Event Detection for Non Intrusive Load Monitoring

Kyle D. Anderson*, Mario E. Bergés†, Adrian Ocneanu†, Diego Benitez‡
and José M.F. Moura *
*Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, Pennsylvania 15217
Email: kda@andrew.cmu.edu
†Department of Civil and Environmental Engineering
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Email: marioberges@cmu.edu
‡Research and Technology Center, Robert BOSCH LLC
Pittsburgh, Pennsylvania 15203

***Abstract*—Monitoring electricity consumption in the home is an important way to help reduce energy usage and Non-Intrusive Load Monitoring (NILM) techniques are a promising approach to obtain estimates of the electrical power consumption of individual appliances from aggregate measurements of voltage and/or current in the distribution system. In this paper, we discuss event detection algorithms used in the NILM literature and propose new metrics for evaluating them. In particular, we introduce metrics that incorporate information contained in the power signal instead of strict detection rates. We show that this information is important for NILM applications with the goal of improving appliance energy disaggregation. Our work was carried out on a publicly-available week-long dataset of real residential power usage.**

I. Introduction

Reducing energy consumption and cutting costs is an important issue today that affects everyone. In residential environments, saving on electricity may be difficult because the user does not often have an easily accessible frame of reference for how much power appliances or activities in their home consume, or even what their overall power consumption is. The only power consumption feedback most residential consumers receive is an aggregate power bill at the end of each month. Studies have shown, however, that providing consumers with real-time power consumption information, at the aggregate level, may help them to change their behavior and save 10-15% on power costs [1]. Providing users with disaggregated power consumption information could enable them to save even more.

Non-intrusive load monitoring (NILM) is a term used to describe a set of techniques for obtaining estimates of the electrical power consumption of individual appliances, based on measurements of current and/or voltage taken at a limited number of locations of the power distribution system in a building. Thus, NILM can provide real-time power consumption feedback to consumers or automated controllers, enabling them to make more informed decisions that could conserve resources and reduce energy cost. The information could also be leveraged by power companies, automated building controllers, appliance manufacturers, and other parties to increase the efficiency of electricity usage and to better study how electricity is used.

Several different NILM techniques have been proposed in the literature. A good review can be found in [2]. Existing approaches can be roughly categorized into event-based and non event-based depending on whether or not they rely on detecting and classifying each appliance-state transition. In this paper, we examine only the event-based approaches, and focus specifically on the detection stage. The main contributions are two: (a) we introduce a set of metrics for evaluating the performance of event detection algorithms for NILM, and (b) we provide a dataset that can be used by others to test event detection algorithms.

The organization of the paper is as follows. Section II gives a survey of event detection approaches used in the NILM literature. In section III we discuss appropriate metrics for evaluating event detection algorithms and describe the detection algorithm that we use. Then, Section IV introduces the dataset that we use for our experiments and release for public use, and in Sections V and VI we present the details of our experiments and the results, respectively. Section VII summarizes our contributions and ideas for future work.

II. Background Work

We review the existing literature on NILM and select those publications that discuss an approach that requires the explicit detection of change-points in a time-series signal. These publications are then further classified by categorizing the approaches into three categories, namely those based on: (a) expert heuristics, (b) probabilistic models and (c) matched filters.

A. Expert Heuristics

Hart [3] was, to the best of our knowledge, the first to show how step-like changes present in the power consumption signal can be used to distinguish between appliances. His approach to detecting these changes first segments normalized power values into steady and transient periods. Steady periods are defined as a predetermined fixed-size set of contiguous samples in which the "input does not vary by more than a specified tolerance". This variation is measured by obtaining

978-1-4673-2421-2/12/$31.00 ©2012 IEEE

the standard deviation of the fixed-size set of samples.

Similarly, in [4], the authors use a rule-based approach to detect the *on* and *off* transitions of a select group of appliances. The "state change detection" rule scanned the difference series of the total power and compared these values to pre-determined ranges for the start and end (i.e., *on* and *off*) power changes associated with each appliance. A very similar approach was presented in [5]. It is worth noting that this procedure requires an initialization process to determine these ranges for each appliance in the home.

B. Probabilistic Models

In the statistics literature, as well as in many other domains, the problem of detecting abrupt changes in time series data is known as *change detection*. Furthermore, the identification of the specific time instance when this change occurs is sometimes referred to as *change-point detection*. A good review of these methods can be found in [6].

There have been a number of techniques from this domain that have been applied to the problem of detecting events in power signals. In [7], for instance, researchers introduce the use of a Generalized Likelihood Ratio (GLR) approach, which calculates a "decision statistic from the natural log of a ratio of probability distributions before and after a potential change in mean". Their approach requires that four parameters be trained offline including the length of the moving windows, the variance of the power data, and a threshold for the detection statistic. In [8], a similar method is proposed which removes the need to specify the variance. This method is then compared to another statistical approach, namely the Goodness of Fit (GoF) statistic, in [9].

C. Matched-Filters

A different approach to detecting events is to use matched filters, which correlate a known signal (usually known as the mask) with an unknown one to detect the presence of the mask in the unknown signal. The masks can be template power signals containing the start-up or turn-off transients of different appliances, and the unknown signal is the total power consumption. In [10], a multi-resolution matched filter for this problem was presented. This same approach, with minor modifications, was the main detector used in subsequent work presented in [11] and [12].

III. Event-Based Non-Intrusive Load Monitoring

A typical event-based NILM approach consists broadly of four processes: power measurement, event detection, event classification, and energy estimation. This can be seen in the block diagram presented in Figure 1. By monitoring the electrical activity in the home (i.e., power measurements), the system attempts to detect when an appliance has changed its state by looking for changes in this signal (event detection) and then attempts to determine which appliance caused this change (classification) in order to track its electric power consumption (energy disaggregation).

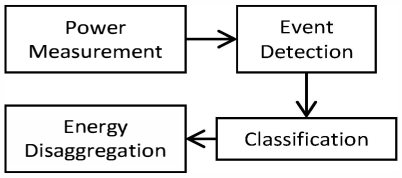

Fig. 1. Block diagram showing the steps involved in a generic event-based NILM algorithm.

To evaluate the performance of an event detection algorithm, it is necessary to have the ground truth for the data (i.e., power measurements) on which the test is being run. In an event-based NILM application this ground truth is a list of timestamps indicating when events actually occur. An event detection algorithm, hence, takes a power signal as its input and returns a list of timestamps of detected events. There are many such algorithms, as discussed in II. For details about the specific detector we utilize, see section III-B below.

A. Event Detection Metrics

We have found that there is no standard framework in the literature for evaluating the performance of event detectors in NILM applications. We provide here a discussion on possible metrics and evaluate their performance on the BLUED dataset [13] in section VI below.

The ultimate goal of the NILM system we present is to disaggregate power consumption by appliance. With this in mind, the best possible event detector would be the one that leads to the best disaggregation. If E_k is the energy consumed by appliance k and $\widehat{E}_k$ is the estimate of appliance k's energy, then, for a particular event detection algorithm, the best detector would be that which solves equation (1).

$$\min_{\psi \in \Psi} \sum_{k=1}^{N} \|E_k - \widehat{E}_k\| \tag{1}$$

We denote the parameter space of a particular event detection algorithm and its elements as Ψ and ψ, respectively.

Unfortunately, this minimization is not trivial, and it is also intrinsically linked with the performance of the classification stage in the NILM algorithm (see figure 1). Instead of solving the minimization directly and finding the best event detection algorithm based on the final energy approximation, we would like a metric that decouples the event detection performance from the performance of the classification stage.

We now present four different metrics for achieving this goal. First we describe each of the metrics and provide some intuition about what type of performance their optimal detectors will have. In section VI we will show their performance on a real dataset.

1) True Positive Rate: A simple metric for evaluating an event detector is its receiver operating characteristics (ROC), typically visualized in the ROC curve. This metric finds the detector that has the best tradeoff between its true positive rate (TPR) and false positive rate (FPR), defined below, where TP is the number of true positives, FP is the number of false

positives, TN is the number of true negatives, and FN is the number of false negatives or misses.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \in [0,1] \quad (2)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \in [0,1] \quad (3)$$

The perfect detector would have a TPR of 1 and a FPR of 0 and, given a particular ROC curve, the best detector, $\widehat{\psi}_{\mathrm{Rate}}$ is that which is closest to the point $(0,1)$.

$$\begin{aligned}\widehat{\psi}_{\mathrm{Rate}} &= \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \|(0,1)-(\mathrm{FPR},\mathrm{TPR})\|^2 \\ &= \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \mathrm{FPR}^2+\mathrm{TPR}^2-2\cdot\mathrm{TPR}+1 \quad (4)\end{aligned}$$

Note that in equation (4) there is a tradeoff between the false positive and true positive rates. In residential power data, however, the events are sparsely distributed. For an event detector with reasonable performance, the sparsity of actual events means that the number of true negatives will be much higher than the number of false positives, $\mathrm{TN} \gg \mathrm{FP}$. When this is the case equation (3) goes to 0.

$$\mathrm{FPR} \xrightarrow{\mathrm{TN}\gg\mathrm{FP}} 0 \quad (5)$$

Thus the minimization in equation (4) becomes

$$\widehat{\psi}_{\mathrm{Rate}} = \underset{\psi\in\Psi}{\operatorname{argmin}} \quad f(\mathrm{TPR}) \quad (6)$$

$$f(\mathrm{TPR}) = \mathrm{TPR}^2-2\cdot\mathrm{TPR}+1 \quad (7)$$

On the interval $[0,1]$ this is a monotonically decreasing function. Therefore the best event detector, $\widehat{\psi}_{\mathrm{Rate}}$ will be the one with the highest true positive rate, regardless of how many false positives it has (provided that the assumption $\mathrm{TN} \gg \mathrm{FP}$ holds).

2) True Positive Percentage: In this metric, we compare the percentage of events correctly detected to the ratio of false positives to total number of actual events. For visualizing the curve, we divide each by the total number of events, E, in the dataset and define the true positive percentage (TPP) and false positive percentage (FPP):

$$\mathrm{TPP} = \frac{\mathrm{TP}}{E} \quad (8)$$

$$\mathrm{FPP} = \frac{\mathrm{FP}}{E} \quad (9)$$

Note that the latter can not be properly called a percentage because the number of false positives can be larger than the number of actual events.

Similar to the rate metric above, the perfect detector would have a TPP of 1 and a FPP of 0. Therefore the optimal detector, $\widehat{\psi}_{\mathrm{Perc}}$ is that which is closest to the point $(0,1)$. We can also see that this is equivalent to an optimization that considers just the count of true and false positives.

$$\begin{aligned}\widehat{\psi}_{\mathrm{Perc}} &= \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \|(0,1)-(\mathrm{FPP},\mathrm{TPP})\|^2 \quad (10)\\ &= \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \|(0,E)-(\mathrm{FP},\mathrm{TP})\|^2 \quad (11)\end{aligned}$$

We expect this metric to be a better measure of performance than the rate metric because there is an actual trade-off. However, there is more information in the power signal that is still not being used. The next two metrics will take advantage of this information.

3) Total Power Change: The first two metrics considered only the relative number of true and false positives; there is an implicit assumption in them that all events are of equal importance. However this is not a good assumption because some appliances have higher power consumption than others.

As a metric for incorporating information about the power change of events, we consider the tradeoff between the sum of the power changes of all the misses and the sum of the power changes of all the false positives.

Given an edge, e, we define its power change ΔP_e in equation (12). Qualitatively speaking it is the difference of the mean of the signal a short time after the event and just before.

$$\Delta P_e = \frac{1}{w_3}\sum_{i=e+w_2+1}^{e+w_2+w_3} P(i) - \frac{1}{w_1}\sum_{i=e-w_1}^{e-1} P(i) \quad (12)$$

The terms w_1, w_2, and w_3 are window lengths; w_1 and w_3 are used for calculating the pre- and post-event means and w_2 is used to allow a delay for the transient to end and reach a more steady state.

With this definition, we can now define a new metric based on the changes in power of both the misses and false positives. In the following equations, let $\mathcal{M}$ be the set of all misses and let $\mathcal{F}$ be the set of all false positives. The total power changes for the misses and false positives are ΔP_{M} and ΔP_{FP} respectively, described in equations (13) and (14) below.

$$\Delta P_{\mathrm{M}} = \sum_{m\in\mathcal{M}} |\Delta P_m| \quad (13)$$

$$\Delta P_{\mathrm{FP}} = \sum_{f\in\mathcal{F}} |\Delta P_f| \quad (14)$$

The optimal detector in this metric, $\widehat{\psi}_{\Delta P}$, is that which is closest to the point $(0,0)$.

$$\widehat{\psi}_{\Delta P} = \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \|(\Delta P_{\mathrm{FP}}, \Delta P_{\mathrm{M}})\|_2^2 \quad (15)$$

4) Average Power Change: This metric also balances the tradeoff between the power of misses and false positives, but instead of evaluating the total power, it seeks to minimize the average power of each. We define the average power for the misses and false positives as follows:

$$\overline{\Delta P}_{\mathrm{M}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |\Delta P_m| \quad (16)$$

$$\overline{\Delta P}_{\mathrm{FP}} = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} |\Delta P_f| \quad (17)$$

The optimal detector in this metric, $\widehat{\psi}_{\overline{\Delta P}}$, is that which is closest to the point $(0,0)$.

$$\widehat{\psi}_{\overline{\Delta P}} = \underset{\psi\in\Psi}{\operatorname{argmin}} \quad \left\|\left(\overline{\Delta P}_{\mathrm{FP}}, \overline{\Delta P}_{\mathrm{M}}\right)\right\|_2^2 \quad (18)$$

5) Score Function: In order to combine the above metrics, we define a score function, $S(\psi)$ that rates the overall performance of a particular detector, ψ, across the set of the four different metrics, M. The score function is given below in equation (19)

$$S(\psi) = \frac{1}{4} \sum_{\mu \in M} \frac{D_\mu(\widehat{\psi}_\mu, \theta_\mu)}{D_\mu(\psi, \theta_\mu)} \tag{19}$$

In this equation, D_μ is the Euclidean distance in the space of the tradeoffs of the metric μ, and θ_μ represents the optimal detector in this space. For example, in the total power change metric, the space is that of the total power change of the misses vs. the total power change of the false positives and $\theta_{\Delta P} = (0, 0)$.

B. Modified GLR Detector

To test the effectiveness of these performance metrics, the event detector that we use is a modified generalized likelihood ratio (GLR) that uses a log likelihood ratio test and voting [8]. It has five tunable parameters that comprise its parameter space Ψ: a power threshold, P_{thr}; pre- and post-event window lengths, w_{pre} and w_{post}, respectively; a voting window length w_{vote}; and a vote threshold, v_{thr}.

For each point in the power signal, $P[k]$, we compute the log likelihood ratio $l[k]$:

$$l[k] = \ln\left(\frac{\mathcal{N}(P[k]|\mu_{\text{post}}, \sigma_{\text{post}})}{\mathcal{N}(P[k]|\mu_{\text{pre}}, \sigma_{\text{pre}})}\right)$$

where μ_{pre}, σ_{pre} and μ_{post}, σ_{post} are the sample mean and sample standard deviation of the pre- and post-event windows, respectively. The log likelihood ratio of all points that fail to satisfy the power threshold, i.e., $|\mu_{\text{post}} - \mu_{\text{pre}}| \leq P_{\text{thr}}$ are set to 0. In the experiments carried out for this paper we used a threshold of 30 Watts but a parameter search could be done.

A voting window of length w_{vote} is then slid over the resulting time-series of log-likelihood ratios. A point receives a vote if its log likelihood ratio, $l[k]$ from equation (20), is the highest among all points in the voting window. Each point receiving at least v_{thr} votes is flagged as an event.

We make no claims as to the superiority of the GLR detector for the NILM event-based approach. The purpose of this paper is to provide a set of metrics for evaluating and comparing different detection algorithms against one another. We use the GLR detector as a simple detection algorithm to illustrate the validity of the metrics we propose.

IV. Dataset

As part of ongoing efforts to unite researchers working on NILM [14], our research team recently released the Building-Level fUlly-labeled dataset for Electricity Disaggregation (BLUED) dataset [13], which is available for download at http://nilm.cmubi.org.

Using a data collection system similar to the one used in [15], one week of aggregate voltage and current measurements were collected at 12 kHz for the two voltage sources in a split-phase electrical panel at a home in Pittsburgh, PA. From these measurements, real and reactive power were computed at a rate of 60 Hz. In addition to these measurements, plug-level power meters coupled with light-intensity sensors located near overhead lights were used to obtain ground-truth information about the appliance-specific events (i.e., a timestamp of when each event occurred, and what appliance it belonged to). Thus, our ground truth is available at the appliance level.

TABLE I
SUMMARY OF EVENTS IN EACH PHASE.

Phase	# of Appliances	# of events	Sample Appliances
A	10	872	Refrigerator, Pump
B	22	1548	Microwave, TV

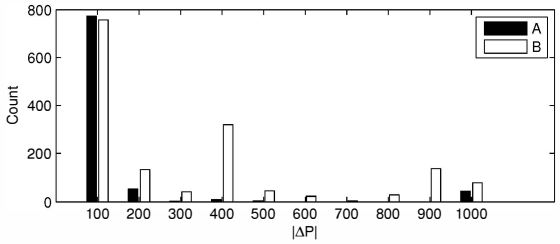

Fig. 2. Histogram of the distribution of events on phases A and B.

There are approximately 30 individual appliances in the BLUED dataset. In a split-phase electricity distribution system, each appliance can be powered by one or two of the voltage sources (commonly referred to as *phases*) coming into the building. Phase A, in our dataset, was feeding a relatively lower number of appliances than Phase B, thus resulting in a *noisy* Phase B, as the analysis will show. Table I summarizes the properties of the data collected from each phase and figure 2 shows a histogram of the distribution of power changes of events on phases A and B. For a more in-depth description of the dataset and summary of its contents, please refer to [13].

V. Experimental Setup

The main results are an analysis of multiple runs of the modified GLR event detection algorithm described in section III-B over the week-long dataset described in section IV. Four of the five detection parameters were varied and the other fixed for a total of 1062 different parameter combinations. Table II shows the ranges for each of the parameter combinations tested. Each configuration was run on the full week of data to produce a list of potential events. Each event was then compared with the ground truth from the dataset and labeled either as a true positive, false positive, or a miss. The power change for each event was also computed according to equation (12). With this data the performance of each detection run was computed according to the metrics described in section III-A.

VI. Experimental Results

After completing our experiments, we found that each metric produced a different set of optimal detection parameters $\widehat{\psi}$. Table III shows the optimal parameters found under each

TABLE II
PARAMETER RANGES USED FOR TESTING THE GLR DETECTOR.

Parameter	Range
w_{vote}	120–330 samples (2–5.5 seconds)
$w_{\text{pre}} = w_{\text{post}}$	15–w_{vote} samples (.25–5.5 seconds)
v_{thr}	15–w_{vote} samples
P_{thr}	30 Watts

TABLE III
OPTIMAL GLR PARAMETERS UNDER EACH METRIC.

	$\widehat{\psi}_{\Delta P}$	$\widehat{\psi}_{\overline{\Delta P}}$	$\widehat{\psi}_{\text{Perc}}$	$\widehat{\psi}_{\text{Rate}}$
w_{vote} (samples)	210	300	120	120
$w_{\text{pre}} = w_{\text{post}}$ (samples)	15	300	60	15
v_{thr} (samples)	30	90	75	15
P_{thr} (Watts)	30	30	30	30

metric within the range of parameters tested as shown in table II. Furthermore, we found that the optimal detector for one metric did not necessarily perform well according to the other metrics. This is particularly interesting because a detector that is perfect in any of the metrics will be perfect in all of them.

For each metric we selected the Pareto frontier to show the tradeoffs of the objective functions of the metrics. Figure 3 shows this frontier for the output of the $\widehat{\psi}_{\Delta P}$ detector on phase B. On this plot, the total power performance of all the optimal detectors is shown. We clearly see that each of them is different and that none have the same performance. Figure 4 shows the frontier for the output of the $\widehat{\psi}_{\text{Perc}}$ detector. Comparing the two figures shows how the different optimal detectors vary in performance under different metrics. This motivates the merit of the score function across all the metrics.

Figures 5 and 6 show histograms of the distribution of detected events for the $\widehat{\psi}_{\Delta P}$ and $\widehat{\psi}_{\text{Perc}}$ detectors respectively.. The x-axis shows the spread of power changes that were detected for true positives, misses, and false positives. Comparing the two reveals that the $\widehat{\psi}_{\Delta P}$ detector has more overall false positives that than the $\widehat{\psi}_{\text{Perc}}$ detector, but it does not miss as many of the larger powered events. There is a tradeoff, and with the assumption that missing events with large power changes will affect the energy estimation more than small false positives, the $\widehat{\psi}_{\Delta P}$ detector would be preferable.

Table IV shows the score for each of the optimal detectors

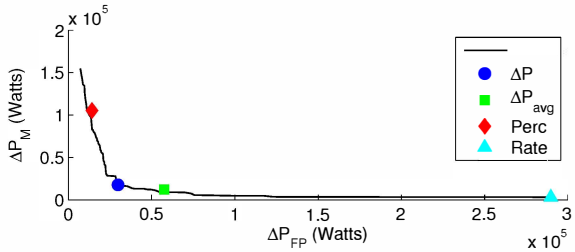

Fig. 3. Pareto frontier for the total power metric shows the tradeoff between the total missed power and the total power contained in the false positives. The operating points of the optimal detectors for each metric are shown as well.

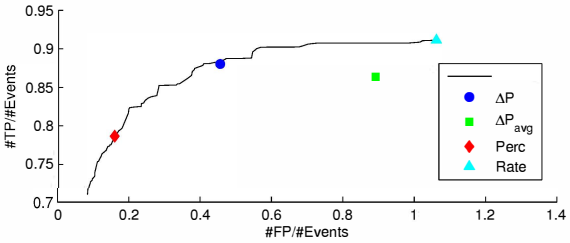

Fig. 4. Pareto frontier for the percentage metric shows the tradeoff between the percentage of true positives and false positives.

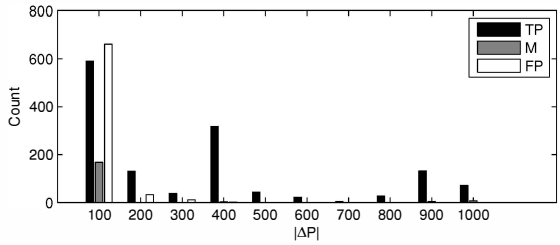

Fig. 5. Histogram of the optimal total power detector, $\widehat{\psi}_{\Delta P}$.

TABLE IV
SCORE FUNCTION FOR EACH METRIC ON THE DIFFERENT PHASES.

	Phase A	Phase B
Total Power	0.80	0.74
Avg Power	0.44	0.63
Percentage	0.78	0.49
Rate	0.30	0.44

on each phase. We note that the $\widehat{\psi}_{\Delta P}$ and $\widehat{\psi}_{\text{Perc}}$ detectors both score well on phase A, but on phase B, the $\widehat{\psi}_{\Delta P}$ detector clearly performs the best.

VII. CONCLUSIONS AND FUTURE WORK

We have provided metrics and a framework for evaluating event detection algorithms for use in non-intrusive load monitoring applications. This is important because there is not currently a unified way of quantizing algorithm performance among different researchers for this problem. Our results indicate that the total power change metric had an overall better performance than the other three metrics we investigated. Of particular interest is the fact that this metric incorporates information in the power signal and not just the number or rate of true and false positives. It is important to include

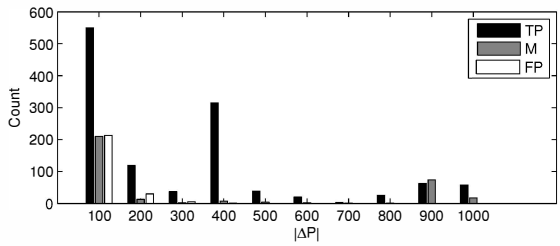

Fig. 6. Histogram of the optimal percentage detector, $\widehat{\psi}_{\text{Perc}}$.

power in the metric for selecting an event detection algorithm because the overall goal of a residential NILM system is to disaggregate the energy consumed by each appliance. An event detection metric that does not place a higher weight on events of larger power consumption misses important information.

We have evaluated our metrics using a modified generalized likelihood ratio detector, but we plan to further validate the results by evaluating the performance of other event detection algorithms on the same dataset. The week-long dataset of actual electrical usage for a residential building has been made publicly available and we plan to release additional datasets of voltage and current measurements along with timestamped events so that the research community can continue to compare and contrast different algorithms.

Our longer term plan is to incorporate energy disaggregation as a metric for selecting NILM algorithms. This is not a trivial extension as it on the ability to track when an appliance turns both on and off. Incorporating power change information into the metrics is an important step towards finding a metric that also involves total energy disaggregation.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge the support from the Robert Bosch LLC Research and Technology Center North America and the National Science Foundation (NSF) grant #09-30868. The opinions expressed herein are those of the authors and not of the NSF.

REFERENCES

[1] S. Darby, "The effectiveness of feedback on energy consumption," Environmental Change Institute, University of Oxford, Oxford, UK, Tech. Rep., 2006.

[2] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 1, pp. 76 –84, Feb. 2011.

[3] G. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[4] L. Farinaccio and R. Zmeureanu, "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses," *Energy and Buildings*, vol. 30, no. 3, pp. 245–259, Aug. 1999.

[5] M. Baranski and J. Voss, "Detecting patterns of appliances from total load data using a dynamic programming approach," in *Fourth IEEE International Conference on Data Mining (ICDM'04)*, Brighton, UK, 2004, pp. 327–330.

[6] M. Basseville and I. Nikoforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[7] D. Luo, L. Norford, S. Shaw, and S. Leeb, "Monitoring HVAC equipment electrical loads from a centralized location - methods and field test results," *ASHRAE Transactions*, vol. 108, no. 1, pp. 841–857, 2002.

[8] M. Berges, E. Goldman, H. S. Matthews, L. Soibelman, and K. Anderson, "User-centered non-intrusive electricity load monitoring for residential buildings," *Journal of Computing in Civil Engineering*, vol. 25, no. 6, pp. 471–480, 2011.

[9] Y. Jin, E. Tebekaemi, M. Berges, and L. Soibelman, "A time-frequency approach for event detection in non-intrusive load monitoring," Orlando, Florida, USA, 2011, pp. 80 501U–80 501U–13.

[10] S. Leeb, S. Shaw, and J. Kirtley, "Transient event detection in spectral envelope estimates for nonintrusive load monitoring," *Power Delivery, IEEE Transactions on*, vol. 10, no. 3, pp. 1200–1210, 1995.

[11] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong, "Power signature analysis," *Power and Energy Magazine, IEEE*, vol. 1, no. 2, pp. 56–63, 2003.

[12] S. Shaw, S. Leeb, L. Norford, and R. Cox, "Nonintrusive load monitoring and diagnostics in power systems," *Instrumentation and Measurement, IEEE Transactions on*, vol. 57, no. 7, pp. 1445–1454, 2008.

[13] K. Anderson, A. Ocneanu, D. Benítez, D. Carlson, A. Rowe, and M. Bergés, "Blued: A fully labeled public dataset for event-based non-intrusive load monitoring research," in *Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012.

[14] M. Berges and Z. Kolter, "1st international workshop on non-intrusive load monitoring," http://www.ices.cmu.edu/psii/nilm/, May 2012. [Online]. Available: http://www.ices.cmu.edu/psii/nilm/

[15] J. Kolter and M. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD)*, San Diego, CA, 2011.