Towards Automated Non-Intrusive Load Monitoring Performance Evaluation

Lucas Pereira
Madeira-ITI, Caminho da Penteada, 9020 - 105 Funchal, Portugal
lucas.pereira@m-iti.org

ABSTRACT
In the past couple of years Non-Intrusive Load Monitoring has been gaining special attention from the research community playing an important role in the solution of problems related to energy sustainability and smart-grids.

However despite its notoriety, until recently no efforts were made towards having formal methods for evaluating its performance. In this paper we present our hardware-software framework for creating extensive fully labeled datasets for non-intrusive load monitoring evaluation where raw Current and Voltage waveforms are stored in audio files and a commercial sub-metering solution is used to collect appliance level ground-truth data.

Towards the end we present our insights on what should be the future of NILM performance evaluation and present a four-step approach towards the automation of this process.

Keywords
Non-intrusive load monitoring, datasets, ground-truth, algorithm performance evaluation.

1. INTRODUCTION
Non-Intrusive Load Monitoring (NILM) is one of the most well know techniques to identify and monitor the energy consumption of individual appliances that co-exist in a building’s electrical circuit. However, and despite the fact that NILM is undergoing extensive research towards making it a viable solution for realistic environments, it is evident that there are still some research avenues that need to be explored. For example, most of the existing studies were done in laboratory settings with small and unbalanced data sets and the almost total inexistence of studies on the human-computer interaction issues that can be generated by the new kind of data that we are getting from NILM techniques.

This research is a follow up on previous work done at Madeira Interactive Technologies Institute (M-ITI) where we studied the effects of deploying electric energy eco-feedback in households for long periods of time (from 9 weeks up to 18 months). These studies were conducted using a custom made event-based non-intrusive load monitor [11, 12] that we now wish to extend and use to answer two important research questions that we believe have been, for various reasons, largely ignored by both NILM and eco-feedback / HCI research communities: 1) How effective is NILM in real world scenarios? And 2) what are the best strategies to manually train event-based NILM systems?

In this paper we will be focusing on the first question. We start by introducing NILM and the most relevant literature on this field. Next we explore the foundations of the question, explaining the why and how we plan to conduct our research in the effectiveness of NILM solutions in real world scenarios.

2. NON-INTRUSIVE LOAD MONITORING
In very broad terms NILM can be defined as a set of techniques used to obtain estimates of the electrical consumption of individual appliances with the application of sophisticated signal processing and machine learning algorithms to Current and / or Voltage measurements, taken at a limited number of locations of the power distribution in a building.

Attempts to monitor and disaggregate electric energy from a single point dates back to the early 1980’s and were first introduced by Schwepe and Hart [3] who coined the term Non-Intrusive Load Monitoring. Their main assumption was that every change in the total power consumption of a building would happen in response to an electric device changing its state (e.g. a hair dryer going from low to high), therefore the initial approaches consisted mostly of trying to match the amount of change in power related metrics (e.g. real and reactive power) to appliance states.

NILM has been subject to some extensive research in the past few years and today it is common two find the existing techniques being categorized as belonging to one of two approaches namely: event-based and non-event based. These are two very different approaches but in very broad terms we can say that the main difference between both is mostly due to the fact that the former relies on keeping track of every appliance state transition (by means of event detection and classification assuming that the system was previously trained) [2, 9], while the later normally does not assume previous knowledge of the existing appliances and attempts to disaggregate the individual loads from the total power (normally at low sampling rates) by means techniques like prior models of general appliance types [10] or temporal motif mining [14]. An extensive review of the existing approaches can be found in [15].

Nevertheless, and despite all the research efforts to the date some considerable challenges are still present. In one hand researchers have to deal with the problem of identifying the individual loads in the ever-growing complexity of the domestic electric circuits where they need to account for very different loads like variable, multistate and always on. On the other hand there is the problem of not having a formal method to evaluate NILM performance
making it impossible to generalize research findings across different problems.

In this paper we focus in the problem of properly accessing the performance on the existing approaches.

3. NILM PERFORMANCE EVALUATION

As said above one of the current challenges of NILM research is the need to properly access the performance of the existing solutions and it is believed that the main reasons for this are: i) the lack of agreement upon what performance metrics should be used to properly compare algorithm results [15] and ii) the almost inexistence of publicly available datasets like the ones that exist for other machine learning problems like character recognition and spam detection [5, 6].

In an attempt to address this, the research community has turned their attention into finding a common metrics [4, 7] as well as on creating public NILM datasets.

To the best of our knowledge there are currently four datasets that are briefly described in the next section.

3.1 Publicly Available Datasets

The Reference Energy Disaggregation Data Set (REDD) [5], which was primarily released for the non-event based approach, consists of whole-home and circuit / device level consumption data from a number of US houses collected over several months' time. The whole-house data was recorded at a high frequency (15 kHz) while the circuit / device data was recorded at 0.5 Hz for the individual circuits and 12 Hz for the single electronics.

The Building-Level fully-labeled dataset for Electricity Disaggregation (BLUED) [6] was, on the other hand, especially tailored for the evaluation of event-based NILM approaches. It consists of one week of whole-house current and voltage high frequency measurements (12 kHz) from one US home. Additionally, a list of labels (e.g. timestamp and appliance identifiers) is provided for each state transition occurred in the dataset for a total of 43 appliances.

Another public dataset is the UMass Smart* Home Data Set [1], that despite not being specifically collected for energy disaggregation, provides power data for three sub metered houses in the US. Power measurements are taken at the mains panel and individual circuits at the frequency of 1 Hz. There is also power data available for individual appliances collected every few seconds.

Finally there is also Tracebase [13] in which individual appliance consumption data was collected at 1 Hz frequency from 158 appliances instances in a total of 43 different appliance types. The power consumption traces of each individual appliance are stored in individual files for each day.

3.2 Main Challenges

It is our believe that the best way to evaluate the performance of NILM systems is by testing the existing algorithms using the same performance metrics under common datasets that would be a faithful representations of real world scenarios.

However it is well known that mimicking these real-world scenarios is not bound to be an easy task especially if we consider all the dynamics of the electric grid in a modern house, for example a proper NILM system will need to account for the presence of unknown and / or malfunctioning appliances, the various operation modes of a single appliance (e.g. a microwave can be used to make popcorn or to defrost food) as well the different conditions of operation (e.g. on weekends normally people do the laundry and therefore the system will need to take into consideration the several cycles of the washing machine while for example the oven is being used to cook lunch and the vacuum cleaner is being used to clean the car).

In addition to this it is also important to take into consideration the fact that such a dataset will need to accommodate several different algorithms and that those algorithms will, on the other hand, require different evaluation metrics that can only be calculated using data from the dataset itself. We refer to this as the NILM Performance Evaluation Loop, and it is summarized in figure X.

Figure 1. The NILM Performance Evaluation Loop: A dataset needs to accommodate several different algorithms that will require their own evaluation metrics which can only be calculated using data from the original dataset.

Dataset establishments for NILM research have gained some interest in the past few months. In [8] authors introduced what they consider the three most important properties of a dataset for NILM research, namely: i) be informative, in a sense that the data in the dataset must contain enough information such that it can serve as training and test sets for as much methods as possible; ii) be diverse, such that the learning methods are able to capture the dynamics that householder actions will have on the electric grid; and iii) be scalable in a sense that new data (e.g. location of events and steady state areas) has to be easily incorporated in an existing dataset.

To these three properties we have added two others that we consider of equal importance. They are: iv) low overhead and easy access to the data; and v) extended ground-truth annotations and will be detailed in the next section where we also propose a new kind of dataset for NILM.

3.3 Proposed Dataset

After using the existing datasets for a while it became clear to us that they have a very high footprint, especially when considering the amount of required space in the hard-drive and the volume of code we had to produce just to interface with the data. Therefore we believe that if a dataset is to be successful it will need to present a low overhead by maximizing the information / file size.
ratio and providing easy mechanisms to access the raw waveforms data and the ground-truth labels.

We also believe that it is of major importance to have as much details of the dataset as possible. For example, the dataset should provide extended appliance activity labeling (e.g. which appliance, when and how much energy was used), individual appliance working cycles (it is important to know when cycling appliances like clothes washers or vacuum cleaners are being used as these can strongly affect event detection algorithms). Additionally we also advocate that it would be good to have labels for household activities (e.g. cooking, cleaning or playing) and other sorts of metadata like the weather conditions when the dataset was collected or detailed information about the householders (e.g. how many residents and the number of children).

With this in mind we have decided to build our own public dataset consisting of single appliance consumption data, whole home and individual circuits’ waveforms and extended ground-truth annotations.

### 3.3.1 Single Appliance Dataset

For the appliance-level data collection we have used the Plugwise\(^1\) system (this was also used in [13]), which is a commercially available, distributed sub-metering platform. The system is made up of three components namely the Circle (also called Module), the Stick and the Source software. Each circle is connected between the appliance being measured and the wall outlet. The Stick is then used to wirelessly (using the ZigBee wireless protocol) interface each deployed circle with a computer (that has the source software installed) that processes and displays the consumption data collected from each individual circle.

Before displaying it to the users, the measured data is aggregated (by hour, day or month) which is appropriate if we have the sole purpose of showing the total power consumption of each individual appliance. Still this is far from offering a level of granularity that is enough to understand in detail the individual behaviors of each appliance, e.g. if it is On or Off and when these transitions happened. Consequently we had to develop our own Application Programming Interface (API) to provide us with an easy access to the consumption data of each individual appliance in the loop. This API is based on the Plugwise Template Engine (PTE), which runs inside multi-threaded server that is running inside the source software.

The PTE works by request, i.e. the clients will request a file with one of the extensions handled by the web-server (we are using XML) and this file will be parsed by the PTE replacing the template tags with the last known values from each plug.

Our API follows the event-based software architecture and works by requesting the state of the plugged-in modules once every second and dispatching events upon changes in the modules’ internal state. The choice of this architecture has proven useful as it offers the possibility of listening and reacting to each module individually according to our needs.

From the available events we like to stand out the \texttt{AppliancePlugged/ApplianceUnplugged} events that are triggered when an appliance is plugged or unplugged from a module and also the \texttt{CurrentPowerConsumptionChanged} event that is dispatched whenever there is a change in the power consumption of a given module between two consecutive readings.

Additionally we have also added the possibility of programming custom event detectors for each individual module or groups of modules, which is especially important if we wish to give each plug the functionality of automatically identifying the appliances that are connected to it.

For the ground-truth collection process we developed an application on top of our API that will collect and store the consumption data of each module in a local database in pre-defined intervals of time. Additionally we also store the aggregated consumption (i.e. the sum of the instantaneous power readings of each individual module) in another table, as this may become important if we want to test low-resolution NILM approaches like the ones mentioned in section 2. Figure 2 shows the power trace of a refrigerator during 24 hours.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{refrigerator_trace.png}
\caption{Partial trace of the refrigerator ground-truth measurements showing that in 24 hours the engine became active 15 times.}
\end{figure}

### 3.3.2 Whole-house and Individual Circuit Dataset

For the whole-house data collection we have taken advantage of our custom made Non-Intrusive Load Monitoring framework that is thoroughly described in [11, 12].

The system consists of a netbook that is installed in the main power feed (see Figure 3) covering the entire house consumption. The audio input from the netbook soundcard is used as the data acquisition module using its stereo capabilities (2 channels, one for Current and the other for Voltage).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{current_and_voltage_sensors.png}
\caption{Current and voltage sensors are installed in the main breaker box hence covering the entire house consumption.}
\end{figure}

\footnote{\url{www.plugwise.com}}
Despite the Soundcard based NILM have proven to be a very effective solution the fact that it only supported two channels was a major disadvantage. Therefore we have extended the framework such that it would support the usage of a multi-port Data Acquisition board (DAQ).

With this new DAQ (LabJack U6\(^2\), which offers 14 analog input ports with 16-bit resolution for a maximum sampling rate of 50 kHz) our system gained the capability of monitoring not only the whole-house consumption but also offered the opportunity to monitor the home’s individual circuits or even multiple houses (as shown in figure 4).

![Figure 4. Current sensors installed in the main building breaker box, hence covering the whole-house consumption of six houses from one single location.](image)

### 3.3.3 Putting it All Together

The acquired Current and Voltage waveforms are persisted to the hard-drive in the wave audio format.

The main reason behind using such an audio format is its reduced file size (24 hours of Current and Voltage waveforms sampled at 8 kHz with a sample size of 16 bits will take roughly 1.5 GB of disk space) when compared for example with the typical Comma Separated Values (CSV) text format where the same 24 hours of Current and Voltage may take up to 100 GB depending on the used precision.

Another advantage is the possibility of easily adding information about the files through the use of metadata, which can be used to write blocks of text identifying the contents of the file.

Another feature of this audio file format is the possibility of adding markers directly in the file. These markers are references to positions in the waveforms and can be used to supplement the power measurements with information about individual appliance activity (as shown in figure 5) therefore merging raw waveforms with ground-truth data in a single location.

![Figure 5. Wave files offer the possibility of embedding markers that can be used to supplement the power measurements with ground-truth data (e.g. appliance and user activities).](image)

### 4. TOWARDS AUTOMATED NILM PERFORMANCE EVALUATION

We are convinced that the ability to merge raw waveforms with ground-truth data is a very important step towards simplifying the process of evaluating different NILM approaches.

Still we also believe that to achieve a successful evaluation framework the whole evaluation process needs to be straightforward, alleviating researchers from the burden of having to worry about how to access the datasets, what metrics to use and how to compute them. This is why we defined Automated NILM Performance Evaluation as being a four-step process.

![Figure 6. The four steps that comprise the envisioned automated NILM performance evaluation.](image)

In the first step (Preparation), the researcher will have to select which algorithms will be tested and under which conditions and performance metrics that will happen. In the second step (Execution) the evaluation framework would take to itself the responsibility of running the algorithms against the datasets selected in the previous step. The same would happen in the third step (Evaluation), where the system would compute the metrics selected in the first step and compare with the results of the execution in step two. Finally in step four (Analysis) the researcher would be offered the possibility of looking at the final results thus having the opportunity to understand where and why the algorithms have failed or succeeded.
5. CONCLUSIONS

In this paper we have presented the challenge of properly evaluating Non-Intrusive Load Monitoring technology. We introduced four datasets NILM datasets, two of them specifically created for performance evaluation purposes. We then discussed in more details the problems behind NILM performance evaluation and introduced our own hardware-software framework for creating NILM datasets.

We have also argued that although having proper datasets is of fundamental importance, one cannot neglect how these large data sets will be accessed and used by other researchers. Therefore in the last chapter we present our research plans where we envisioned a futuristic framework in which researchers will be able to test and evaluate their algorithms with very little effort.

We are currently planning a long-term deployment of our platform in one or two houses for a couple of weeks in order to obtain some validation. After that we will be looking at ways to automatically annotate the waveforms with data from the individual appliance usage as well as activities from the householders.

Once this is done we will be looking at the best possible ways to make this dataset public to other researchers.

6. ACKNOWLEDGMENTS

This research has been partially supported by the Portuguese Foundation for Science and Technology (FCT) grant SFRH/BD/77856/2011 and the SINAIS project (CMU – Portugal) CMU-PT/HuMach/00064/2008.

7. REFERENCES


