test the performance of both the event detector and classification algorithms.

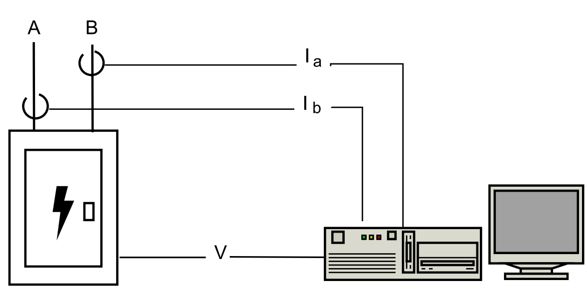

Fig. 4. Diagram for the initial prototype system; I_a and I_b are the currents on phases A and B; V represents the voltage on one of the two phases

Experimental Results

Having described the prototype system and its components and the data sets that were obtained, we present some preliminary results. The prototype system's algorithms were implemented by using LabVIEW (LabVIEW 2010) and MATLAB (MATLAB 7.9) and were running on a computer with a general-purpose data-acquisition card that was sampling voltage and current at 15 kHz, as shown in Fig. 4. Power metrics were computed at different rates depending on the data set (see Table 1).

Event-Detection Performance

The event-detection algorithm requires setting certain parameters: the number of samples in the event-detection window (w^e), the preevent and postevent window sizes (w_b^l, w_a^l), and the minimum number of votes ($V_{\min}$). Exploring the whole parameter space would be almost impossible; however, to find appropriate settings, some ranges were selected for each of the parameters based on how the algorithm works and initial tests, as shown in Table 2. Because the data sets have different sampling rates for the power metrics, these values are expressed in number of seconds. To obtain the equivalent number of samples, the reader should simply multiply by the sampling rate for the power metrics in each data set.

By using Data Set A, all possible combinations of these settings were tested to identify the one that produced the most satisfactory results on the basis of the true positive and false positive rates (TPR and FPR, respectively). Correct detections (true positives) considered were those that were within 2 s of the ground truth. For comparing the results obtained with each setting, these values were plotted in the receiver operating characteristic (ROC) space and looked for the lowest FPR and highest TPR.

Through this process, we arrived at the following values for the parameters of the event detector: (1) the detection window (w^e) was set to 160 samples, which at 20 Hz indicates 8 s; (2) the post- and preevent window sizes (w_a^l, w_b^l, respectively) were set to 40 samples (2 s); (3) the voting threshold ($V_{\min}$) was set to 10 votes. By using these settings alone, many small changes in power were detected as events. These changes may correspond to measurement noise or to actual appliance-state changes that are not of interest (e.g., computer hard drives). To eliminate these false

Table 2. Values of Parameters Used during Performance Test of Event Detector

Parameter	Values
w^e	4, 5, 6, 7, 8 seconds
w_a^l, w_b^l	1, 1.5, 2, 2.5 seconds
$V_{\min}$	10, 20, 30 votes

detections the likelihood ratio (l_n) was set to zero every time $\mu_{\text{after}} - \mu_{\text{before}} < t$, in which t was set to 30 W. After correcting this, we evaluated the performance of the detector on Data Set B and obtained the following values for TPR and FPR: 94.13% and 0.26%, respectively.

This technique works well for abrupt changes in power, but it does not detect slower transients accurately. For example, the heat pump in Data Set B has a very slow start-up that can last more than a minute. During this time, the detector produces most of the false detections. Also, if two events occur too close to one another, the stronger of the two will dominate the voting process, and the other will go undetected. For these scenarios, other detection schemes, such as the multiresolution approach described in Luo et al. (2002), are being explored.

Classification Performance

We compared the accuracy of three different machine learning algorithms: Gaussian naïve Bayes (GNB) (Mitchell 2005), decision trees (DT) (Mitchell 1997) and k-nearest neighbors (k-NN) (Bishop 2007). Before applying these algorithms, two tasks needed to be completed. First, extract the transient profiles (fixed-size collection of power samples around each event) from the overall power signals. Second, separate the transients that were used for validation purposes from the ones that were used during training/testing. For the former, after performing cross-validation experiments with varied parameters on Data Set A and making the assumption that longer windows are generally better unless they are so long that they contain another transient, we finally decided on a preevent window of 2 s and a postevent window of 3 s.

Approximately one transient was randomly picked for each class in each data set (34 for Data Set A and 44 for B) to be used for validation. The remaining transients were used for training the classifiers. Each of the classification algorithms was tested by using different linear-regression basis functions and model orders.

Besides the linear regression coefficients, we also experimented with two other features: a delta metric and the full transient profile. The former and the simplest is the result of subtracting the average value over the samples in a postevent window from the average over a preevent window. This would give the size of the change in the signal and would work best for appliances that quickly reach a steady-state power draw after going through a state transition. The second feature set, the transient profile, is obtained by simply storing the complete set of sample points that describe the transient. Each sample was considered to be a separate feature.

Each combination of feature sets and classification algorithms provided different results, but the best validation results were obtained by using a Fourier basis and the k-NN classifier, with $k = 1$. A summary of these results can be found in Table 3. The order of the model was varied for the basis expansions, but the table only shows the results of the best value. These global accuracy rates are not very meaningful unless one inspects the specific rates for individual appliance-state transitions. To address this, we applied the algorithms to Data Set C. The 62 transients in the data set were first presented to the event detector (by using the parameters previously described), which detected 71 events, including all of the 62 ground-truth events. Upon closer inspection, the nine false positive detections were, in fact, distinct changes in $P_1[n]$ that were not properly labeled. So, effectively, the detector had a 100% TPR. Then, the k-NN algorithm, by using previously obtained signatures for the same appliance-state transitions (i.e., not the transients in the data set) correctly classified 51 of the detected transients. This indicates an overall classification accuracy of 82% (i.e., 51/62). By using the confusion matrix that resulted from this multiclass classification task, we computed the F-measure for each class (appliance-state transition), as shown in Table 4.

Table 3. Validation Result by Algorithm and Feature Set

Validation results (Accuracy in %)		GNB	k-NN, $k = 1$	DT
Data Set A	Delta	53%	68%	62%
	Whole transient	38%	74%	59%
	Polynomial coefficients	59%	68%	53%
	Fourier coefficients	65%	79%	65%
	RBF coefficients	68%	68%	65%
Data Set B	Delta	48%	74%	43%
	Whole transient	10%	74%	48%
	Polynomial coefficients	62%	81%	57%
	Fourier coefficients	50%	81%	55%
	RBF coefficients	48%	76%	55%

Discussion and Conclusions

Even when using simple features (e.g., first-order Fourier basis) to describe the power transients associated with every appliance-state transition, nonintrusive load monitoring is achievable in the laboratory and in residential buildings. Moreover, by providing the homeowners with the opportunity to interact with the system on-site, no pretraining was required.

The results shown are generally promising. The event detector has a low false-positive rate, although slow start-up transients still affect its performance. Classification performance results also indicate that the larger and nonresistive appliances were easily recognized, but the fully resistive loads (e.g., incandescent light bulbs) were many times misclassified. This may be partly because of the fact that the shapes of these transients are not as informative as devices also containing inductive and capacitive loads. In the case of incandescent lights, in most instances, the algorithm confused them with other lights. This could conceptually be resolved by merging all the lights into a single class (e.g., lighting), depending on how the appliance-level information will be used in the end.

However, two larger questions were left unanswered by this analysis. First, the algorithm-specific performances (i.e., event detector, classifier) by themselves are not enough to assess the effectiveness of NILM systems. If the end goal is to obtain appliance-level energy consumption information, perhaps a measure of how well the system estimates these values would be more appropriate. We are currently investigating the use of an energy identification ratio (EIR), defined as the ratio between the estimated energy and the actual energy consumed by an appliance or the home under study.

Second, because of the difficulty of obtaining ground-truth data, the data sets were limited in the number of events, the types of appliances present and the amount of data recorded. Thus, the results of this analysis may not be the best indicators of the true performance of NILM systems. For example, it is important to develop methods that can deal with loads with a continuously varying power draw, dimmers, and other similar situations. The approach presented in this paper would fail under conditions like these.

Future Work

We plan to further explore the use of the system in a real-world setting, with data that is currently collected in a handful of occupied residential buildings. Our intention is to experiment with the features and algorithms presented in this paper, with the possible

Table 4. F-Measure for Classification Test on Data Set C, by Appliance-State Transition

Appliance	State transition (from, to)	Steady-state power change (watts)	Number of events	F-measure
Incandescent light 1	On, off	35	2	100%
Refrigerator light	Off, on	45	2	100%
Refrigerator light	On, off	45	2	80%
Incandescent light 2	Off, on	50	3	80%
Incandescent light 2	On, off	50	3	0%
Incandescent light 4	Off, on	65	1	0%
Incandescent light 4	On, off	65	1	40%
Incandescent light 7	Off, on	65	2	80%
Incandescent light 7	On, off	65	2	100%
Incandescent light 5	Off, on	80	1	50%
Incandescent light 5	On, off	80	2	100%
Television	Off, on	90	2	0%
Television	On, off	90	3	80%
Fluorescent lamp with ballast	Off, on	110	3	86%
Fluorescent lamp with ballast	On, off	110	4	86%
Incandescent light 3	Off, on	120	3	86%
Incandescent light 3	On, off	120	3	86%
Incandescent light 6	Off, on	160	3	100%
Incandescent light 6	On, off	160	3	100%
Refrigerator	Off, on	260	1	100%
Refrigerator	On, off	260	1	0%
Microwave	Off, on	1,000	2	0%
Microwave	On, off	1,000	1	100%
Toaster	Off, on	1,300	3	86%
Toaster	On, off	1,300	3	100%
Steaming iron	Off, on	1,500	3	100%
Steaming iron	On, off	1,500	3	100%

additions of others, given that we expect to be dealing with different transients. For instance, approaches to solve the issues related to appliances with longer transients and overlapping transients need to be developed. A multiresolution event detector could be used to solve the former, whereas a second-stage optimization process that bounds the predicted power consumption to the measured levels could help solve the latter.

More importantly, we will create fully labeled benchmarking data sets that we can utilize to compare different algorithms and refine the EIR metric for this purpose. Additionally, further research is needed on the human-computer interaction issues that such disaggregated data sets for the power consumption of buildings would bring, such as the ones that an implementation of NILM techniques would generate. How much information should be presented to the users of a facility? What is the appropriate way to display it? Which pieces of information are more effective for modifying behavior and reducing energy consumption? In the context of this last question, it may be relevant to investigate the potential benefits of building automation systems (BAS), given that relying on a fully manual response from the users may not be the most effective strategy. NILM systems can provide important information to BAS and reduce the number of sensors that they need to rely on.

Finally, we would like to investigate the use of other inexpensive sources of information present in modern buildings (i.e., other than overall voltage and current) to enhance the disaggregation and continue with the nonintrusive approach to solving the problem. Examples of these are light intensity, audio levels, temperature, and motion sensors.

Acknowledgments

The authors would like to gratefully acknowledge the support of the Robert Bosch LLC Research and Technology Center and the National Science Foundation (NSF) Grant #09-30868. Additional support was provided by the Pennsylvania Infrastructure Technology Alliance (PITA). The opinions presented here are those of the authors and not of the National Science Foundation, Bosch, or PITA.

References

Baranski, M., and Voss, J. (2004). "Genetic algorithm for pattern detection in NIALM systems." *Proc., 2004 IEEE Int. Conf. on Systems, Man and Cybernetics*, IEEE, New York, Vol. 4, 3462–3468 ⟨http://ieeexplore.ieee.org/iel5/9622/30424/01400878.pdf⟩ (Oct. 1, 2007).

Berges, M., Goldman, E., Matthews, H. S., and Soibelman, L. (2010a). "Enhancing electricity audits in residential buildings with non-intrusive load monitoring." *J. Ind. Ecol.*, 14(5), 844–858.

Berges, M., Soibelman, L., and Matthews, H. S. (2010b). "Leveraging data from environmental sensors to enhance electrical load disaggregation algorithms." *Proc., 13th Int. Conf., Computing in Civil and Building Engineering*, Nottingham University, UK.

Berges, M., Soibelman, L., Matthews, H. S., and Goldman, E. (2010c). "Evaluating the electric consumption of residential buildings: Current practices and future prospects." *Proc., Construction Research Congress 2010*, ASCE, Reston, VA.

Bishop, C. M. (2007). *Pattern recognition and machine learning*, 1st Ed., Springer, New York.

Chang, H., Lin, C., and Yang, H. (2008). "Load recognition for different loads with the same real power and reactive power in a non-intrusive load-monitoring system." *Proc., 12th Int. Conf., Computer Supported*

Cooperative Work in Design 2008. IEEE, New York, 1122–1127. ⟨http://ieeexplore.ieee.org/iel5/4530852/4536938/04537137.pdf?arnumber=4537137⟩ (Oct. 1, 2007).

Committee on America's Energy Future; National Academy of Sciences; National Academy of Engineering; National Research Council. (2009). *America's energy future: Technology and transformation*, Summary Ed., National Academies Press, Washington, DC.

Cox, R., Leeb, S. B., Shaw, S. R., and Norford, L. K. (2006). "Transient event detection for nonintrusive load monitoring and demand side management using voltage distortion." *Proc., 21st Annual IEEE Conf.: Applied Power Electronics*, New York.

Darby, S. (2006). *The effectiveness of feedback on energy consumption*, Environmental Change Institute, Univ. of Oxford, Oxford, UK. ⟨http://www.eci.ox.ac.uk/research/energy/electric-metering.php⟩ (Aug. 17, 2009).

Electric Power Research Institute (EPRI). (1997a). "Low-cost NIALMS technology: Market issues and product assessment." *TR-108918-V1*, Palo Alto, CA. ⟨http://mydocs.epri.com/docs/public/TR-108918-V1.pdf⟩ (Apr. 30, 2009).

Electric Power Research Institute (EPRI). (1997b). "Low-cost NIALMS technology: Technical assessment." *TR-108918-V2*, Palo Alto, CA. ⟨http://mydocs.epri.com/docs/public/TR-108918-V2.pdf⟩ (Apr. 30, 2009).

Energy Information Administration. (2001). "End-use consumption of electricity by end use and appliance." Washington, DC. ⟨http://www.eia.doe.gov/emeu/recs/recs2001/enduse2001/enduse2001.html⟩ (Feb. 19, 2009).

Farinaccio, L., and Zmeureanu, R. (1999). "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses." *Energy Build.*, 30(3), 245–259.

Fischer, C. (2008). "Feedback on household electricity consumption: A tool for saving energy?" *Energ. Effic.*, 1(1), 79–104.

Greene, K. (2010). "Technology review: Home sensor startup snapped up." Massachusetts Institute of Technology (MIT), Cambridge, MA. ⟨http://www.technologyreview.com/energy/25205/⟩ (Oct. 6, 2010).

Hart, G. (1989). "Residential energy monitoring and computerized surveillance via utility power flows." *IEEE Technol. Soc. Mag.*, 8(2), 12–16.

Hart, G. (1992). "Nonintrusive appliance load monitoring." *Proc. IEEE*, 80(12), 1870–1891.

Intel. (2010). "Intel's experimental sensor analyzes appliance power consumption from single outlet." ⟨http://www.engadget.com/2010/04/15/intels-experimental-sensor-analyzes-appliance-power-consumption/⟩ (Oct. 6, 2010).

Labview 2010 (Build number: 10.0.0.4032) [Computer software]. Austin, TX, National Instruments.

Lam, H., Fung, G., and Lee, W. (2007). "A novel method to construct taxonomy electrical appliances based on load signatures." *IEEE Trans. Consum. Electron.*, 53(2), 653–660.

Laughman, C., et al. (2003). "Power signature analysis." *IEEE Power Energ. Mag.*, 1(2), 56–63.

Lawrence Livermore National Laboratory. (2009). "Energy flow charts." ⟨https://publicaffairs.llnl.gov/news/energy/energy.html⟩ (Mar. 29, 2010).

Lee, K. D. (2003). "Electric load information system based on non-intrusive power monitoring." Ph.D. thesis, Dept. of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA. ⟨http://dspace.mit.edu/handle/1721.1/29633⟩ (Mar. 26, 2009).

Leeb, S., Shaw, S., and Kirtley, J. (1995). "Transient event detection in spectral envelope estimates for nonintrusive load monitoring." *IEEE Trans. Power Delivery*, 10(3), 1200–1210.

Luo, D., et al. (2002). "Monitoring HVAC equipment electrical loads from a centralized location—Methods and field test results." *ASHRAE Transactions*, 108(1), 841–857.

MATLAB 7.9. Natick, MA, MathWorks.

Matthews, H. S., Soibelman, L., Berges, M., and Goldman, E. (2008). "Automatically disaggregating the total electrical load in residential buildings: A profile of the required solution." *Proc., Intelligent Computing in Engineering (ICE08)*, European Group for Intelligent Computing in Engineering, Lausanne, Switzerland, 381–389. ⟨http://www.eg-ice.org/Workshops/ICE08/papers%20pdf/P014V03.pdf⟩.

Mitchell, T. (1997). *Machine learning*, 1st Ed., McGraw Hill Higher Education, Burr Ridge, IL.

Mitchell, T. (2005). "Generative and discriminative classifiers: Naïve Bayes and logistic regression." *Machine learning*. ⟨http://www.cs.cmu.edu/%7Etom/mlbook/NBayesLogReg.pdf⟩.

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

Paris, J. (2006). "A framework for non-intrusive load monitoring and diagnostics." M.S. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. ⟨http://dspace.mit.edu/handle/1721.1/37081⟩ (Nov. 14, 2008).

Parker, D., Hoak, D., Meier, A., and Brown, R. (2006). "How much energy are we using? Potential of residential energy demand feedback devices." *Proc., 2006 Summer Study on Energy Efficiency in Buildings*, American Council for an Energy Efficiency Economy, Washington, DC.

Peschiera, G., Taylor, J. E., and Siegel, J. A. (2010). "Response-relapse patterns of building occupant electricity consumption following exposure to personal, contextualized and occupant peer network utilization data." *Energy Build.*, 42(8), 1329–1336.

Prudenzi, A. (2002). "A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel." Power Engineering Society Winter Meeting, 2002. IEEE, New York, Vol. 2, 941–946.

Shaw, S. R., Leeb, S. B., Norford, L. K., and Cox, R. W. (2008). "Nonintrusive load monitoring and diagnostics in power systems." *IEEE Trans. Instrum. Meas.*, 57(7), 1445–1454.

Smith, V. A., Norford, L., and Leeb, S. (2003). "Final report compilation for equipment scheduling and cycling." California Energy Commission. ⟨http://www.archenergy.com/cec-eeb/P2-Diagnostics/P2-Diagnostics_Reports/P2/D-2.2-Compiled.pdf⟩ (Apr. 30, 2009).

Srinivasan, D., Ng, W., and Liew, A. (2006). "Neural-network-based signature recognition for harmonic source identification." *IEEE Trans. Power Delivery*, 21(1), 398–405.

Wichakool, W., et al. (2007). "Resolving power consumption of variable power electronic loads using nonintrusive monitoring." *Proc., Power Electronics Specialists Conf.*, IEEE, New York, 2765–2771. ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4342456&isnumber=4341940⟩ (Jul. 11, 2008).