CALIFORNIA ENERGY COMMISSION

Final Report Compilation for

Equipment Scheduling and Cycling

TECHNICAL REPORT

October 2003
P-500-03-096-A2

Gray Davis, *Governor*

CALIFORNIA ENERGY COMMISSION

Prepared By:
Architectural Energy Corporation
Vernon A. Smith
Boulder, CO

Massachusetts Institute of Technology
Leslie Norford
Steven Leeb
Cambridge, MA

CEC Contract No. 400-99-011

Prepared For:
Christopher Scruton
Contract Manager

Nancy Jenkins
PIER Buildings Program Manager

Terry Surles
PIER Program Director

Robert L. Therkelsen
Executive Director

Acknowledgements

Les Norford, Steve Leeb, K. Douglas Lee, Peter Armstrong, and Chris Laughman with MIT conducted the research. Lanny Ross with Newport Design Consultants provided field support.

Preface

The Public Interest Energy Research (PIER) Program supports public interest energy research and development that will help improve the quality of life in California by bringing environmentally safe, affordable, and reliable energy services and products to the marketplace.

The Program's final report and its attachments are intended to provide a complete record of the objectives, methods, findings and accomplishments of the Energy Efficient and Affordable Commercial and Residential Buildings Program. This attachment is a compilation of reports from Project 2.2, *Equipment Scheduling and Cycling*, providing supplemental information to the final report (Commission publication #P500-03-096). The reports, and particularly the attachments, are highly applicable to architects, designers, contractors, building owners and operators, manufacturers, researchers, and the energy efficiency community.

This document is one of 17 technical attachments to the final report, consolidating three research reports from Project 2.2:

- *Detection of HVAC Equipment Turn On-Turn Off Events with Non-Intrusive Electrical Load Monitoring (Aug 2001)*
- *Fully Automated Analysis of Equipment Scheduling and Cycling (Oct 2001)*
- *Development of a Functioning Centrally Located Electrical-Load Monitor (May 2003)*

The Buildings Program Area within the Public Interest Energy Research (PIER) Program produced this document as part of a multi-project programmatic contract (#400-99-011). The Buildings Program includes new and existing buildings in both the residential and the nonresidential sectors. The program seeks to decrease building energy use through research that will develop or improve energy-efficient technologies, strategies, tools, and building performance evaluation methods.

For the final report, other attachments or reports produced within this contract, or to obtain more information on the PIER Program, please visit www.energy.ca.gov/pier/buildings or contact the Commission's Publications Unit at 916-654-5200. The reports and attachments, as well as the individual research reports, are also available at www.archenergy.com.

Abstract

Project 2.2, *Equipment Scheduling and Cycling*.

Tracking electrical usage of HVAC system components, such as fans, pumps, chillers, and cooling towers, has traditionally required expensive submetering. This project used an innovative, single electrical monitor that can be installed at a motor control center or a building electrical service entrance to detect when multiple pieces of equipment turn on and off. The electrical monitor, called a Non-Intrusive Load Monitor (NILM), has been under development at MIT for over a decade. It is essentially an inexpensive computer that samples the electrical voltage and current at a very high frequency and uses software techniques to recognize the signature of each piece of electrical equipment.

- Motor loads such as reciprocating chillers, fans, and pumps can be identified and tracked using a single NILM.
- Loads turning on or off outside of the normal schedule can be detected, as well as abnormal cycling.
- The MIT research team achieved a major breakthrough by developing a method to detect certain loads driven by variable speed drives. In addition, they made considerable progress in automating the detection and tracking algorithms.

This document is a compilation of three technical reports from the research.

Detection of HVAC Equipment Turn On-Turn Off Events with Non-Intrusive Electrical Load Monitoring
Deliverable 2.2.1(b) Report

Submitted to Architectural Energy Corporation

Under the

California Energy Commission's

Public Interest Energy Research (PIER) Program

Energy-Efficient and Affordable

Small Commercial and Residential Buildings

California Energy Commission Contract 400-99-011

L.K. Norford, H. Xing and D. Luo*
Massachusetts Institute of Technology
August 9, 2001

* Dr. Luo is now with the Carrier Corporation, Syracuse, NY

Table of Contents

Detection of HVAC Equipment Turn On-Turn Off Events with Non-Intrusive Electrical Load Monitoring

1. Introduction

This deliverable is an interim report of field tests of MIT's centralized or non-intrusive electrical load monitor (NILM). The tests were performed at the Iowa Energy Center's Energy Resource Station, on the campus of the Des Moines Area Community College in Ankeny, Iowa. The report assesses the NILM's ability to detect turn-on and turn-off events for HVAC equipment, including pumps, fans, and a chiller, and to estimate equipment-level energy consumption. This is a final report and succeeds an interim report submitted on March 12, 2001.

This report begins with a brief description of the test building and the NILM. An extension to the NILM's sampling algorithm is then presented, which has enabled the NILM to reliably detect the cycling operation of the reciprocating chiller. Next, the report describes the detection of the start-up of variable-speed drive fans and presents an empirically determined lower limit on the size of detectable loads. The report shows that the extended sampling algorithm successfully detected a few artificially introduced faults in the air handler and the chiller, using electrical on-off events as the sole measurement. Finally, the report compares changes in power as recorded by the NILM and by submeters, shows NILM-derived estimates of energy consumption, and offers brief conclusions.

More information about the NILM and the test building can be found in the references listed at the end of this report.

2. Test Building

Tests were conducted in a one-story building that combines laboratory-testing capability with real building characteristics and is capable of simultaneously testing two full-scale commercial building systems side-by-side with identical thermal loads. The building is equipped with three variable-air-volume air-handling units: AHU-A, AHU-B, and AHU-1. AHU-A and B are identical, while AHU-1 is similar but larger to accommodate higher thermal loads. AHU-A

and AHU-B each serve four test rooms with controllable lighting loads and heaters to simulate occupants and equipment. AHU-1 serves classrooms and offices occupied by researchers who operate the building. The major components of the AHUs are the recirculated air, exhaust air, and outdoor air dampers; cooling and heating coils with control valves; the supply and return fans; and ducts to transfer the air to and from the conditioned spaces. Air from the AHUs is supplied to VAV box units, each having electric or hot-water reheat. Cooling is provided by a single on-site two-stage reciprocating chiller and by the campus chilled-water loop. Figure 1 diagrams the building plan, the chilled-water circuit, and an air-handling unit. The AHUs and the chiller are well instrumented with electrical-power meters and with sensors for all controlled variables in common HVAC systems.

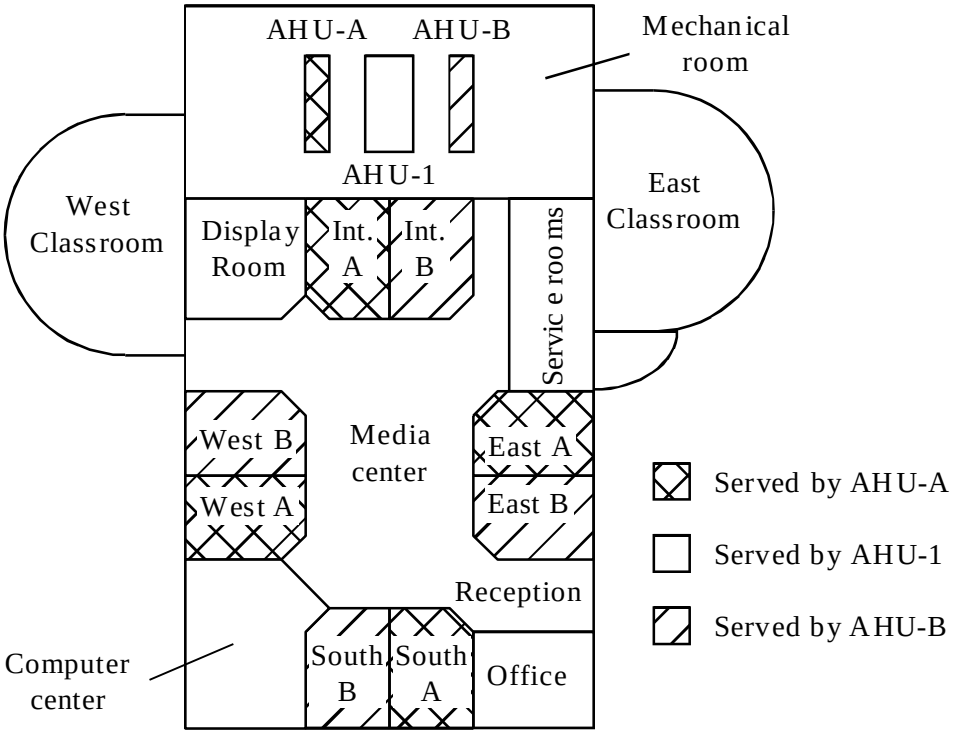


Figure 1a. Schematic of the test building

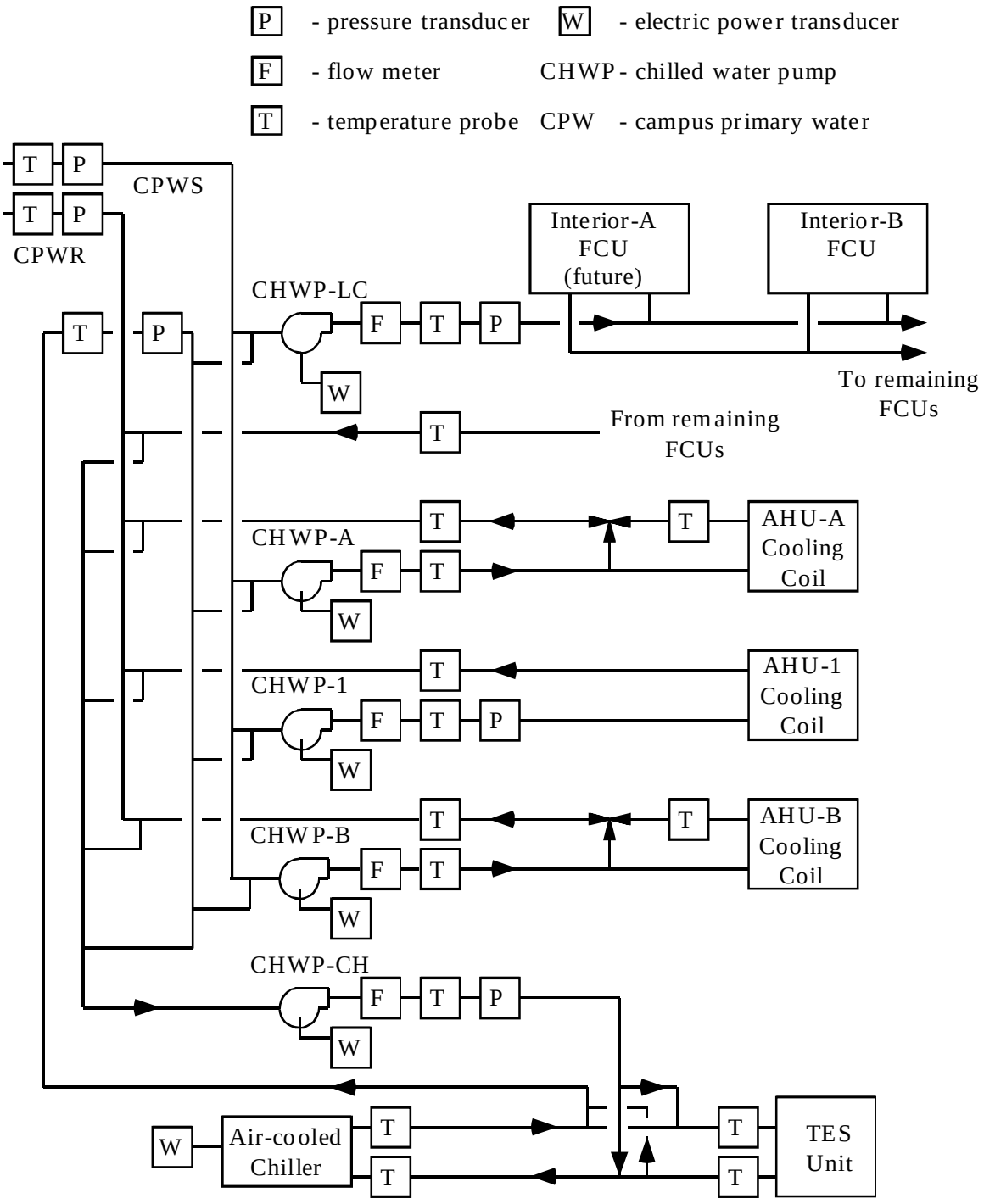


Figure 1b. Schematic the chilled water flow circuit

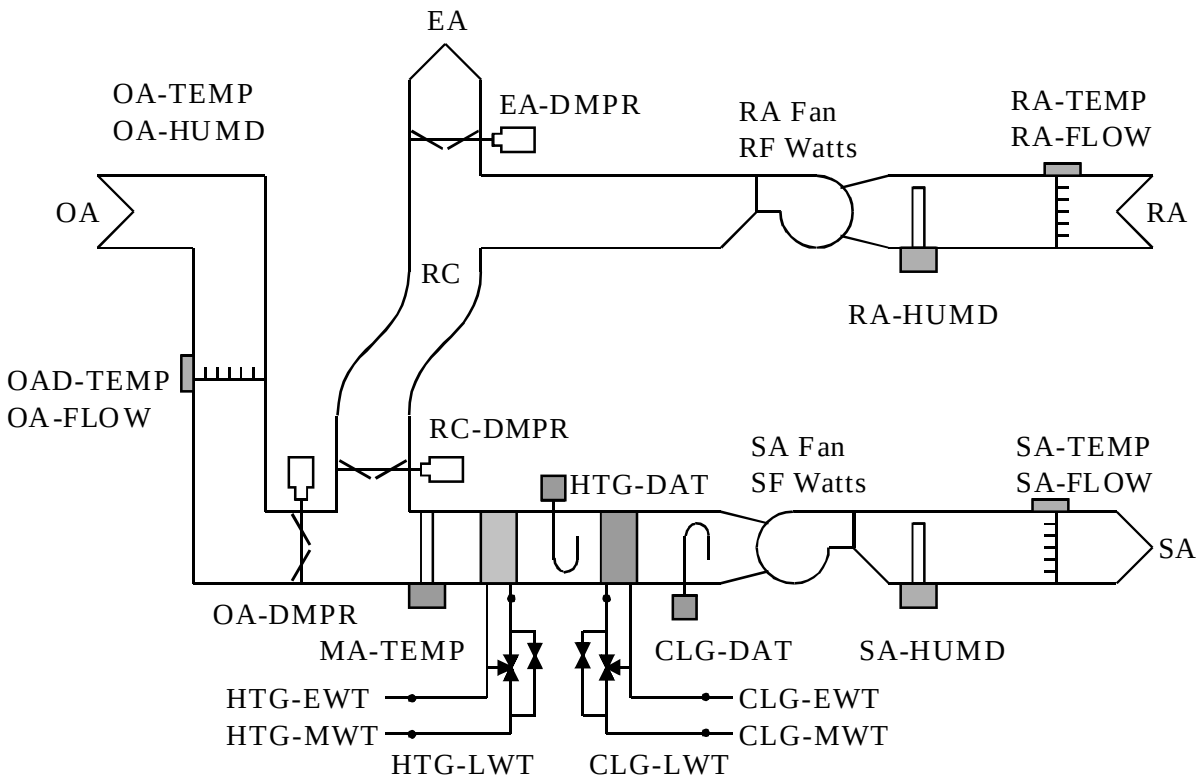


Figure 1c. Schematic the air handling units

A motor-control center provides electrical power for the six fans and 10 pumps, with individual capacities ranging from 0.5 to 5.0 kW and listed in Table 1. The pumps in AHU-A and B are equipped with constant speed motor drives while the other pumps and all the fans have variable speed motor drives. The total power input of the motor control center is recorded by a NILM meter named Remote-2 and each component is monitored with a power submeter for the tests. The cooling load is handled by a two-stage air-cooled reciprocating chiller with a nominal input of 10 kW. The power input of the chiller used for this test is supplied through the building's general electricity distribution circuit, which is monitored by another NILM called Remote-1.

The NILM's detection algorithm, an extension of the Generalized Likelihood Ratio (GLR), is based on the probability distribution of data points in a detection window relative to a distribution of data in a window used to establish a pre-event mean signal. If there is a high probability that data in the detection window differ from those in the pre-event window, an on/off event is flagged. The reset window shown in Figure 2 is used to purge the data after an event has been detected.

Table 1. Power capacity and motor types of fans and pumps served by the motor control center.

Equipment	Power capacity (kW)	Motor drive speed
Supply fan 1	5.0	Variable
Return fan 1	2.0	Variable
Supply fan A	5.0	Variable
Return fan A	2.0	Variable
Supply fan B	5.0	Variable
Return fan B	2.0	Variable
Hot water pump 1	0.75	Variable
Hot water pump A	0.4	Constant
Hot water pump B	0.4	Constant
Hot water pump LA	0.4	Variable
Hot water pump LB	1.0	Variable
Chilled water pump 1	1.0	Variable
Chilled water pump A	0.4	Constant
Chilled water pump B	0.4	Constant
Chilled water pump LC	1.0	Variable
Chilled water pump CH	1.0	Variable
Total	32.25	

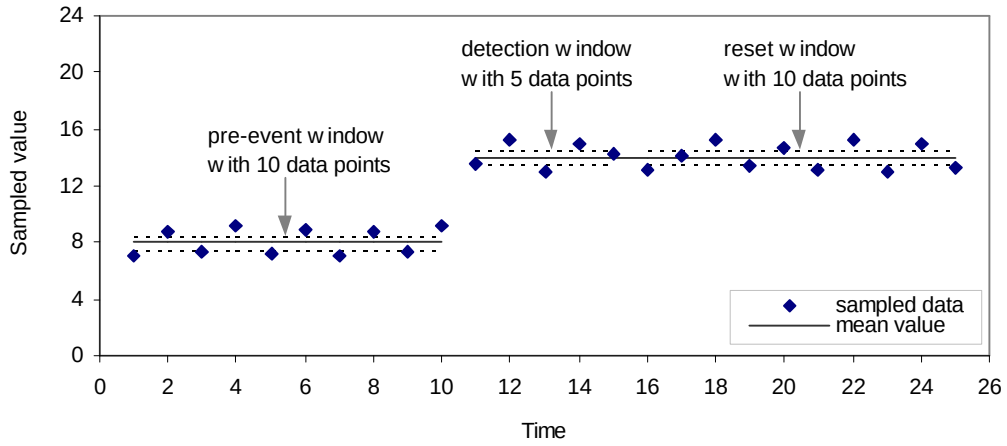


Figure 2. Finite Moving Average windows used with the NILM detection algorithm.

3. Development of Multi-Sampling Rate Non-Intrusive Electrical Load Monitor

The sampling rate of the data fed to the detector has been found critical in keeping track of the events in a system. In this section, after comparing the detection output of the NILM for single- and multi-rate sampling, a detector for change detection with multiple sampling intervals is developed and discussed with power data from the test building.

3.1 Change detection by multi-rate sampling of the total power data

On/off events occur in a wide range of time intervals in HVAC systems. In order to find as many events as possible, a short sampling period should be used to avoid missing of events due to sampling. This is because a shorter sampling interval, or a faster sampling rate, can help to distinguish closely separated events. However, a fast sampling rate also tends to result in more false alarms because the calculation is more susceptible to high-frequency transients and disturbances. Moreover, a too fast sampling rate may cause more than one alarm for a single event, especially for some relatively slow on/off events, e.g., a supply fan with a variable speed motor drive. On the other hand, a long sampling interval, or, a slow sampling rate, is likely to miss changes with short time duration, such as the startup of a constant-speed water pump, and sometimes mix up closely separated events. Therefore, a reasonable range of sampling intervals should be determined before applying the NILM's detection algorithm to a data series.

3.2 Order of magnitude analysis of the sampling interval

In general, the sampling interval should be shorter than the time lapse between two most closely spaced events. Such a limit can be obtained by examining the total power data at a fast enough sampling rate or by observing the signals from the control system, if possible. For example, in the tests with a real HVAC system, the shortest time period between two consecutive events has been found to be about 1 second by analyzing the building's total power data. Figure 3 shows the whole-building power data during a typical day. The effect of sampling rate on the data pattern fed to the detector is illustrated in Figure 3. All the data were collected with the Remote-1 NILM, from which data were extracted at 24 Hz. Data are plotted here with a sampling interval of 10 seconds due to limitations of the plotting software.

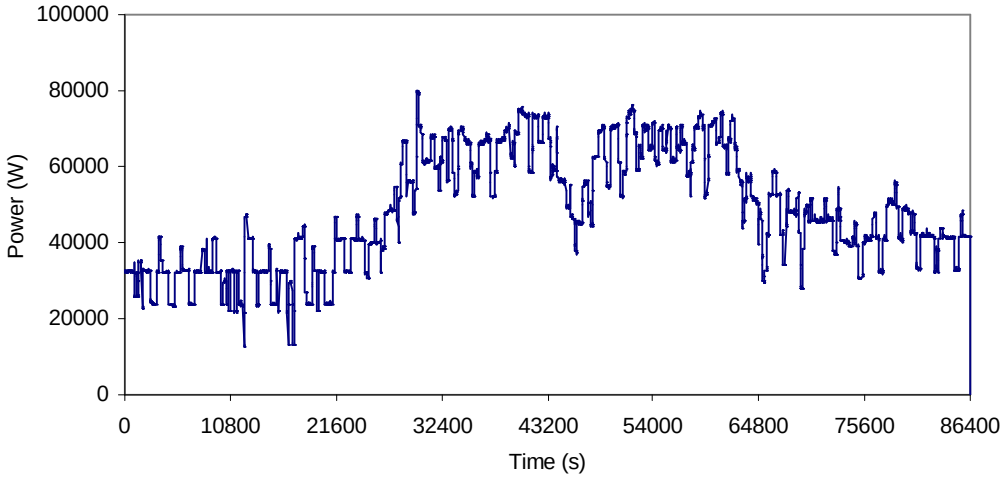
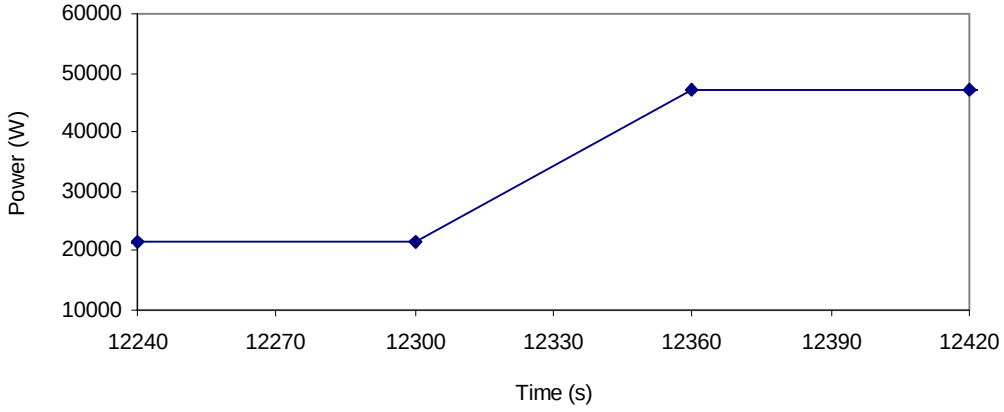
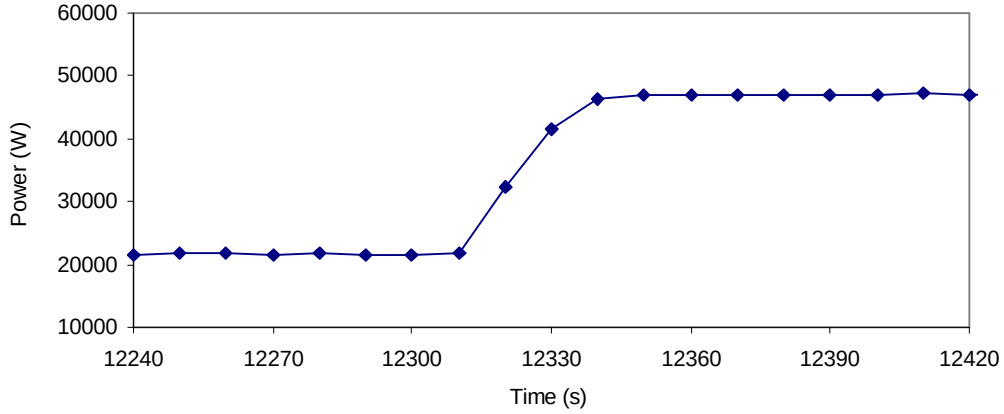


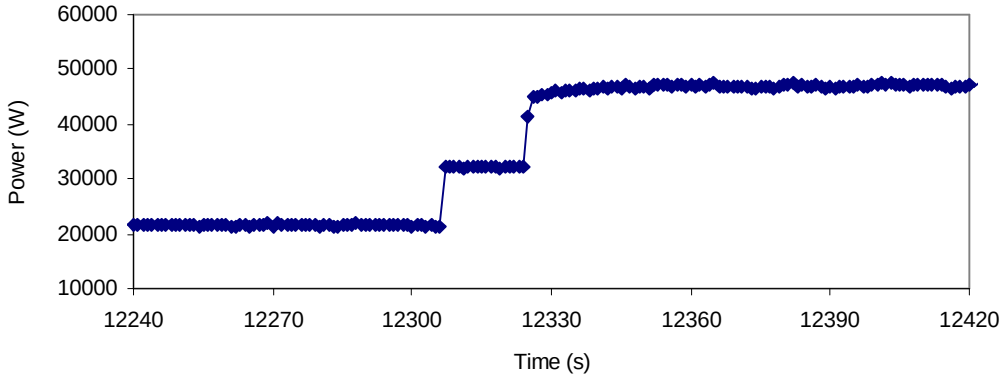
Figure 3. Whole-building electrical power, sampled at 24 Hz, plotted with 0.1 Hz (one data point every 10 seconds).



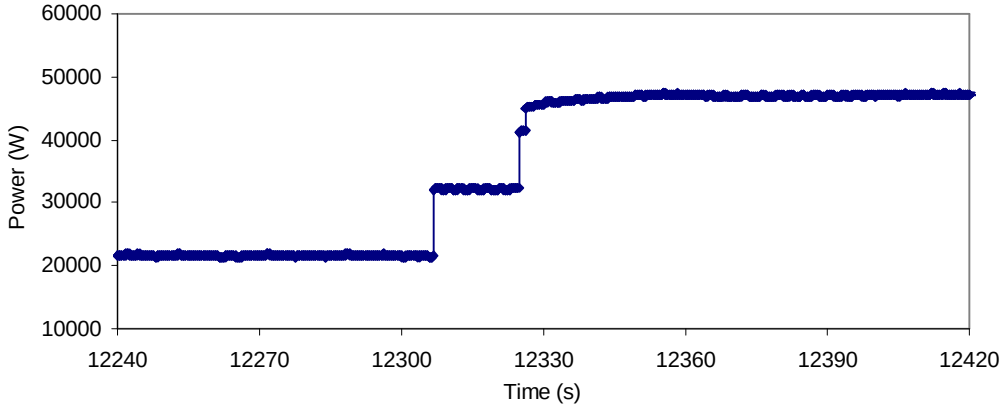
(a)



(b)



(c)



(d)

Figure 4. Effect of different data sampling intervals on detectability of rapidly occurring equipment start-up events.(a). 60 seconds; (b). 10 seconds; (c). 1 second; (d). 0.125 seconds.

There were more than 100 on/off switches of equipment in the system during the displayed 24-hour period. During a selected period of 30 seconds (12300-12330) three components were turned on sequentially. The power data series with sampling intervals of 60 10, 1, and 0.125 seconds are shown sequentially in Figure 4.

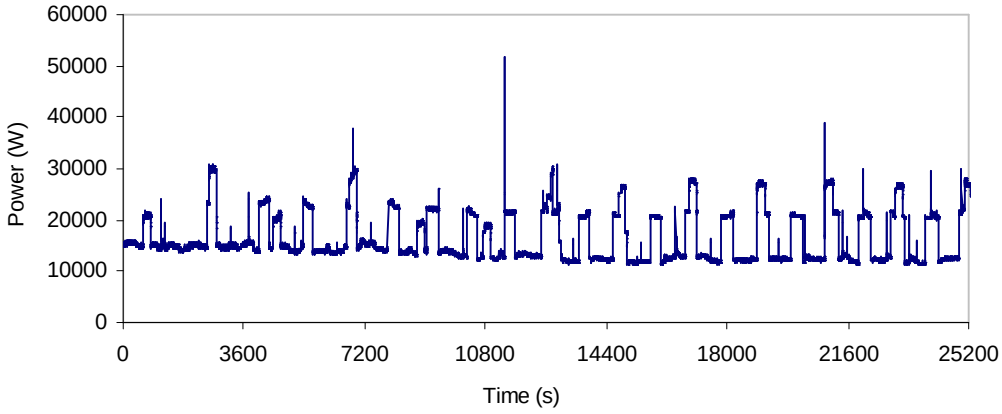
As displayed, if an interval of 60 seconds is used to sample the data, then these three turn-on events are taken as only one step change. With the sampling interval reduced to 10 seconds, the events are only sampled as one point each, which still cannot be used as a solid proof for a step change. When the sampling interval is reduced by another order of magnitude, i.e., to 1 second, then the first event can be clearly recognized by eye and by the detector as well. This decrease of the sampling interval makes the second event discernible to eye, but not to the detector, which needs an appropriate length of data window to build up the sufficient statistic. The second event remains ambiguous to the detector until the sampling interval is reduced to 1/8 second, i.e., a sampling rate of 8 Hz. However, tests to detect the cycling of a reciprocating chiller in the given system showed that the false alarm rate increases drastically when the sampling interval dropped below 1 second. Based on the tests and analysis of the sampling

effect on the detection quality, the appropriate lower limit of the sampling interval for GLR detection has been found to be of the same order of magnitude as the shortest time duration of the on/off events or the shortest time interval between two consecutive events of interest, whichever is shorter. The upper limit should be longer than the lower limit by about 1~2 orders of magnitude. In this case, the lower and upper limits are 1 and 30 second(s) respectively. In addition to a reduced false alarm rate, another significant advantage of using such a range compared to the higher sampling rate of 8 Hz is the more acceptable execution time of the detection. This is more meaningful when the automatic detector is developed with multiple sampling intervals as discussed later. The running time of the detector, as implemented on a 233 MHz personal computer, was reduced from 90 minutes to less than 5 minutes when the lower limit of the sampling intervals increased from 0.125 seconds to 1 second for the test with one day's data, making the detector more desirable for on-line detection.

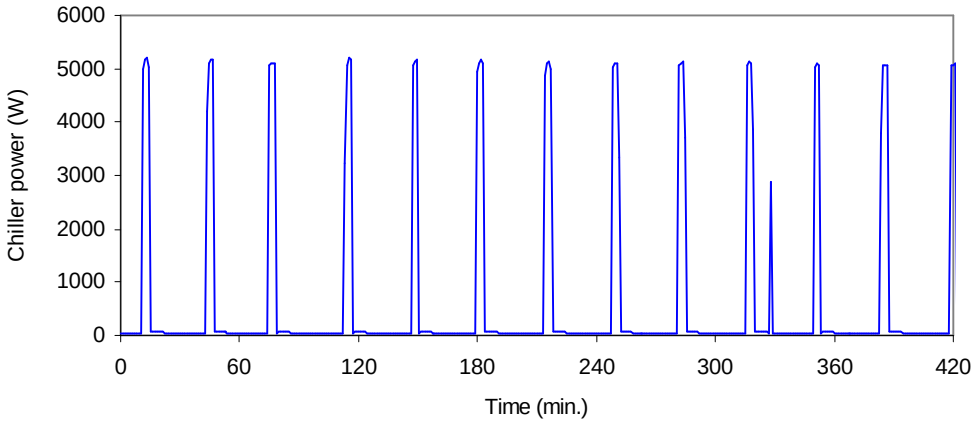
One problem with such a moderate sampling range is that some very closely separated events might be missed or misinterpreted, such as the above event with an interval of less than two seconds. But such events seem to happen by coincidence and are rarely seen in common HVAC systems because the switches of major electrically-driven components are typically staggered to alleviate electrical surge. In practice, there is generally an interval of not less than 30 seconds even for some coupled on/off equipment, such as the supply and return fans associated with the same air handler.

3.3 Detection of on/off changes with single sampling rate

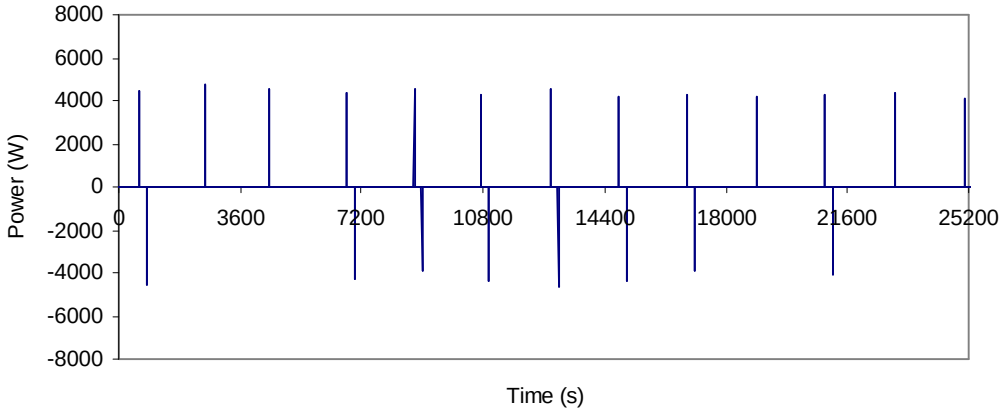
Because the data patterns supplied to a detector and consequently the 'visibility' of an event to the detector vary significantly with the data sampling frequency, the sampling interval must be carefully determined based on the above order-of-magnitude analysis to achieve desirable quality of the detection. Many tests have been conducted to search for the optimal sampling rates to detect all the on/off events in an HVAC system. In the following analysis, one example of such tests is demonstrated with the power data from a test building over six days. The tests were started with the detection of the on/off cycling of a reciprocating chiller from the building's total power data. Figure 5 illustrates the first seven hours of the total power data on a single day and the GLR detection output for the chiller's on/off switches with different sampling rates. From the original power data shown in Figure 5(a), it is difficult to visually distinguish the chiller cycles from the whole-building power signal. The first chiller cycle, aligned with data from an electricity submeter used to check the GLR results, is readily discerned. Most of the following switches are obscured by noises or other events.


60000
50000
40000
30000
20000
10000
0
Power (W)
0
3600
7200
10800
14400
18000
21600
25200
Time (s)

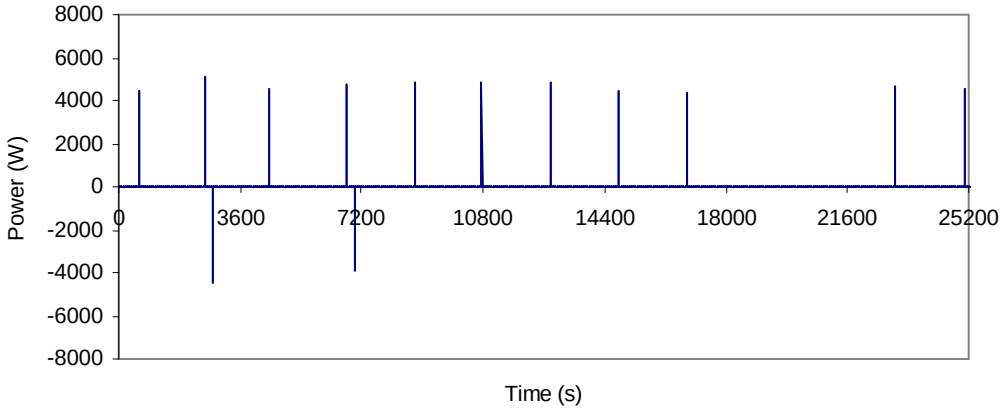
(a)


6000
5000
4000
3000
2000
1000
0
Chiller power (W)
0
60
120
180
240
300
360
420
Time (min.)

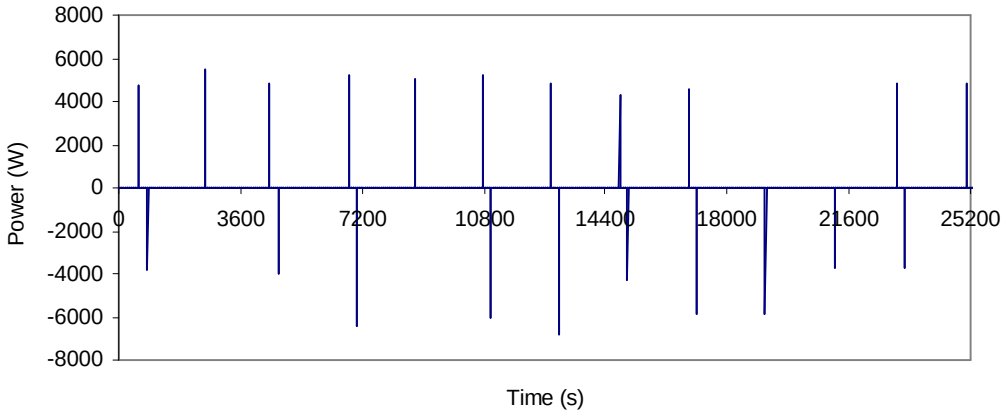
(b)


8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
Power (W)
0
3600
7200
10800
14400
18000
21600
25200
Time (s)

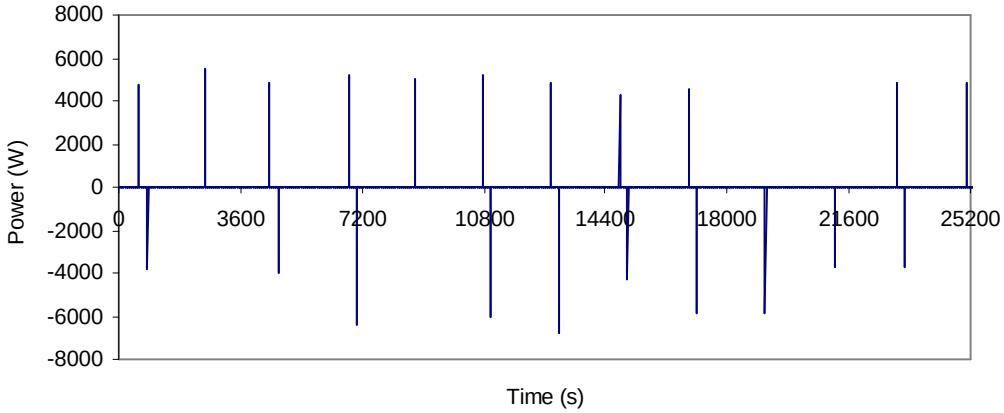
(c)


8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
Power (W)
0
3600
7200
10800
14400
18000
21600
25200
Time (s)

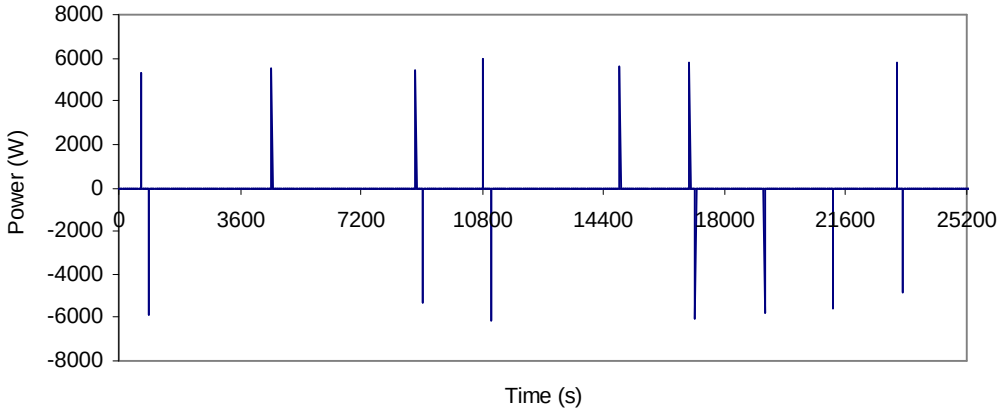
(d)


8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
Power (W)
0
3600
7200
10800
14400
18000
21600
25200
Time (s)

(e)


8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
Power (W)
0
3600
7200
10800
14400
18000
21600
25200
Time (s)

(f)



(g)

Figure 5. Demonstration of the detection of chiller on-off cycles from the total power data of a test building with different sampling intervals. (a). total power data, 00:00-07:00, from the Remote-1 NILM; (b). chiller power data from a submeter during the same seven-hour period; (c) ~ (g). detection output of the chiller's on/off switches with sampling intervals of 1, 2, 5, 10, and 20 seconds respectively.

All other sampling intervals between 1 and 60 seconds produced results similar to those shown above. Apparently, with a single sampling interval, the detector is not able to find all the on/off switches correctly as shown in the submetered data. However, it can be seen from the above plots that all the events can be found by matching the on's and off's among outputs with different sampling intervals. In this case, on/off matching among the outputs with the sampling intervals of 1, 2, and 5 seconds identifies all the switches without any false alarms or missing events. Other combinations of sampling intervals may yield the same results, such as the combination of 1, 2, and 10 second(s). Similar detection patterns with different sampling intervals have been found with the remaining five days' data. This indicates that the quality of detection can be significantly improved if the detection can be performed by properly integrating the outputs among different sampling rates.

Difficulties with a single sampling rate arise because, first, different components in a system require different time durations for the turn-on switches. Second, for given sizes of the pre- and post-event windows, the ideal sampling interval should be longer than any of the on/off transients and shorter than the time interval between two on/off events. Unfortunately, this condition cannot always be met in commercial applications where VSD drives and load controls are widely used, which lead to more gradual changes in the system's total power series. As a consequence, an on/off transient process may last longer than the interval between two events. Moreover, as can be seen from Figure 5(a), the quality of the total power data makes it difficult to isolate changes under different conditions with a single sampling rate even for the same equipment.

3.4 Detection of on/off changes by automatic matching among multiple sampling intervals

In this detector, the data series is sampled with discrete integer intervals between the lower and upper limits and then supplied to the GLR detector. The detection output with each interval is sorted by the magnitude of the changes, assigned to the equipment with that specific magnitude, and matched with the outputs with other sampling intervals. Figure 6 displays the output of this automatic detector for the data shown in Figure 5 without missed events or false alarms.

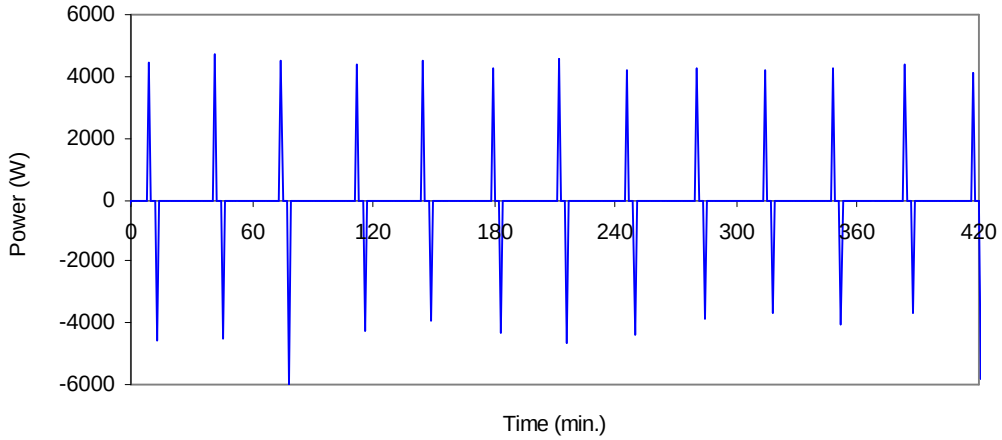
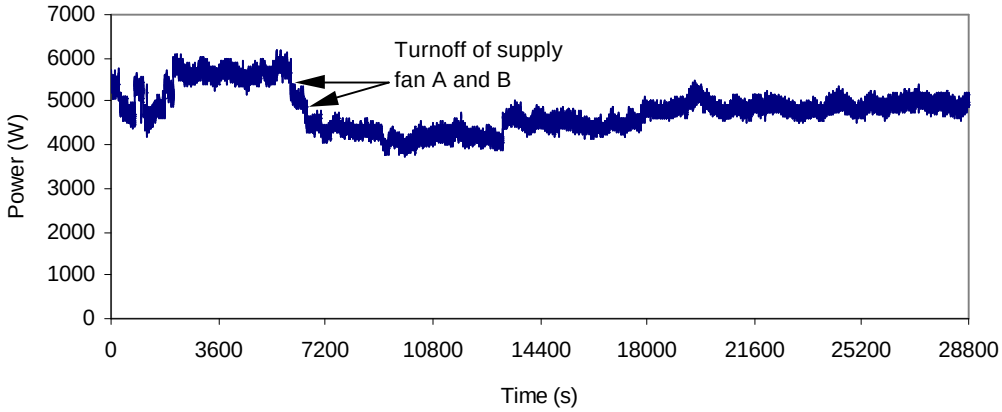


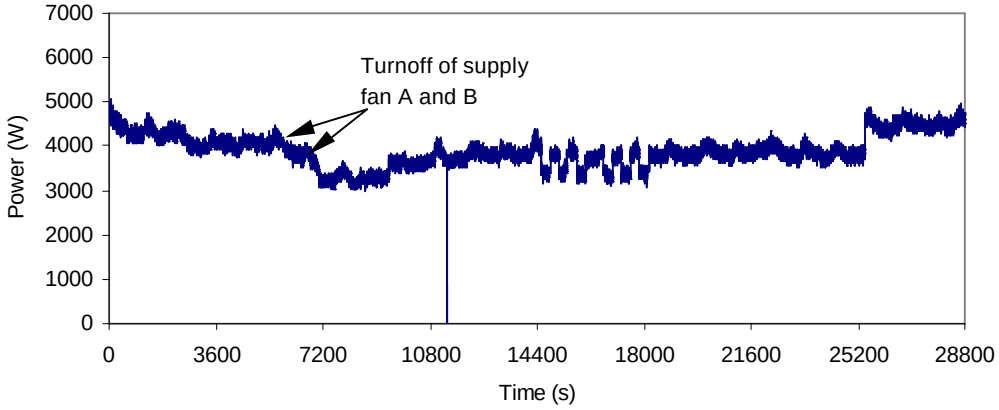
Figure 6. NILM detection output for the chiller on/off switches with the building's total power data by automatic matching among different sampling intervals: 1, 2, 3, 4, 5, 10, 20, and 30 seconds.

Additional data filtering criteria, if available, may be applied for specific equipment to reduce the false alarm rate, especially with the presence of other components with close magnitude of change. For the chiller in the test building, for example, these criteria include minimum off-time between cycles, which is typically set within the chiller controls to prevent unnecessary equipment cycling, and minimum expected on-time. Such limits have been incorporated in the detector designed for the given system.

Therefore, data sampling with multiple intervals not only sharpens the abruptness of changes, which makes the event clearer for the NILM detector to "see", but also helps to distinguish abrupt on/off cycles among a gradually changing process which for example, can be a slow startup of a VSD fan. With only one sampling rate, such obscured changes cannot be found without false alarms. As shown in Figure 7, fan turn-off events as recorded by the NILM logger vary in their apparent abruptness, as influenced by changes in other loads. All four events were successfully detected with multi-rate sampling but some were always missed with any single-rate detection.



(a)



(b)

Figure 7. Demonstration of the various abruptness of the turn-off transient of the same fans in two different days. Multi-rate sampling produces at least one data stream with the 'visible' or ' abrupt' event for the GLR detector despite of the noise variations in the data environment, which could not be achieved with any single rate. (a). total power data of the HVAC system over 20 hours beginning at 8:07 p.m., showing the turnoff of the fans among multiple events; (b). total power of the HVAC system, demonstrating the obscured abruptness of the turnoff events in the total power.

3.5 GLR Training Guidelines

Guidelines for training the parameters involved in the GLR detection are summarized as follows.

1). Record electrical power for the circuits monitored by the data logger for one day under typical operating conditions. The sampling rate should be between 1 and 10 Hz for common HVAC systems;

2). Locate the events from the abrupt changes in the total power data and estimate the fastest and the slowest events;

3). Determine the sampling rates for detection. The base sampling rate will be used as the fastest sampling rate if multi-rate sampling is employed. Therefore, power data sampled at this rate should be used to discern each event of interest by eye. The lower limit of the sampling rate should be able to pick up the longest duration of events of interest. Then other sampling rates can be decided between these two limits with 2~5 rates selected within each order of magnitude;

4). Determine the window lengths. The length of the detection window should contain at least two data points. It should not be longer than the interval between two consecutive events and never shorter than a disturbance. The length of the pre-event window is 2~4 times longer than the detection window. A third window with the same length as the pre-event window is used for data reset;

5). Calculate and estimate the maximum of the standard deviation of the power data as a fraction of the current total power data, *f*. Note that periods with abrupt changes should be avoided in the calculation of *f*. The lower limit of the standard deviation should be set one magnitude lower than the estimated value *f*.

6). Estimate the threshold for the detection statistic. Without knowledge about the minimum expected change V_m in the total power data, a reasonable base value for the threshold is $1/f^2$. If V_m is known in advance, then the training of the threshold can be started with $(s/V_m)^2$, where s is the average standard deviation of the samples from the training data. The threshold for the GLR can be trained by adjusting this value until all events of interest can be seen by the detector, with a minimum number of false alarms. The lower limit of the GLR can be started with $1/(10*f)^2$ or $s/(10*V_m)^2$.

Note that the starting values for the training of the parameters are the estimated thresholds, not the exact value. Moreover, exact values are not expected for the above parameters due to the statistical properties of the detection method. Slightly different combinations of these parameters may yield equally acceptable detection output.

The detector developed is applied on the test cases in this report. Because the rules for the window lengths are for common HVAC systems, similar window lengths can be used in different applications with minor adjustments and the major training task is usually reduced to determination of the thresholds for the sufficient statistic and the standard deviation.

The multi-sampling rate GLR works in a way that for each discrete integer-sampling interval between the upper and lower limits, the data series is sampled with that interval and then supplied to the GLR detector. The detection output for each interval is sorted by the magnitude of the changes, assigned to the equipment with that specific magnitude, and matched with the outputs with other sampling intervals.

4. Detection Of Turn-On and Turn-Off Events

4.1 Detection of Start-Up of Fans with VSDs

The start-up of fans with variable speed drives usually takes several minutes. The multi-sampling rate NILM at very slow speed is used to detect the start-up event. Detection and analysis were conducted in the following steps.

a) Test based on submetered data

The submetered data for AHU-A SF, AHU-A RF, AHU-B SF and AHU-B RF were recorded and added up to form a single signal, SUM. All the five signals are plotted for four typical days, shown in Figure 8. The return fan power levels are the smallest and the supply-fan powers are an order-of magnitude larger.

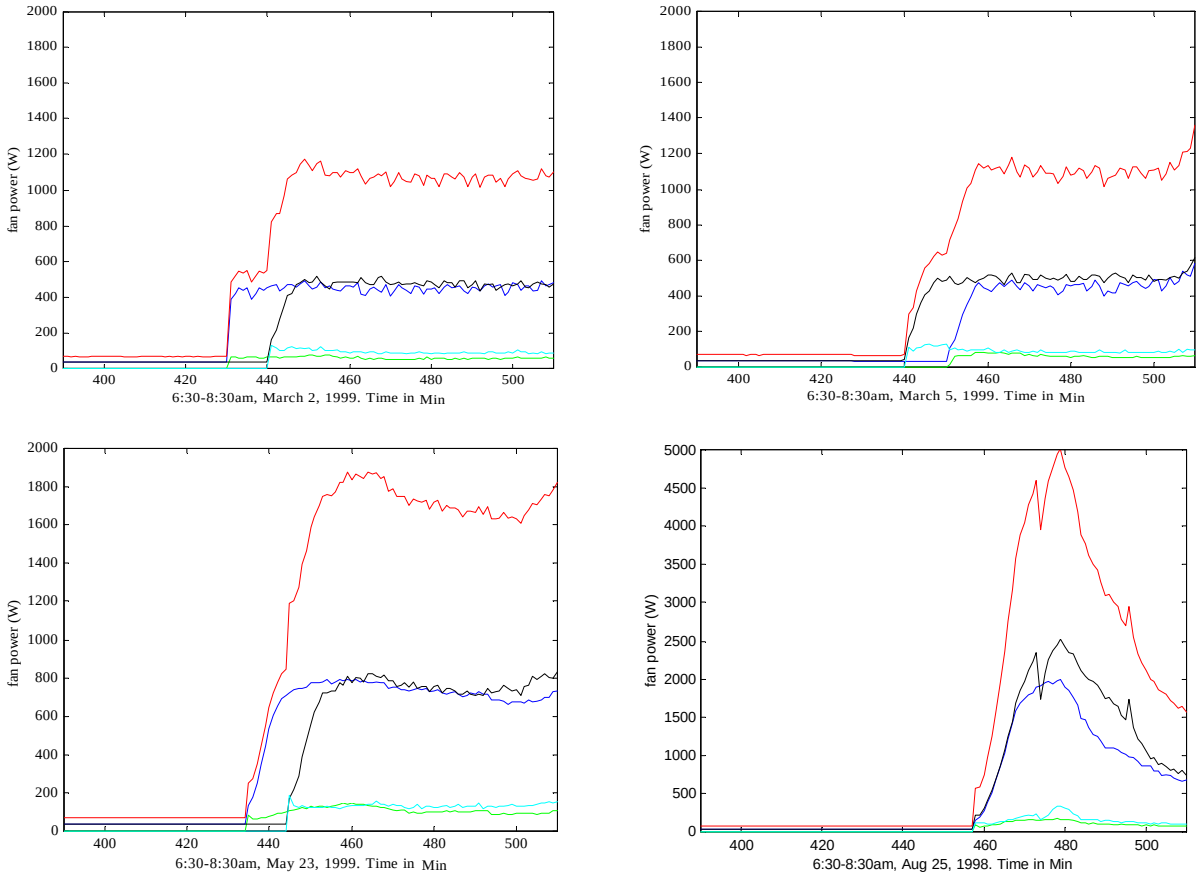
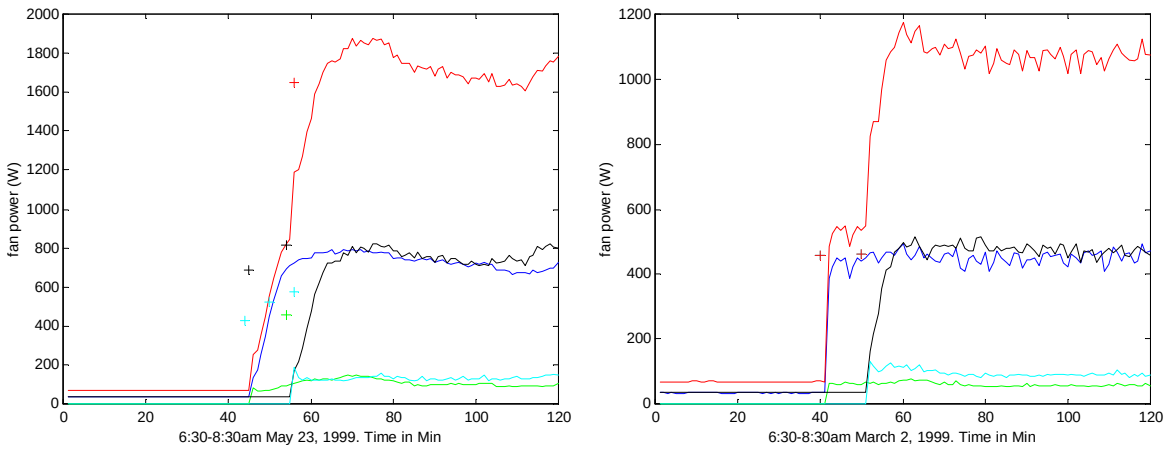


Figure 8. Submetered data of supply fans and return fans for AHU-A and AHU-B during the start-up period. a) 03/02/99; b) 03/05/99; c) 05/23/99; d) 08/25/98.

As we can see from the power plots, the supply fan and the return fan for the same AHU are turned on at the same time. The power usage of the return fan is much smaller than that of the supply fan. Therefore it is hard to individually identify the SF and RF. Visually, it appears

possible to separate the two AHUs in Figures 8.a) and 8.b), but the start-up signals are not distinct in Figure 8 c). Fans for two AHUs started at the same time in Figure 8 d) and it would be impossible to tell which AHU is on from the aggregated power.

The multi-sampling GLR was applied to the aggregated submetered power to detect the start-up of VSD fans on May 23, 1999 and March 2, 1999, shown in Figures 9 a) and b) respectively. The symbol of '+' represents a detected change. The time when the change happens is indicated by the horizontal axis and the magnitude of the change by the vertical one. Since all the data are fan powers, the detected change is a detected fan start-up. In Figure 9 a), the detected changes in power levels vary with sampling rate, as shown by the different magnitudes of the '+' symbol. (The four '+' symbols of lowest magnitude are associated with one-second sampling intervals, the intermediate two '+' symbols are from three-second samples, and the largest change was from a five-second sampling interval.) The multi-sampling GLR was unable to reliably detect the VSD fan start-up in this case, in the absence of an *a priori* basis for choosing one sampling rate in lieu of another. In Figure 9 b), the symbols of '+' are the results of applying multi sampling GLR on the submetered data. Multi-sampling rates give consistent detection results. It can be seen that the symbols of '+' match the submetered fan start-ups in both timing and magnitude (the first '+' is the AHU-A supply fan and return fan start-up; the second '+' is that of AHU-B). Therefore, the multi-sampling GLR succeeds in this case.



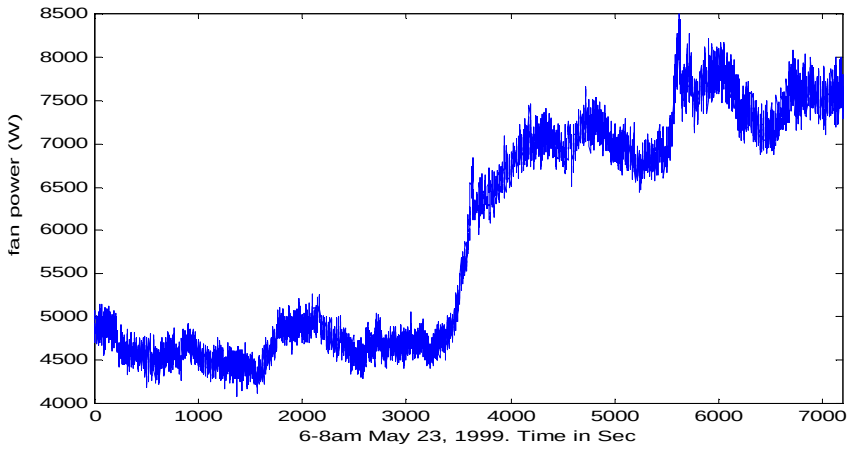
a) b)

Figure 9. Detect VSD fan start-ups based on submetered data. a) 6:30-8:30am, May 23, 1999; b) 6:30-8:30am, March 2, 1999.

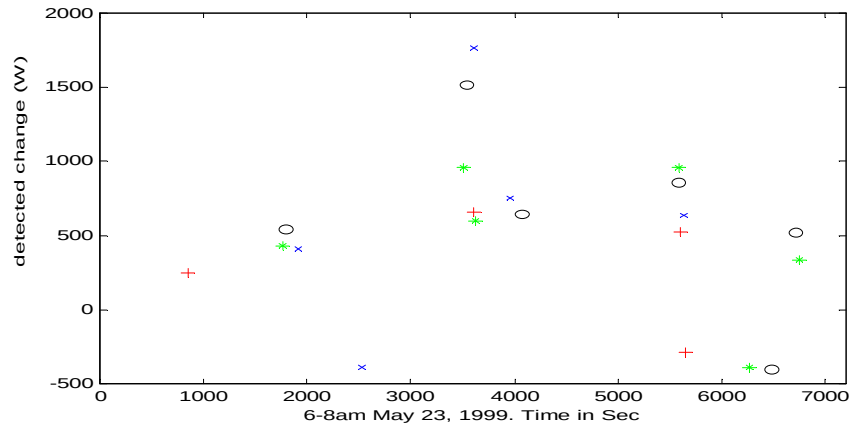
b) Test based on the NILM data

The NILM data of the two examples in Fig.9 were input to the multi-sampling GLR for another test of change detection. It was expected that NILM would not be able to detect the fan start-ups on May 23,1999 because the detection based on submetered data, which contain much less noise than NILM data, failed to consistently find and quantify the magnitude of power changes associated with the fan start-ups. Using the same logic, it was uncertain whether the NILM would be able to find the fan start-ups on March 2, 1999.

For May 23, 1999, the NILM data are shown in Figure 10 a) and the detection results based on the multi sampling GLR with sampling intervals of 5, 30, 60 and 120 sec are shown in Figure 10 b). The results of 5, 30, 60 and 120 sec sampling interval are indicated by '+', '*', 'o' and 'x' respectively. It can be seen that the results from different sampling rates are very different from each other in terms of time and magnitude. Given that the VSD fans' start-ups are a gradually changing process and the power curve is unknown, we cannot combine the multi-sampling results into a single estimate of power, as could be done with the chiller. Therefore, the VSD start-up could not be detected. This problem also happens to March 2, 1999, although we succeeded in detecting the VSD fan start-ups using the submetered data of that day.



a)



b)

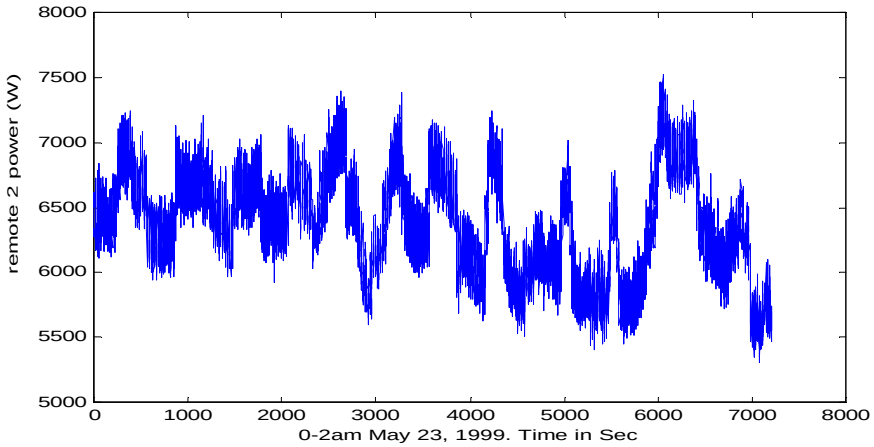
Figure10. Detect VSD fan start-up based on NILM data using multi sampling GLR, 6-8am May 23, 199. a) NILM data; b) Detection results: 5 sec by '+'; 30 sec by '', 60 sec by 'o' and 120 sec by 'x'.*

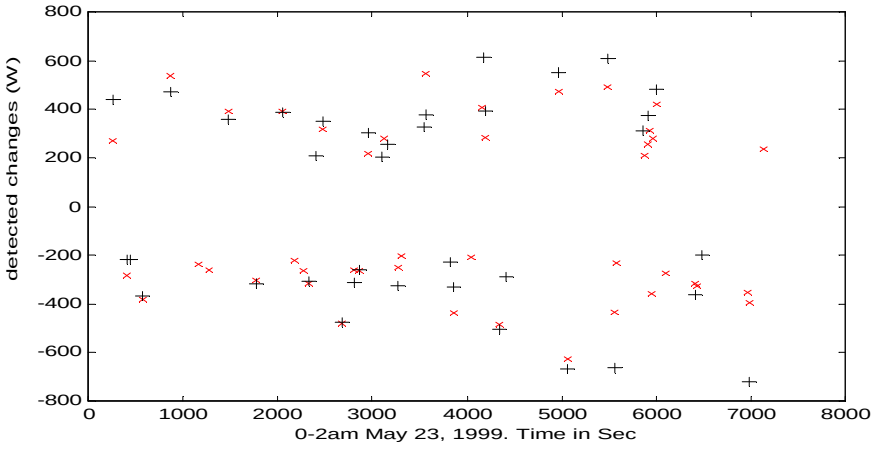
It was not very surprising that the multi-sampling GLR failed to produce consistent results for detection of VSD fan start-ups when using the NILM data. The start-ups of VSD fans usually take several minutes. This gradual change is hard to be detected by the change-of-mean

method, which is the essence of the GLR detector used in the NILM. In addition, NILM data have much more noise in compared to the submetered data. Therefore, more research is needed to consistently detect VSD components; this research will include further attention to sampling rates and, more importantly, use of both start-up signal patterns (without approximation as step changes) and use of harmonics higher than the fundamental (because VSDs generate higher-order harmonics).

4.2 Detection of On/Off Transients of Pumps

During most normal days, chiller-water pumps and hot-water pumps run constantly, so there are no on-off signals. On May 23, 1999, the chilled water pump, which has a power capacity of 0.4 kW, was turned on and off a couple of times from midnight to 2am. The detection of the on and off events for this pump was tested within this period of time in order to get some sense as to whether the pump power capacity is too small to be detected using NILM. As shown in Figure 11, the chilled water pump power capacity is too small. To detect its power change, the threshold has to be set so small that many other changes from noise get detected as well. It should be understood that since all the fans are turned off during the nighttime, the detectability of the pump at nighttime is only a lower bound of that during the daytime, i.e., if the pump's on/off events cannot be detected at nighttime, it would be impossible to detect them during the daytime.





b)

Figure 11. Detect chilled water pump on/off, 0-2am, May 23, 1999. a) remote 2 NILM data; b) detection results with different sampling rates: 'x' 5 sec sampling interval; '+' 10 sec.

4.3 Detection of On-Off Events in A Complex Load Environment

For the Iowa project, there are two operating conditions in general. One is the normal day condition in which there are not many on/off events except that the fans are turned on early in the morning and turned off late at night. (The on-site chiller can cycle frequently during the day but is not powered from the motor-control center and is not considered in the analysis of this section of the report.) The VSD fans' start-ups have been addressed before. The fan turn-offs will be discussed in a comprehensive example in this section. Therefore, the normal day condition is not going to be studied separately. Our focus will be the other operating condition called artificial adjustment, in which fans and pumps are manually turned on and off a couple of times during some specific days and the multi-sampling GLR was applied to the NILM data to detect these on/off events.

As an example of the tests conducted in this research, Figure 12 and Table 2 compare the detection output for changes of different magnitudes with single- and multi-rate sampling of power data from the motor control center of a test building against that measured by submeters. During this test period, there were 17 on/off switching events of two hot-water pumps, two supply fans, and two return fans among a total of six fans and 10 pumps served by the motor-control circuit monitored by the Remote-2 NILM.

The sampling interval used in the single-rate detection was 10 seconds while nine intervals, including 1, 2, 3, 4, 5, 10, 20, 30, and 60 seconds, were employed in the multi-rate case. These sampling intervals, selected after sufficient tests with the intervals from 0.125 through 1200 seconds based on power data from different systems and under various conditions, were found to be the most representative values for HVAC power systems. The two limits, 0.125 and 1200 seconds, are observed as the lower and upper limit for the duration of the on/off events in HVAC systems.

a) With single sampling interval

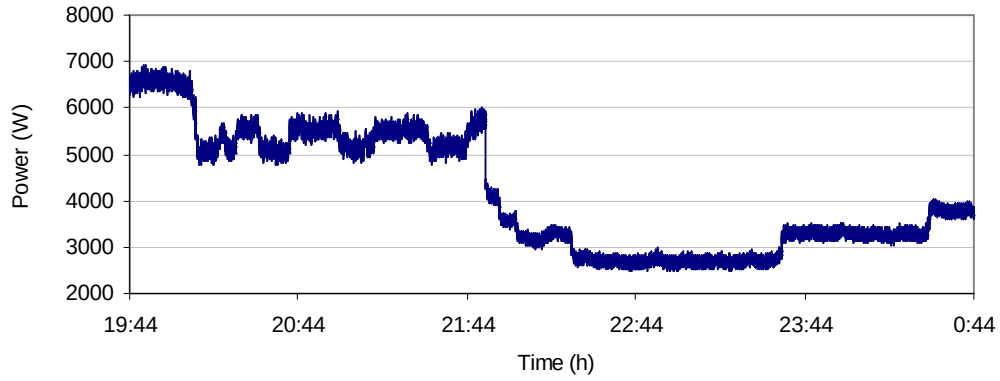
Of the 17 events, 15 were found with different levels of error and two were missed. From Table 2, it can be seen that the quality of the detection depends on the magnitude of the change relative to the total power as well as the current data trend. For example, at 20:34 with the turnoff of pump B, for which the power change of –244 W was less than 5% of the total power, the error was about 150%. At 21:55 with the turnoff of supply fan A, for which the power change of 1400 W was about 25% of the total power, the error was less than 0.3%. When the change in magnitude was very small, the detector was not able to find the event at all, such as the event at 21:56 when return fan A was turned off with a power change of –82 W.

In addition, the shorter the time interval between two changes, the more difficult to find the changes, especially the small ones. This is because the NILM's detector needs some time for the effect of the former event to die out in order to find the subsequent change. This is demonstrated by the output at 20:21 for the turnoff of pump A. With a magnitude of 275 W, the event was still missed due to the gradually changing property of the data caused by the previous change and the noise, as seen from Figure 12.

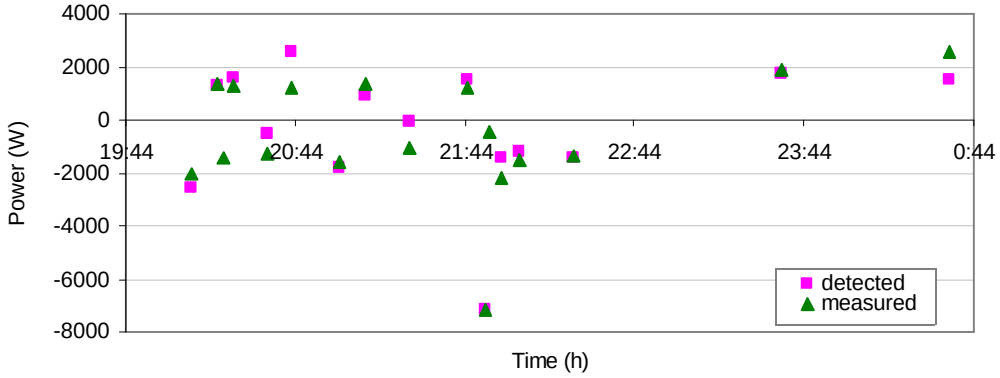
b) With multiple sampling intervals

With a single sampling interval, significant discrepancies of power changes were observed between the submetered data and that detected by the NILM. Sometimes, on/off events are undetectable with a single sampling rate. However, with further techniques like the multiple-sampling rates, the detector was significantly improved in terms of the detection quality, not only in the detectability of the events but also in the resolution of the output. This was tested by running the detector with multiple sampling intervals through the same data set. As shown in Figure 12 and Table 2, with the multiple sampling intervals, 16 out of the 17 on/off events were identified, one of which was undetectable with a single sampling rate that was implemented before. Moreover, the errors of nearly 70 percent of the events were significantly smaller with the multi-rate sampling than that with the single-rate sampling, while only one event at 23:40 showed considerably larger errors with multi-rate samples, as listed in Table 2.

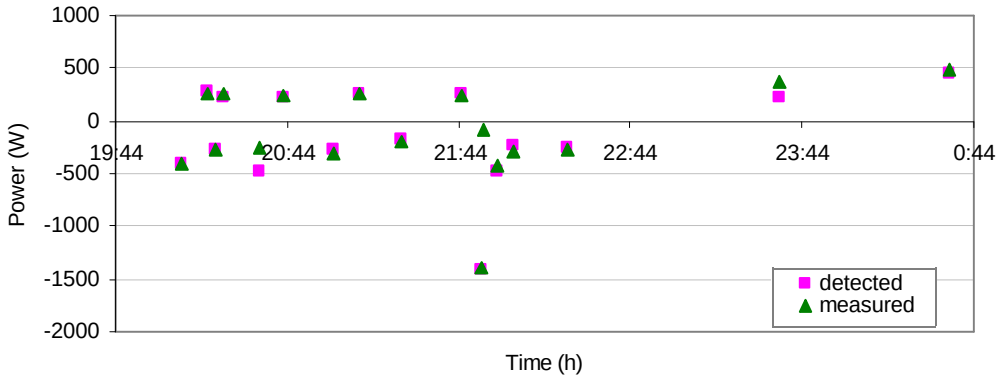
Under some circumstances, the error by the multi-rate detection is larger than that by single-rate detection. To obtain the maximum amount of alarms for the events with each sampling rate, the threshold for the detection must be lowered to the minimum value among all the sampling rates. With the lowered threshold, the alarm points are changed for those with higher thresholds. As a result, the data patterns in the detector's sliding windows of data points are changed, which in turn may cause some variations in the computed magnitude. But the occurrence of this problem is rare and such errors can usually be averaged out among different sampling intervals in the multi-rate detection.



(a)



(b)



(c)

Figure 12 Detection of changes under different conditions with single and multiple sampling intervals in comparison with the values measured by submeters. (a). Five-hour power data sampled at 1Hz from the motor control center of a test building; (b). Comparison of the submetered power with the detected changes in the centralized power data with a single sampling interval of 10 seconds – 15 alarms for 17 events, two missed; (c). Comparison of the submetered power with the detected changes in the centralized power data with multiple sampling intervals of 1, 2, 3, 4, 5, 10, 20, 30, and 60 seconds – 16 alarms for 17 events, one missed.

Table 2. On/off power change detection for a motor control center serving fans and pumps in a test building, as compared with submetered data.

Time	Equipment	Submeter (W)	Total power (W)	Relative power (W)	Detector			
					Single interval		Multiple intervals	
					Power (W)	Error (%)	Power (W)	Error (%)
20:10	Loop-B hot water pump—OFF	-405	5970	6.8	-502	24.0	-401	1.1
20:19	Loop-B hot water pump—ON	264	5325	5.0	249	5.7	284	7.6
20:21	Loop-A hot water pump—OFF	-275	5181	5.3	Missed		-279	1.5
20:25	Loop-A hot water pump—ON	252	5489	4.6	313	24.2	227	10.1
20:34	Loop-B hot water pump--OFF	-244	5125	4.8	-98	59.8	-476	95.1
20:44	Loop-B hot water pump--ON	232	5519	4.2	497	114.2	230	0.7
21:02	Loop-B hot water pump--OFF	-309	5289	5.8	-346	12.0	-282	8.9
21:14	Loop-B hot water pump--ON	267	5105	5.2	171	36.0	252	5.6
21:29	Loop-B hot water pump--OFF	-202	5755	3.5	-11	94.6	-172	15.1
21:47	Loop-B hot water pump--ON	237	5371	4.4	289	21.9	257	8.5
21:55	AHU-A supply fan--OFF	-1400	4224	33.1	-1396	0.3	-1411	0.8
21:56	AHU-A return fan--OFF	-82	4142	2.0	Missed		Missed	
22:00	AHU-B supply fan--OFF AHU-B return fan--OFF	-430	3727	11.5	-274	36.3	-485	12.8
22:06	Loop-B hot water pump--OFF	-300	3338	9.0	-242	19.3	-243	18.9
22:26	Loop-A hot water pump--OFF	-264	2734	9.7	-283	7.2	-244	7.4
23:40	Loop-A hot water pump--ON	342	3092	11.1	373	9.1	220	35.6
00:33	Loop-B hot water pump--ON	295	3789	7.8	494	67.5	451	52.7

In addition to the data sampling technique, characteristics of the power profile play an important role in the detection of changes and the selection of values for the modeling, which can be seen from the above plots and the table as well. Two major factors have been found to influence detection quality, the relative magnitude and the noise level.

1) Effects of the relative magnitude

In principle, the NILM can be used to detect changes of any magnitude. However, tests have shown that in order to minimize the occurrence of false alarms due to random noise in the data, the magnitude of the minimum detectable event should not be less than the current standard deviation of the data window. This minimum value helps to determine the smallest component that can be modeled from the system's total power as well as the appropriate time for the detection of changes to be used for the fitting.

It has been found that the detectability of an event changes with its magnitude relative to the current integrated power value of the system and the detection becomes more accurate when its magnitude accounts for a larger portion in the total power consumption of a given system. The

visibility of a change to the detector diminishes with the increase of the total power and the consequent accumulation of noise as more components are put into operation.

Although no quantitative and concrete correlation has been observed between the detection error and the relative magnitude of a change, a general descriptive trend has been found as the envelope of the detection error under different relative magnitudes, shown in Figure 13. The trend line is composed of the maximum error found at different relative magnitudes in the detection. The randomly distributed points show that the errors vary significantly even for the same relative magnitude, but the dashed line clearly suggests that the maximum error of the monitored event decreases with the increase of the relative magnitude.

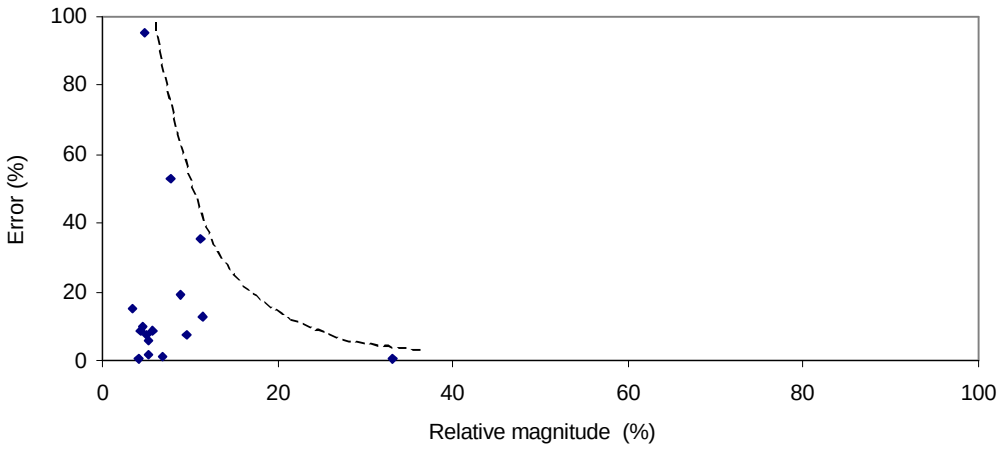


Figure 13. Errors between the detected changes from the total power data and the submetered values vs. the relative magnitude of the submetered values.

A reliable power model of the component cannot be established if the component's power input accounts for a very small portion in the monitored system, e.g., less than 5% in this example.

However, with only the relative magnitude, it is still difficult to choose the appropriate data points without measurements by submeters, due to lack of control signals and the uncertainties in the detection caused by the irregular fluctuations in the power series. In order to choose reliable detection outputs, another index, namely, the standard deviation has been found useful in tracking the noise level of the current environment.

2) Data screening by the standard deviation in the NILM detector's pre-event window

In a real power system, the values of random noise are difficult to define by a function or formula. However, the characteristics of the noise at a given time can usually be reflected by such indices as the approximate period and the peak value for the transient description of a disturbance and the standard deviation for the average noise level. Generally, the error of a detected change increases with the noise level, as measured by the ratio of the standard deviation to the total power. In this research as well, it has been observed that the accuracy of the detected changes is related to the noise level represented by the standard deviation in the detection window.

Test results showed that large relative errors were always accompanied by high levels of noise as the major source of errors in change detection. However, a high level of noise does not necessarily result in large deviations between the NILM results and submetered data. Figure 14 shows this clearly, for the same 17 events shown in Figure 13. The number beside each point is the value of the relative standard deviation at the time of the event. It can be clearly seen that big errors always occur with large relative standard deviations, such as the points at 66.4, 77.2, and 90.1. On the other hand, large relative standard deviations do not necessarily lead to big errors, such as the points at 76.6 and 79.6.

Similar patterns for the distribution of relative error vs. relative standard deviation have been found with all other tests conducted in this research. In spite of the lack of an explicit formula in the relationship between the two variables, such a trend provides a sufficient, but not essential, condition for the selection of data to be used for correlation. Useful data can be chosen from the detected changes by discarding the ones with relatively large values of relative standard deviation in the current detection window. Although this may sacrifice some data points with big values of relative standard deviation but small relative errors, such as the points at 76.6 and 79.6 in Figure 13, this method helps to secure a more accurate model and hence more reliable outputs for future fault detection.

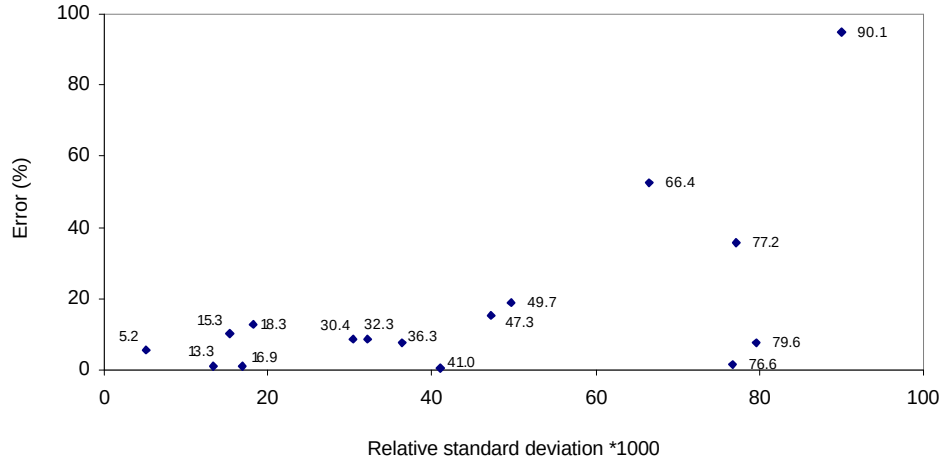
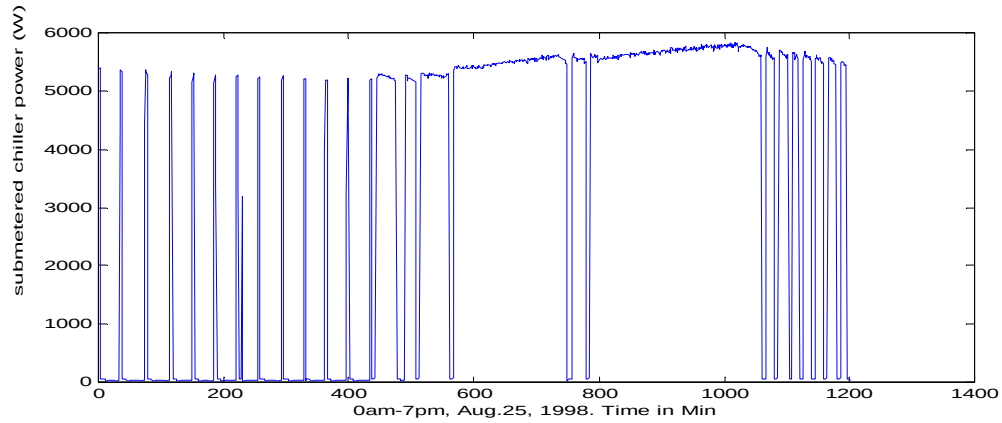


Figure 14. Distribution of the relative error with the relative standard deviation in the detection window.

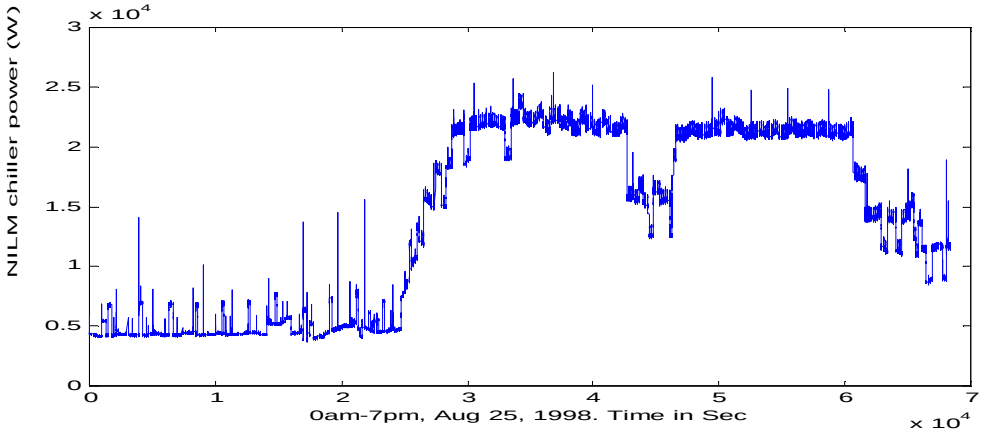
Moreover, such a screening method can be realized without the reference of submetered data. This is because the values of the relative standard deviation are directly calculated with each incoming point in the pre-event averaging window and then compared among all the events involved.

4.4 Detection of On-Off Cycling of Two-Stage Reciprocating Chiller

Luo (2001) has shown that the NILM can detect the chiller's on-off cycling at nighttime, when there are few other loads. We would like to know how the NILM works during daytime when the load condition is more complex, which is more valuable for detection and energy usage estimation. Data from the day of August 25, 1998 were used to test the chiller on-off cycling. The submetered chiller power data and the corresponding NILM data are shown in Figure15 a) and b) respectively.



a)



b)

Figure 15 Data for chiller power, 0am-7pm, August 25, 1998. a) Submetered data; b) NILM data

a) Detection of on-off cycling at nighttime

The chiller was turned on and off frequently at nighttime when there are few loads running. The period of 0-7am on August 25, 1998 was used to detect the chiller's on-off cycling. The submetered data, the NILM data and the NILM detection results using multi sampling rate GLR are shown in Figure 16. A symbol of plus represents a detected change and the sign and the magnitude of a change can be seen from the power axis. According to the submetered data, during 0am-7am on August 25, 1998, there are a total of 11 chiller on-offs. Observing the detection results in Fig.16, where a symbol of '+' represents a detected change, all the 11 on-offs are identified with the satisfactory times and magnitude estimations.

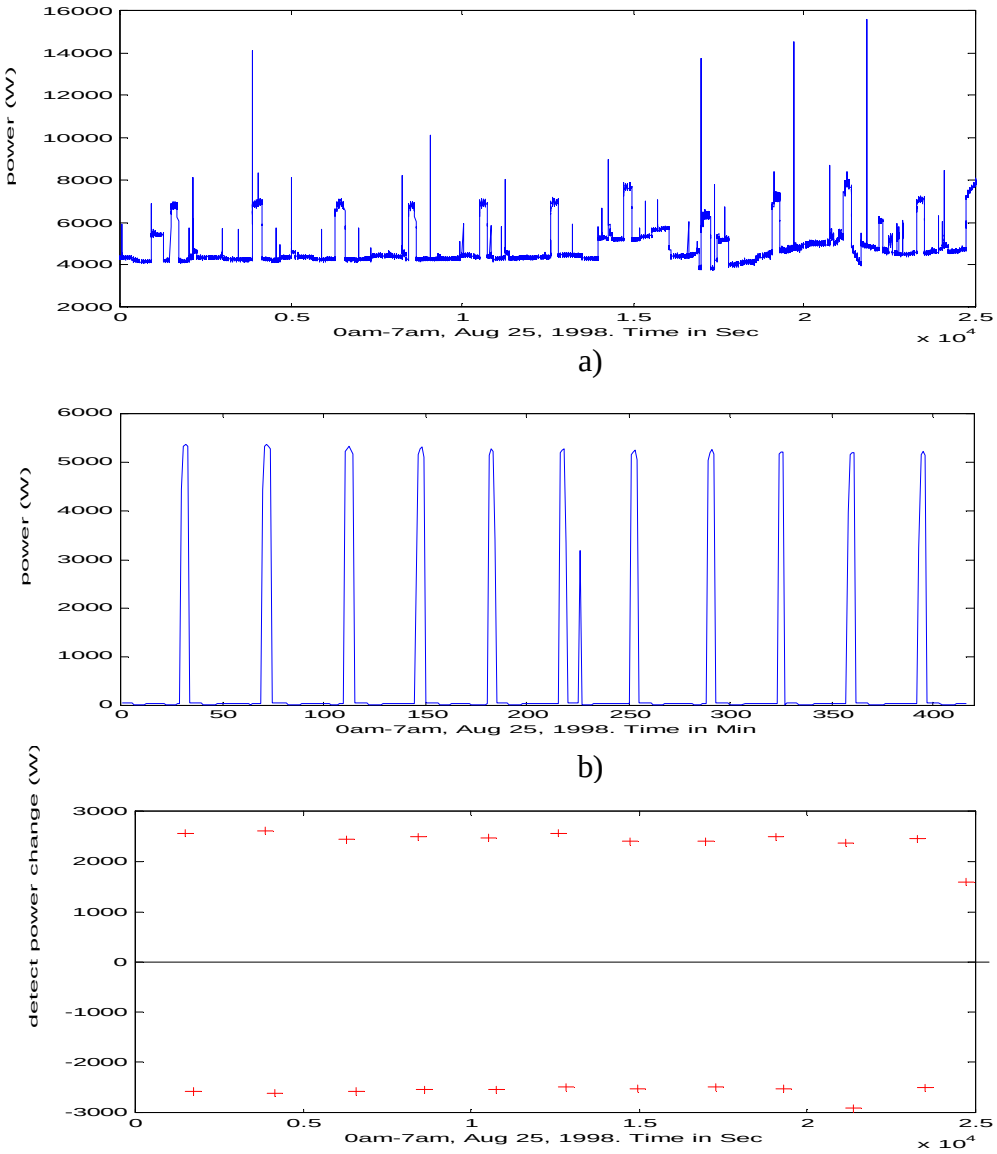
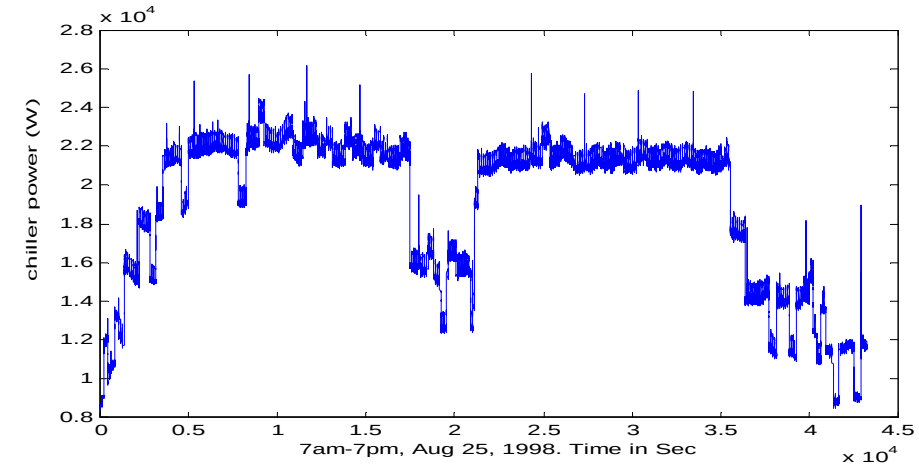


Figure 16 NILM detection result for the chiller on-off cycling at nighttime: 0am-7am, August 25, 1998: a) NILM data; b) submetered data; c) detected power changes.

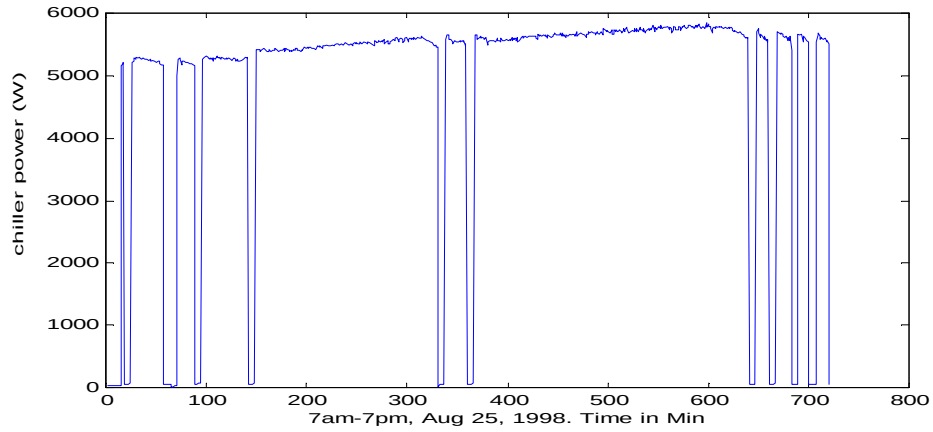
b) Detection of chiller's on/off in the day time

We have shown that the NILM was able to see the chiller on-off cycling during nighttime hours. A more interesting issue is whether or not the NILM is able to detect chiller's on-off cycling during daytime hours when many more other loads operate simultaneously. To test this, the time period of 7am-7pm, August 25, 1998, was input to the multi-sampling GLR program. The submetered data and the NILM data for this specific period are shown in Figure 17 a) and b) respectively and the GLR detection results in Figure 17 c). A symbol of '+' represents a change detected and its magnitude can be found on the power axis.

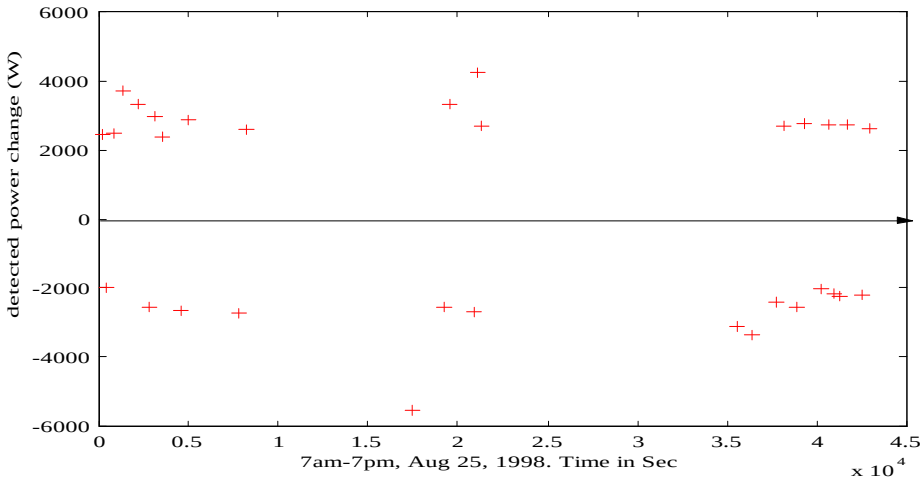
According to the submeteted data, there are a total of 11 on-offs during this period: four in the morning, 2 around the noontime and 5 in the afternoon. The detection results indicate that all the on-offs are detected with the correct magnitudes, as seen in Fig.17 c). However, together with the chiller's on-offs, 3 extra ons are also detected in the morning and 2 extra offs in the afternoon. Comparing the detection results with the NILM data, we know that these extra ons and offs are the system noise that also meets the detection requirements for the chiller. We need to refer to the chiller operating status, which can be obtained from the system control signals, to determine whether or not an on or off is the chiller's. More is discussed in section c) of chiller energy estimation.



a)



b)



c)

Figure 17. Detect chiller on-off cycling during daytime hours: 7am-7pm Aug 25, 1998. a) Submetered data; b) NILM data; c) Detection results of multi-sampling rate GLR.

c) Estimate Chiller's Energy Usage

The chiller on-off cycling can be used to estimate its energy usage. Starting with every detected turn-on event, the product of the power change magnitude and the time interval between this turn-on to the next turn-off is the energy consumption during this period of time. From a turn-off to the next turn-on, the power consumption is zero in theory. The total energy consumption can be obtained by accumulating those energy step changes through the on-off cycling.

It is to be understood that the performance of change detection determines the preciseness of the energy usage estimation. Two key ideas are adopted in the estimation. One is the chiller power change magnitude. Of the detection results, only those changes that have the expected change magnitudes are retained. The other idea is to utilize the information of the chiller operations schedule, i.e., when the chiller is really running. The schedule information is very important in that otherwise the estimation would be totally mistaken by a single misdetected on/off event at the beginning of the testing period.

Given the potential difference in the detectability between nighttime and daytime, the energy usage estimation based on the detector is conducted separately for each period. The one for the nighttime on a normal day is shown in Table 3. The detector-based energy usage estimation is 17.4% lower than the value provided by the submetered data. The error mainly comes from the change magnitude detection because the on-off cycling gets detected precisely, as seen in Figure 16. Table 4 shows the detector-based estimation of chiller power usage during the daytime on the same day and compares that to the submetered data. During the daytime, the detector's result is 17.7% lower than that from submetered data. It should be noticed that the good performance of the detector in the daytime case is largely credited to the chiller operations schedule. As can be seen from Figure 17, there were two major periods of time on August 25, 1998 when the chiller was on. If we got these two parts wrong, the estimation would be far from the reality. In addition, if the chiller is turned on and off fairly frequently during the day and there is no large part of time when the chiller's status is constant, it is expected that the estimation performance would go down. Tests on more days of data and tests at sites in California will further assess the ability of the NILM to quantify the energy consumption of reciprocating chillers.

Table 3. Energy estimation of a chiller for 0-7am, August 25, 1998.

Item	Chiller power based on submeter	Chiller power based on detector
Energy (kWh)	4.2	3.5
Error (%)	0.0	-17.4

Table 4. Energy estimation of a chiller for 7am-7pm, August 25, 1998.

Item	Chiller power based on submeter	Chiller power based on detector
Energy (kWh)	61.7	50.8
Error (%)	0.0	-17.7

5. Detection Of Selected HVAC Faults From On/Off Electrical Data

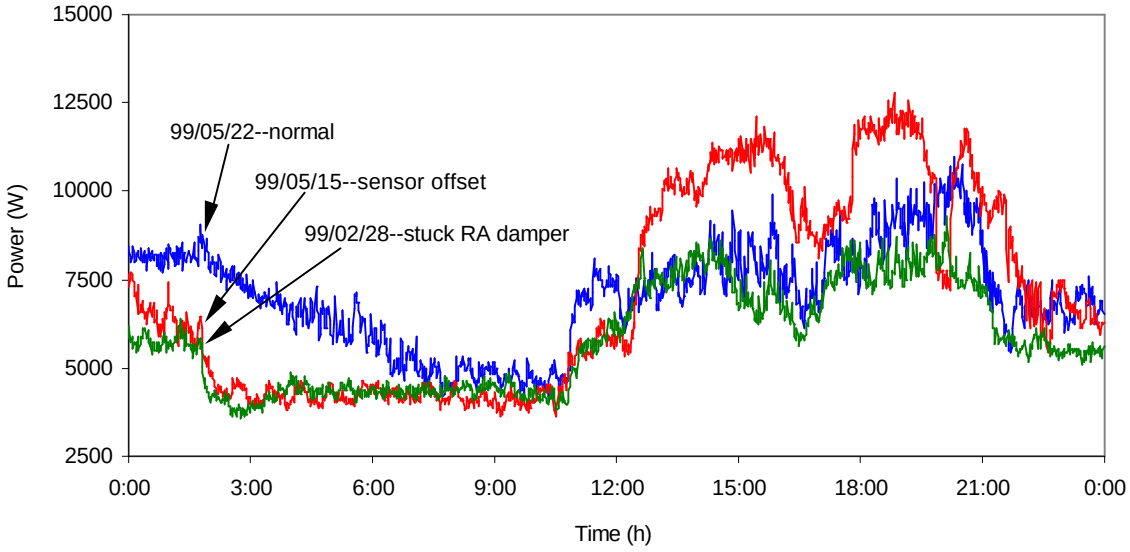
5.1 Fan: Pressure sensor offset and stuck-closed damper

In general, a change in the total power data at turn-off is generally more abrupt and hence easier to detect than during turn-on, especially for equipment with startup protection or variable load control. Therefore, detection of faults related to the magnitude of the change in total power is conducted at turn-off. If the detected change in the total power data caused by turnoff of the equipment at the given time is significantly higher than a trained threshold for this component, a false alarm will be issued for it. For example, a VSD fan tends to consume more energy when the control loop is not working properly. Figure 18(a) shows the total power profile of the motor control center under normal and faulty operating conditions throughout 24 hours (8:00pm – 8:00pm) of 3 different days, with the supply fan turned off at 10:00pm on each day. Figure 18(b) compares the detected magnitude of the changes under normal and faulty conditions with some additional normal days' of data. All the points under 600 watts are the detected turnoff power changes under normal conditions during 16 days. Two days with an artificially introduced supply-air static-pressure-sensor offset clearly show significantly larger values (> 600 W). Therefore, by setting a threshold of 600 W for the change magnitude, abnormal power consumption can be recognized by the detector. Other faults with similar effects can also be found in the same way, such as an unexpectedly restricted airflow path that also leads to higher fan power consumption and hence a larger magnitude of change in the system's total power when the fan is turned off. Figure 18(b) presents such an example of a stuck-closed recirculation-air damper.

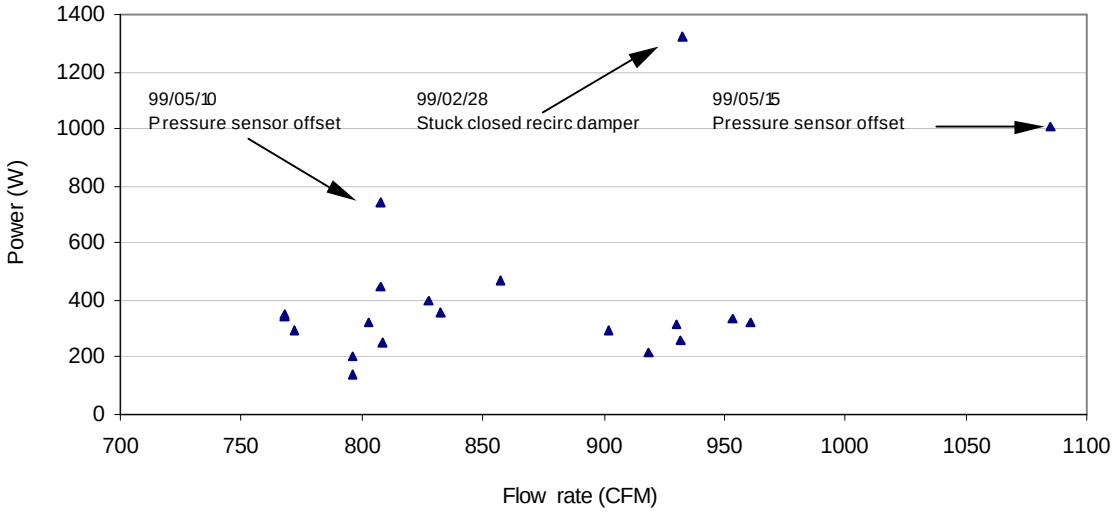
While faults can be detected when an abnormally large magnitude is found at turnoff, extremely small changes in power, or the absence of any change, also indicate potential faulty operation. For example, a fan with a slipping belt may lead to an extremely small magnitude of total power change at turnoff.

5.2 Chiller: Leaky cooling coil valve

Sometimes, a fault does not cause any abnormality in the power magnitude, but, rather, in the on/off cycling frequency of the equipment, especially when the motor runs at a constant speed. This method is very effective because in HVAC systems, especially for commercial applications, there are generally certain restrictions on the on/off cycling of equipment to prevent deterioration by frequent switches.



(a)

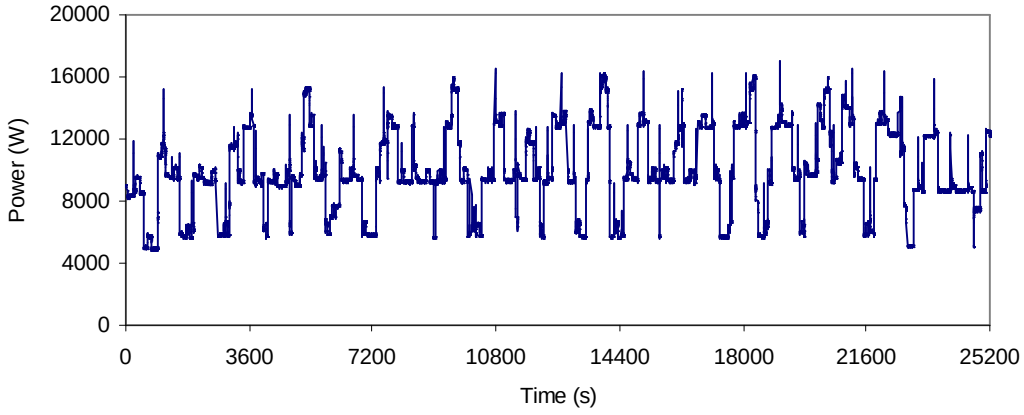


(b)

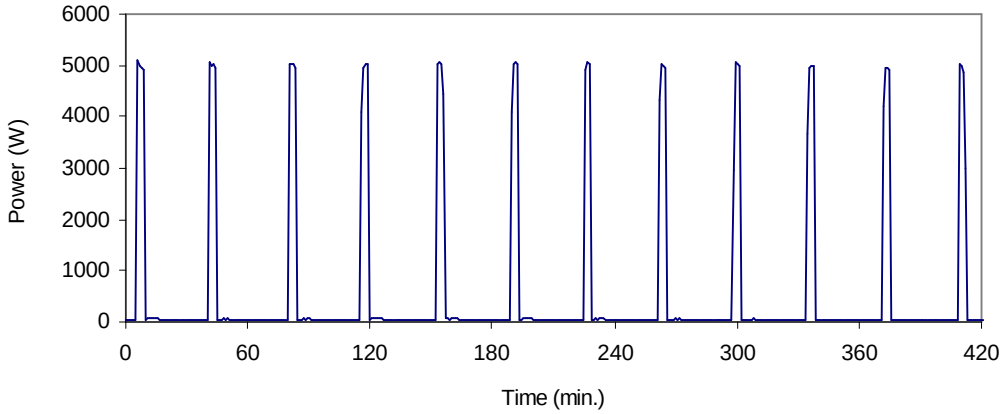
Figure 18. Detection of the faults of pressure sensor offset and stuck-closed damper from the magnitude of changes in the HVAC system's total power input at the turnoff of a supply fan. a). 24-hour total power data of three typical days of an air conditioning system including 10 fans and 6 pumps; b). detection output of the total power change caused by the turnoff of the fan under normal and faulty operation conditions.

For example, chiller operation is generally controlled based on the entering and leaving water temperatures, which are mostly affected by the cooling load introduced through the cooling coil valve. If the load condition stays near a stable state, e.g., during the late night or early morning in an office building, the chiller cycling period should also remain constant under normal operation. When the cooling coil valve is leaking, the load of the chiller tends to increase due to the higher flow rate of water to be processed, which tends to cause shorter off-on intervals for a chiller with stepwise control. Therefore, the leaky valve might be identified if the chiller on/off switches can be correctly detected during periods under appropriate load conditions.

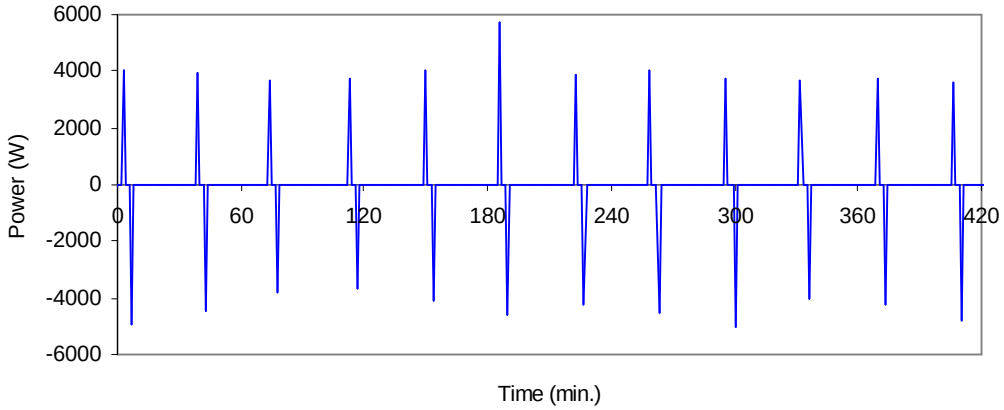
It should be noted that with the chiller power input, the leaky-valve fault can be most reliably identified when the cooling coil valve is closed by the control signal. The following charts show the output obtained by the detector from the total power data of the test building. In this system, the cooling coil valve was closed and the chiller cycled on and off regularly during the early morning hours to maintain chilled-water temperature, which would otherwise rise due to heat gains through the piping. Under normal conditions, the on-off duration and the off-on interval were 4~5 minutes and 38~39 minutes respectively, i.e., about 10 full on/off cycles in total between 0:00~7:00 am when the valve was closed. Therefore, the threshold of a full cycle that consists of the on-off duration and the off-on interval has been calculated to be 42 minutes/cycle, which was then used as the minimum time period for a full on/off cycle. If the valve is leaky, the average period of a full on/off cycle tends to exceed this threshold when the valve is expected to be closed and an alarm will be issued for the excess by the automatic detector. During the FDD tests in spring, the leaky valve fault has been found on three days with the average full cycling intervals of 33.2, 30, and 27.3 minutes respectively. Figure 19 demonstrates the detection output of one day with the leaky valve.



(a)



(b)



(c)

Figure 19. Detection of the fault of a leaky cooling coil valve from the cycling frequency of the chiller with the building's total power data under low load condition during the early morning hours. (a). 7-hour total power data of the system with a leaky cooling coil valve; (b). chiller cycles from the submetered chiller power; (c). chiller cycles as the detected power changes.

6. Estimation of Energy Consumption of Specific Components

Estimation and modeling of the energy consumption of the major components in HVAC systems were derived from the detection of changes in a system's total power profile. The general modeling procedure is summarized and two modeling examples follow. The general concept is to build up a correlation of electrical power with an independent variable (airflow, for example), using power data only from the NILM. To do so requires that the component be cycled on and off at different loads and that the power changes at the time of the transients be recorded by the NILM.

6.1 Principles

Assuming the GLR detector has been trained with the method developed in section 3.5, the component modeling procedure can be summarized as follows:

1). Obtain information about a component of interest, including power capacity, type of motor drive, and general operating schedules from manufacturer's catalogs and/or the system's design specifications;

2). Study the feasibility of the modeling of the equipment's power input with the system's total power based on the component's power capacity relative to the system's total value. Although there is no fixed criterion, the relative magnitude should not be too small, e.g., 1/1000 is too small to be seen by the detector. Tests showed that this value should not be less than 5%, especially in the presence of VSD motors in the system. Further, if the power capacity of the equipment is less than the standard deviation of the total power data when this equipment is in operation, then its power input may not be modeled with desirable accuracy;

3). Determine the parameter to be correlated with the power consumption of the equipment;

4). Select the parameter that can be used as an indicator of the major variations in the load of the system. In common HVAC systems, outdoor air temperature or enthalpy can be used if the heat gain/loss from outside is the major variable part of the total load. If the internal load dominates, then use the schedule of the source as a reference, e.g., number of occupants;

5). Observe the total power data of the monitored system from one normal day's operation and find the time for the on/off events to be detected based on the variations of the load condition;

6). Determine the sampling rates of power data. In addition, check the availability of the control parameters of the equipment, such as the PID coefficients to help to determine the time constant of the process;

7). Apply the detector to the power data and record the values of the reference parameter to be correlated at the on/off switches of the equipment;

8). Choose the power data for correlation based on the calculated values of the relative magnitude of the changes and the relative standard deviation at the event time as discussed in Section 5.3;

9). Set up the model by correlating the power data with the reference data and form the allowable offset range from the fitted function based on a confidence level;

10). Test the model with additional data collected under normal operation and make modifications if needed.

In deciding the time points for the on/off events in Step 5, some preliminary information or observations might be helpful to avoid switches of the equipment that are likely to cause large errors in the detection. In addition, in order to obtain more accurate changes, periods with more dynamics marked by large relative standard deviations such as the transient processes should be avoided. Although the occurrence of large noises and transient processes is rarely known in advance, detection with significant effects of such disturbances can be skipped based on the knowledge of the system's normal control sequence or by looking at the power profile during a typical day's operation.

6.2 Modeling Examples

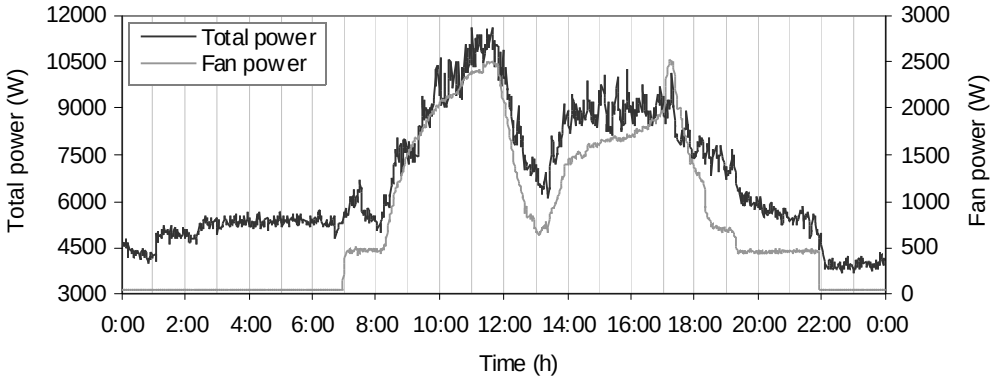
a) Fan

Figure 20 illustrates the modeling process for the power consumption of a supply fan with a variable speed drive and a design capacity of 5 kW in the test building. It is known from the design information of this system that the major factor that causes the variations in the building's cooling load is the outdoor air temperature. The cooling load trend can be

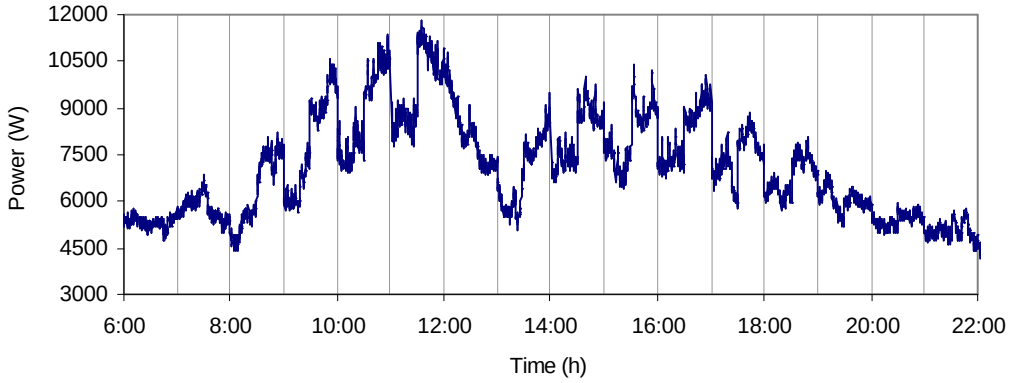
approximately predicted by keeping track of the outdoor air temperature. Table 5 lists the necessary preliminary information for the detection. From the twenty-four-hour data of a typical summer day, it can be seen that during the transient processes between 7:00 – 8:00 when the system was started and 12:00 – 13:00 when the system was turned down by the control system during lunch break, the power data showed significantly more dynamics than the rest of the day. This indicates that manual turnoff should not be implemented during these time periods. For the remaining time, manual turnoff of the fan was simulated with the power data by subtracting the submetered power data at an interval of 1 hour from 8:00 to 20:00 excluding 12:00 and 13:00, as shown by Figures 20 (a) and (b). The 1-Hz data were then fed to the detector for changes at these time points. The standard deviation was also computed simultaneously in the moving window. The detected changes in comparison with the submetered values and the relative error distribution with the relative standard deviation are shown in Figures 20 (c) and (d) respectively. Data points at 14:00, 15:00, and 16:00 corresponding to the three largest relative standard deviations 70.5, 100.3, and 74.6 were further removed for the modeling, although it can be seen from Figure 20 (c) that the error at 14:00 is not the largest among the other nine points. The nine remaining data pairs were finally used for the polynomial power model. From the comparison between the model based on detection with the total power and that from the submetered data as shown in Figure 20 (e) and Table 6, it can be seen that the difference between the two models decreased with an increase in the air flow rate and hence the fan power input, which further verified that the detection accuracy increased with the relative magnitude of the monitored equipment in its host system.

Table 5. List of the information used in the power change detection of a supply fan.

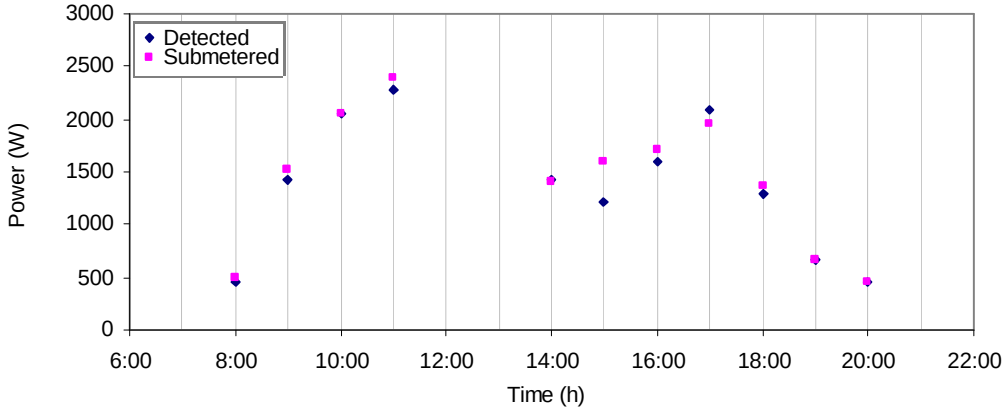
Items		Descriptions	Comments
Fan	Power capacity	5.0 kW	
	Minimum power input	0.5 kW	When the fan is in operation
	Type of motor drive	VSD	
	Minimum speed control signal	20%	Under the zero-load condition
	Operating schedule	7:00 – 22:00	
	Power percentage of system	15.5%	Detectable magnitude
	Load indicator	OA temperature	
	Correlation parameter	Air flow rate	The fan speed control signal can also be used for FDD and energy estimation.
	Confidence level of model	90%	
	Event time	Every hour	8:00 – 20:00, excluding 12 and 1 p.m.
System	Total capacity	32.25 kW	
	Transient period	12:00 –14:00	Lunch break -System turndown
	Range of standard deviation	0.0 – 105.0 W	Higher values occurred at events
	Sampling intervals	1 – 30 seconds	Multi-rate detection



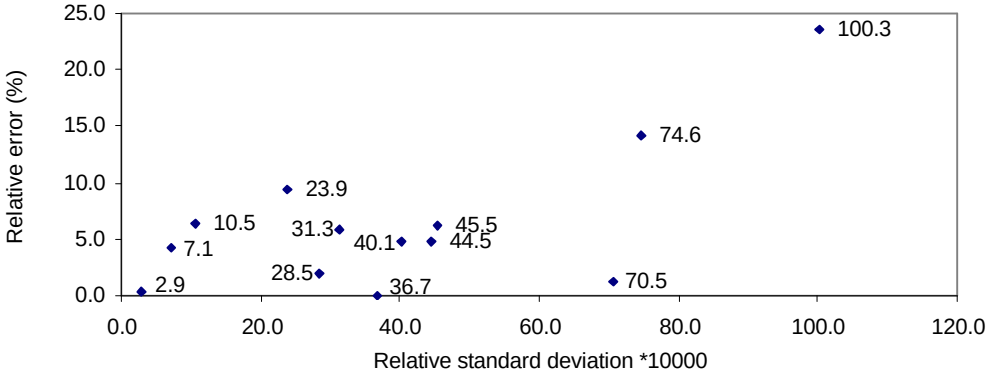
(a)


12000
10500
9000
7500
6000
4500
3000
Power (W)
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time (h)

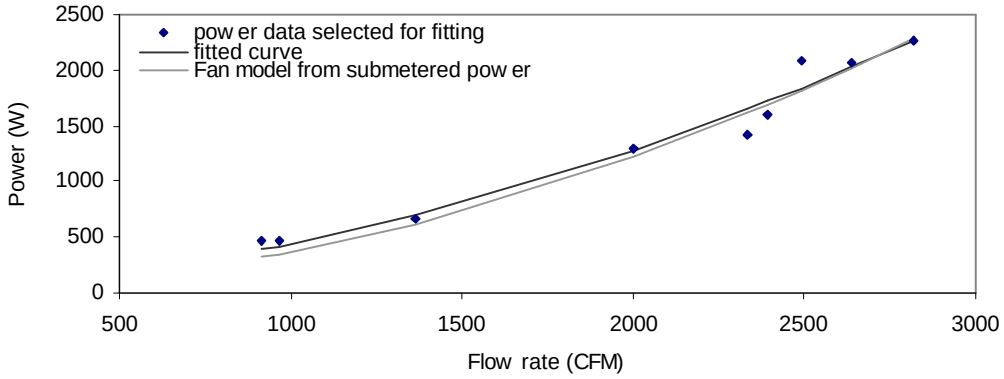
(b)


Detected
Submetered
3000
2500
2000
1500
1000
500
0
Power (W)
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time (h)

(c)


25.0
20.0
15.0
10.0
5.0
0.0
Relative error (%)
100.3
74.6
23.9
10.5
31.3
45.5
40.1
44.5
7.1
28.5
2.9
36.7
70.5
0.0
20.0
40.0
60.0
80.0
100.0
120.0
Relative standard deviation *10000

(d)



(e)

Figure 20. Component power modeling via detection of power changes at the system level--electrical power input vs. air flow rate of a supply fan in a test building. (a). Twenty-four-hour total power of the motor control center and the submetered power of a supply fan; (b). Total power of the motor control center with the manual shutdown at an interval of one hour from 8:00 to 20:00 except 12:00 and 13:00 when the system was turned down for lunch break; (c). Power changes measured from the submeter and that detected from the total power at the shutdown of a supply fan; (d). Error distribution with the relative standard deviation; (e). Comparison between the two models based on the detected changes from the total power data and the submetered fan power data respectively.

The model based on detection of changes by manual shutdown can be used to estimate energy consumption of the related equipment and of the monitored system if a model can be established for each component. For the 24-hour operation of the day used in the above example, as listed in Table 7, the energy consumption of the fan was 19.5 kWh by the submeter and 18.8 kWh from the model with the measured air flow rate. The error was about 3.6% despite an electrical surge around 17:30 as seen in Figure 21, which has rarely occurred in common operation. The errors for other normal days' operations were found less than 3.5%.

Table 6. Comparison between the fan power as detected changes in the total power and submetered measurements at given airflow rates.

Air flow rate (CFM)	Detected power change (W)	Submetered power data (W)	Error (%)
914.1	387.2	317.5	22.0
965.3	418.8	347.2	20.6
1361.6	698.5	620.2	12.6
1999.1	1280.0	1218.8	5.0
2332.5	1648.7	1610.1	2.4
2391	1718.0	1684.3	2.0
2490.4	1838.8	1814.2	1.4
2636.7	2023.9	2014.0	0.5
2820.9	2269.0	2280.4	0.5

Table 7. Energy estimation of a fan based on submeter and detector for a normal day.

Item	Fan power based on submeter	Fan power based on detector
Energy (kWh)	19.5	18.8
Error (%)	0.0	3.6

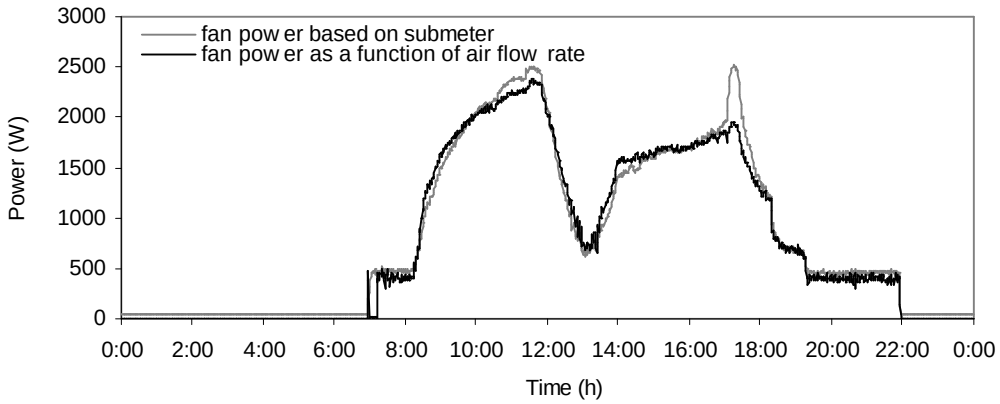


Figure 21. Twenty-four-hour supply fan power profiles during a normal working day by measurements and simulation of power as a function of air flow rate based on detection of changes in the total power data.

In practice, control signals that can be obtained from the building energy management system may also be used to establish power functions for the related equipment by following the guidelines described above. For example, power input of a fan can be modeled as a function of the motor speed control signal. However, for equipment driven by a VSD motor, special care must be taken in data acquisition for the model fitting when the equipment is running under low load conditions, i.e., when the control signal is at a low value. Although the power data should be ideally obtained under a wide range of the control signal to produce an accurate model, detection of the related changes in the total power series may not be applicable for the model. This is because with a small value of the control signal, the relative magnitude of the equipment's power consumption may also become too small to be computed with acceptable accuracy and hence the quality of the model is affected. Therefore, the power model needs to be extrapolated for detection and energy estimation under low load conditions. When the equipment is not in operation, the controlled variable and the power input should be zero. In the test building, however, the minimum fan speed control signal was 20% as listed in Table 5, which indicates zero power input and zero rotational speed are related to 20% fan speed control. Therefore, the zero-load data pair (0.2, 0.0) must be used in the correlation as a constraint for the function for low-load energy calculation. Otherwise, the power input will be interpreted as zero under zero speed control signal by extrapolation, which may lead to considerable errors in energy estimation, as listed in Table 8 for the same data set used in the above example illustrated in Figure 20. Comparison of the fan power consumption between the submetered value and the power model with appropriate low-load constraint is illustrated by Figure 22.

Table 8. Estimation of energy consumption of a fan in one day by the power vs. speed control signal model with different constraints under the zero-load condition.

Item	Submeter	Power functions based on change detection		
		No constraint on the zero-load condition	Zero-speed zero-power	20%-speed zero-power
Energy (kWh)	19.5	14.5	14.6	19.0
Error (%)	0.0	25.6	25.1	2.7

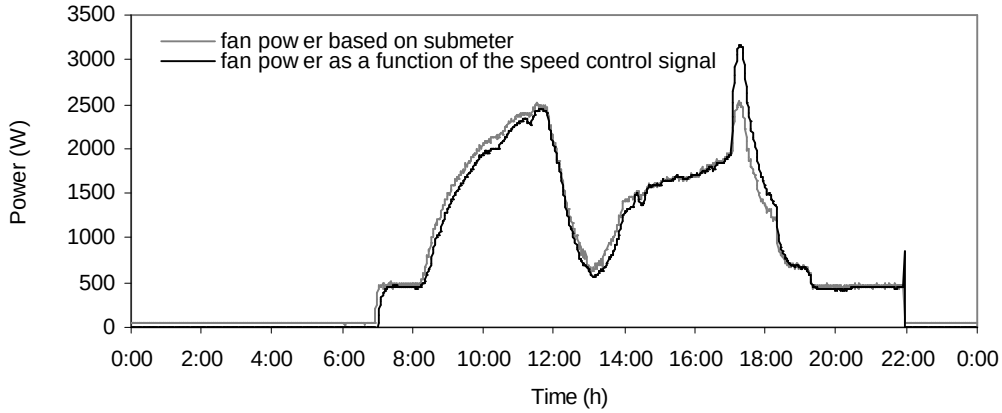
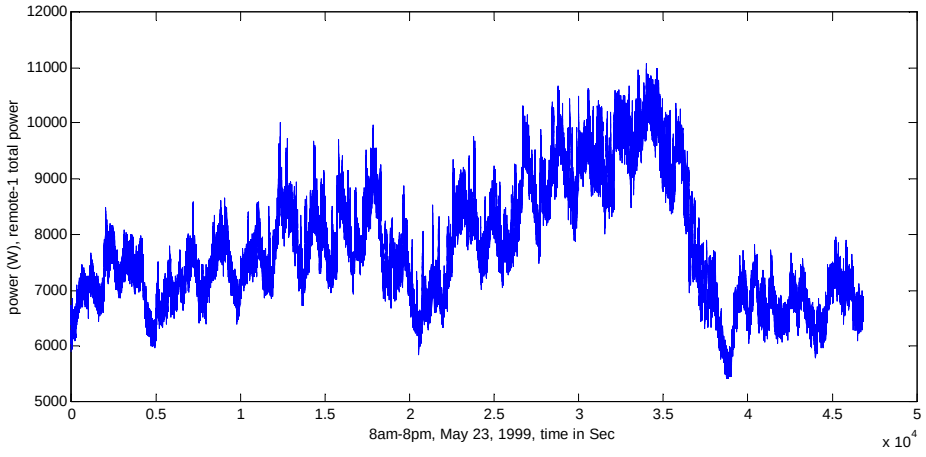


Figure 22. Twenty-four-hour supply fan power profiles during a normal working day by measurements and simulation of power as a function of fan speed control signal based on detection of changes in the total power data.

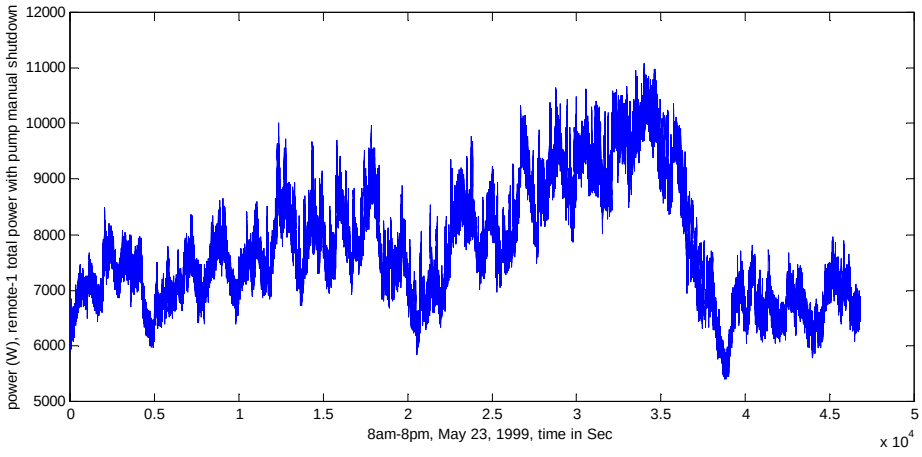
b) Pump

There are ten pumps used in the system, with capacities ranging from 0.4 kW to 1.0 kW. As shown in Table 5, the power capacity of the system, which includes these ten pumps and six fans, is 32.25 kW. Compared to the system power capacity, all pump capacities are below 5% of system power, making detection of power changes a very challenging task. The manual shutdown simulation was conducted on the total power data on a normal day when many other loads in the system were running simultaneously; detection results are shown in Figure 23. Figure 23 a) is the total power profile of the system at the motor control center. Manual shutdown simulation is conducted in Figure 23 b) by subtracting the submetered values of the pump power consumption from the total power at an interval of an hour from 8am to 8pm on May 23, 1999. The updated total power is input to the GLR detector and Figure 23 c) shows the results. The symbol of '+' represents the detected changes using 60-sec sampling interval and 'o' 120-sec interval. It can be seen that the pump power magnitude is so small that most changes are buried in the system noise under both sampling rates. This test was substantially the same as that reported earlier for the pump, in section 4.2, but with the addition of more transitions; the conclusions are the same.

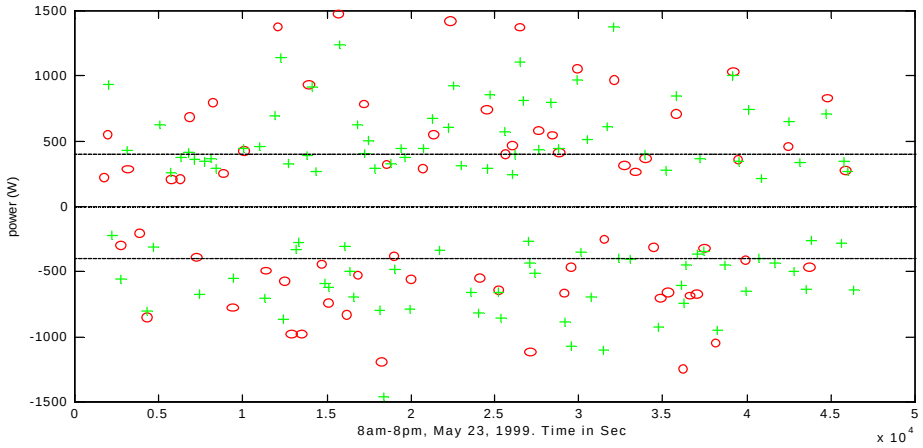
It is hard to detect the operation changes of pumps during daytime, but it is possible to do so at night when most fans are off and the system load is small. The manual shutdown simulation is also conducted for the time period of 2-6am May 23, 1999 and all four hourly shutdowns get detected. To overcome the pump power capacity constraint in modeling the pump power consumption, one approach could be to detect the changes of manual shutdown only at nighttime. The problem is that all the VSD pumps run at a constant speed and therefore nearly constant flow rate and power consumption at nighttime. The lack of operation variety would reduce the effectiveness of the power consumption model. If in practice, we could artificially adjust the operation status of a specific component and then collect the NILM and control signal data, we should be able to build a better power consumption model. In addition, four pumps in the system are constant-speed drive. Their operation status and power consumption are constant over the time; therefore there is no need to model the power consumption.



a)



b)



c)

Figure 23. Detect the hourly manual shutdown of the AHU-A chilled water pump, 8am-8pm, May 23, 1999. a) Original total power profile of the motor control center; b) total power of the motor control center with the manual shutdown at an interval of one hour; c) detected changes based on the power data in b) using different sampling rates, '+' 60-sec interval, 'o' 120-sec interval, dash line - pump power capacity.

6.3 Discussion

As a general rule for equipment to be monitored from the integrated power profile, it has been found that the relative magnitude of the component's power consumption in the system's total power input should not be too small. As seen in the report, most pump changes were not detected due to the small power percentage. A power logger at a lower level is needed in order to find the power changes that are accurate enough for feasible evaluation or modeling of a small component, such as a chilled water pump. For example, a logger for all the pumps in a building would be feasible if the wiring supported such an installation by separating pumps from fans. Although such a monitor is not working on the system's total power system and the detection may require more than one logger for the whole system, it may still be worth doing compared to submeters for each component.

In addition, two potential problems with the tests need to be addressed when applying the detection method. First, for the turnoff power detection, it is assumed there should be no other events happening at the same time. This might produce false alarms if something else actually occurred. One possible solution is to repeat the manual change detection on another day with a similar load condition and remove the data points that differ significantly from that at the same time on the other day(s). Second, the detection in the test building was conducted with a base sampling interval of one second, while the sampling interval of the reference power data by the submeter was one minute. This might introduce deviations between the submetered data and the detected values due to the ever-changing magnitude of the total power even if no other events occurred during the one-minute period. Such a potential difference can be reduced by

using a submeter with a sampling interval that is comparable to the base sampling interval of the detector. Submeters will be used to assess NILM performance in California buildings.

It should be noted that although feasible power models of the major components can be obtained from the total power data at the system level with the method developed in this report (manual switching of equipment to establish a power curve), submeters need to be used for fault detection and monitoring when the equipment must run around the clock, such as in a hospital where the supply fans should never be turned off. Under such conditions, fault detection and diagnosis should be based on submeters if power consumption is used as the major index for monitoring.

7. Recommendations for NILM Placement

Besides system hardwire constraints, NILM placement should consider the power percentage of all the components that are monitored in the system. The goal is to have most of these components 'seen' by the NILM, meaning detectable. If the power consumption of some component is too small compared to other components in systems, we need to either add more NILM box or seek help form the submeter which can be installed for this specific component and all other components can still be analyzed by the NILM.

8. Conclusions

The objective of the NILM is to reduce the cost of obtaining useful information concerning the energy consumption of HVAC equipment and whether or not the equipment is operating properly. This report demonstrates that a number of turn-on and turn-off events for fans and a chiller in a test building in Iowa could be detected with two centrally located power meters. As a general rule for equipment to be monitored from the integrated power profile, it has been found that the relative magnitude of the component's power consumption to the system's total power input should not be too small. With the test building, a lower limit of 5% was used for the power model. This lower limit will be evaluated in other buildings, including those in California. Detection limits will determine the number and location of NILM meters required to obtain useful information about a building and its equipment. For example, fault detection and diagnosis for an air handler may dictate a different meter location than roof-top air-conditioner diagnostics or first-cut whole-building energy disaggregation.

Correct operation of the NILM requires careful selection of a number of parameters, concerning the size of detection windows, data sampling rate, detection threshold, minimum power level of interest, and signal standard deviation. Tests in other buildings, as planned, will help determine the extent to which these parameters are suitable without adjustment and will also serve as a means of evaluating a tuning procedure that has been developed but not yet evaluated.

References

Leeb, S. B. 1993. "A conjoint pattern recognition approach to non-intrusive load monitoring." Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Luo, D. 2001. "Detection and Diagnosis of Faults and Energy Monitoring of HVAC Systems with Least-Intrusive Power Analysis." Ph.D. thesis, Department of Architecture, Massachusetts Institute of Technology.

Norford, L. K. and S. B. Leeb. 1996. "Nonintrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms." *Energy and Buildings* 24, pp.51-64.

Norford, L. K., J. A. Wright, R. A. Buswell, and D. Luo. 2000. "Demonstration of fault detection and diagnosis methods in a real building." Final Report for ASHRAE Research Project 1020-RP.

Project 2.2
Equipment Scheduling and Cycling

Task 2.2.2
Fully Automated Analysis of Equipment Scheduling and Cycling

Deliverable 2.2.2(b)
Intermediate Report on Automated Diagnostics

Submitted to Architectural Energy Corporation

Under the

California Energy Commission's

Public Interest Energy Research (PIER) Program

Energy-Efficient and Affordable

Small Commercial and Residential Buildings

California Energy Commission Contract 400-99-011

K. D. Lee and L.K. Norford
Massachusetts Institute of Technology
July 12, 2001
Revised October 23, 2001

Intermediate Report on Automated Diagnostics

Progress Report- 160 Sansome Street, San Francisco
NILM Installation and Training Data

A Non-Intrusive Load Monitor (NILM) was installed on the 19th floor mechanical room of the Hong Kong Bank building (160 Sansome Street) in San Francisco on March 26, 2001. The hardware setup and the software configuration was completed on the following day and the training data was collected in the morning of March 28, prior to the routine occupation of the building. HVAC loads connected to the motor control center (MCC) were cycled on and off while recording the power and the reactive power draw using the *prep* program of the NILM software with 120 Hz of sampling frequency. Each pump or fan was repeatedly turned on and off for about 10 to 60 seconds. The on/off sequence was repeated five times, where possible. The supply and return fans were operated longer because they have longer start transients. The chillers were not tested due to the difficulty in bypassing the control circuitry.

Site and HVAC Equipment

The Hong Kong Bank building is a 100,000 ft^2 class A office building with 18 floors of office space. It is 250 ft tall and was constructed in 1964. Purchased steam is used for heating. Hot water is generated by converting the steam through a heat exchanger and supplied to zone heating hot water coils on the main air system. For cooling, chilled water is produced at a central plant. Conditioned air is distributed through the supply and return ducts. Ten of the eighteen rentable floors have been converted to the variable air volume (VAV) system. The rest still use the constant volume (CV) air handling system. Table 1 lists the HVAC equipments and their capabilities.

Table 1. HVAC equipment

Chiller	2 EA, 225 ton, Centrifugal, 0.8 KW/ton at full load
Supply Fan (SF)	Variable frequency drive (VFD) controlled, 100 hp, rated at 125000 cfm
Return Fan (RF)	VFD controlled, 75 hp, rated at 100000 cfm
Cooling Tower	Single tower with two cells, each with single speed 20 hp centrifugal fan
Chilled Water Pump	2 EA, 10 hp, rated at 450 gal/min, centrifugal, const. speed
Condenser Water Pump	2 EA, 10 hp, rated at 615 gal/min, centrifugal, const. speed

NILM Installation

Three primary current transducers (CTs) were installed to monitor a phase of the three-phase 480 VAC power input to the HVAC equipments. One CT is installed on the main line (MCC) and two others are installed on the elevator line and the transformer line (used to supply 120 VAC to 18th floor), each with reversed polarity. The current reading on the secondary sensor is the main load minus the elevator and the transformer, which is the HVAC load. The reading is fed to the analog input card of the NILM host computer and sampled by the NILM software. The NILM is currently configured and programmed to monitor phase-A current and phase-A-to-neutral potential and to compute spectral envelope time series of the fundamental and second through seventh harmonics of phase-A real and reactive power[1]. With the installation of additional CTs and the modification of the software, it will be possible to monitor three phases together or selectively. The CTs are rated at 1500 A and produce 5 A at the secondary winding (1500:5). A three-phase 480 V potential transducer (PT) is also installed to measure voltage.

The NILM may be programmed to accept user input of the CT (primary) amp rating and convert it to an appropriate scale factor. However, the necessary program change has not been made yet and the location within *prep/setpcl* of the existing scale factor is not documented and was therefore not changed before the training tests. The numbers presented in this article must be multiplied by 16.67 to obtain A-phase power and by 50 to obtain an estimate of three phase power.

Although the installation and software configuration was completed in March, the NILM host was not on line until June 1, when a DSL line was connected to the NILM to overcome the logistic problem of providing an internet connection in the mechanical room. Real time power plots can be viewed on a Netscape web browser by typing http://66.120.20.99/. A software modification is underway to recognize the URL in addition to the IP address.

[1] Real power ($P = I^2R$) is the power consumed by the resistive component (R) of the complex impedance, $Z = R + jX$. Reactive power ($Q = I^2X$) is the power periodically stored and returned by the reactive component (X) of the complex impedance. A pure inductive load gives rise to $X = \omega L$ and a pure capacitive load $X = -1/\omega C$. Apparent power (P_A) is the complex sum of the two powers ($P_A = P + jQ$).

Training Data

HVAC loads were cycled on and off, one load at a time, while recording at 120Hz the power, P, and reactive power, Q, fed to the MCC. Second and higher harmonic spectral envelopes were not recorded during these training tests.

Plots of all training time-series data are presented in this article to document the essential character and repeatability of each load transient. The real and reactive power traces for a given load can be distinguished by the fact that real power is always positive while the reactive power can be either positive or negative depending on the load characteristic. Typically an inductive load gives a positive Q value (lagging power factor), while a capacitive one has a negative Q (leading power factor)[2]. Buildings typically have lagging power factors due to the highly inductive motor loads. However, in NILM set up, negative Q indicates inductive load.

Cooling Tower

Each cooling tower cell (1 and 2) consists of a 20 hp centrifugal fan and a 10 hp condenser water pump. Figure 1 shows the graph of the cooling tower pump 1 test. Five on/off step inputs are applied to the pump. Figure 2 shows the detailed view of the fourth step test. The pump shows a large start transient, typical of the fixed-speed electric motor. The start load drops rapidly as the motor reaches the steady state speed. Figure 3 shows the detail of the start transient. A small amount of oscillation on top of the transient peak is clearly visible. Small glitches frequent in Figure 1 are noise and assumed to be from elevator operations.

[2] Power factor is defined as a ratio of the real power to the apparent power ($pf = P/P_A$). Utility meters record P value only. But there typically is a power factor penalty in the rate schedule, because the utility supplier should have the generation and transmission capacity to meet the apparent power demand, not the real power demand recorded by the meter.

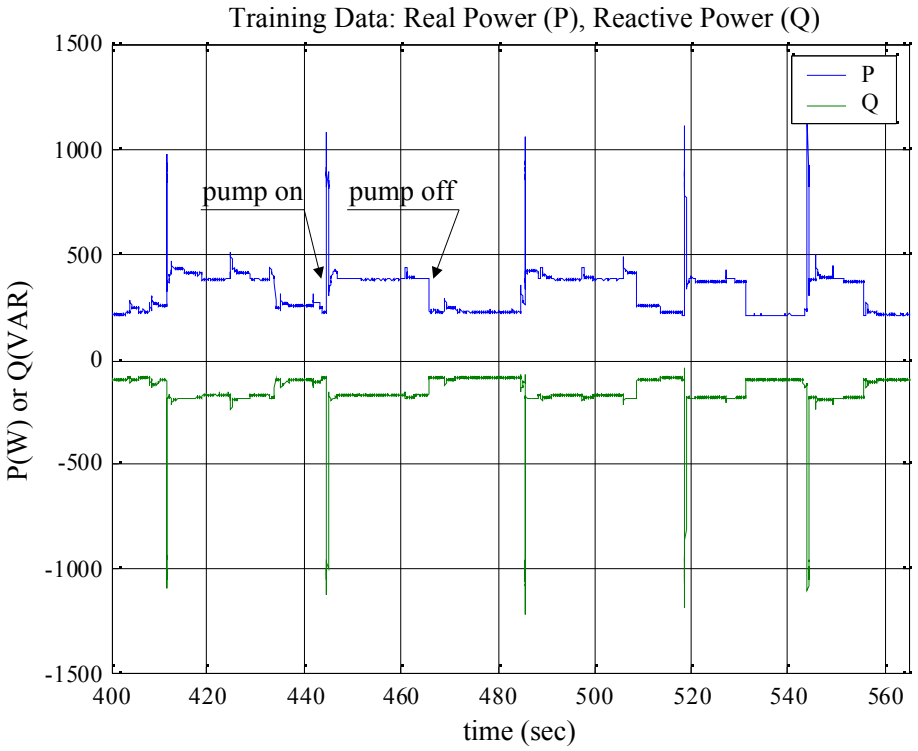


Figure 1. Training data of the cooling tower (condenser) pump 1

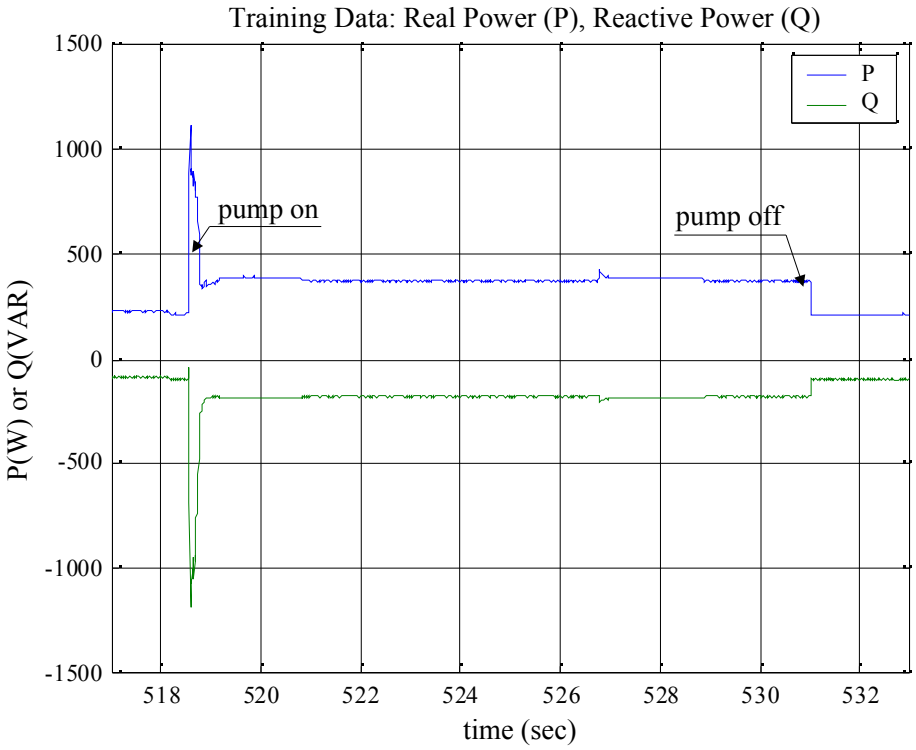


Figure 2. A step test of the cooling tower pump 1

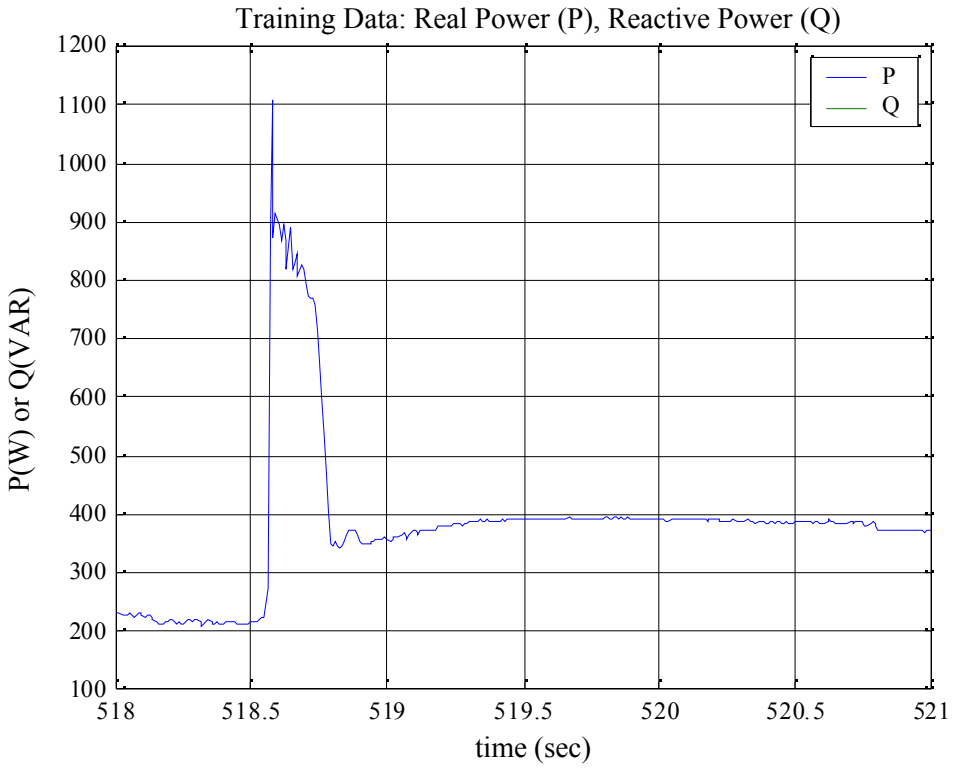


Figure 3. Start transient of the cooling tower pump 1

Figure 4 shows the training data of the cooling tower fan 1. Again five on/off sequences were applied. Figure 5 shows one of the step responses. Because it has a 20 hp motor, it shows higher peak and steady state P value compared to the condenser pump. Interestingly, P shows a plateau of 1150 W during the transient, which lasts about a second. Figure 6 shows the enlarged view of the step transient. This plateau is not observed in the pump step transient. Thus it is assumed that the fan motor start circuit has a type of starting current limiting device (such as an autotransformer which reduces the voltage across the winding during start up and disconnected once the motor accelerates, or a part-winding controller in which only part of the motor winding is initially connected to the line and the rest of the winding is connected to the line after a time interval of 1 to 3 sec.).

Figure 7 shows the case when the cooling tower fan 1 is turned on while the pump 1 is running. Each distinctive step transient profile is clearly seen. Thus each signal can be superimposed on top of another and can be distinguished if they are separated by a reasonable amount of time (on the order of each step transient interval).

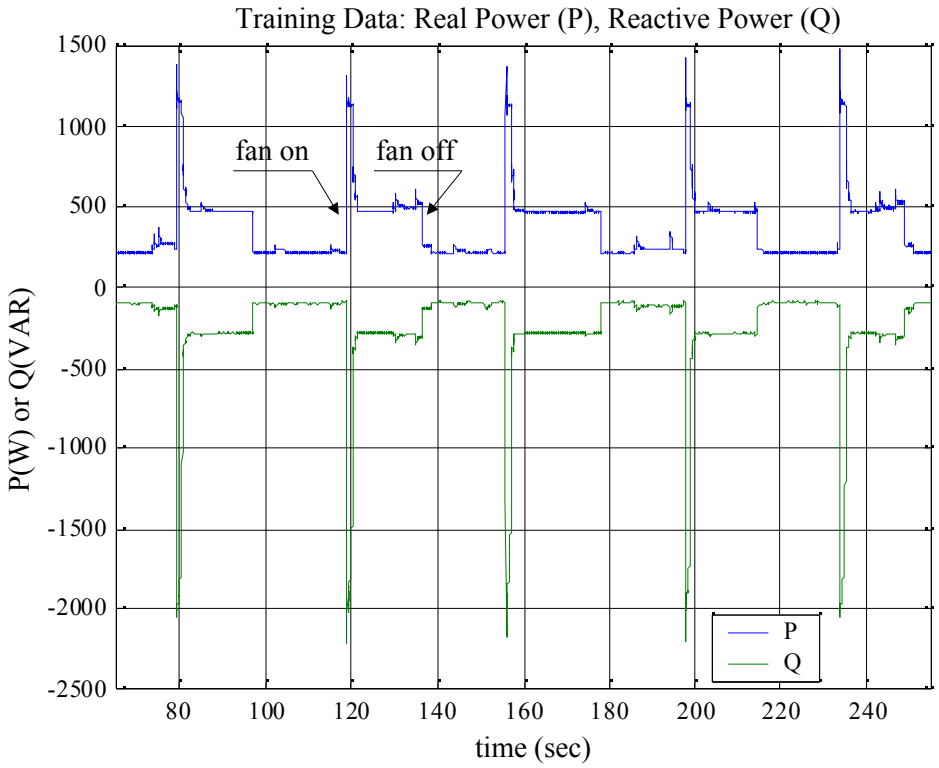


Figure 4. Training data of the cooling tower fan 1

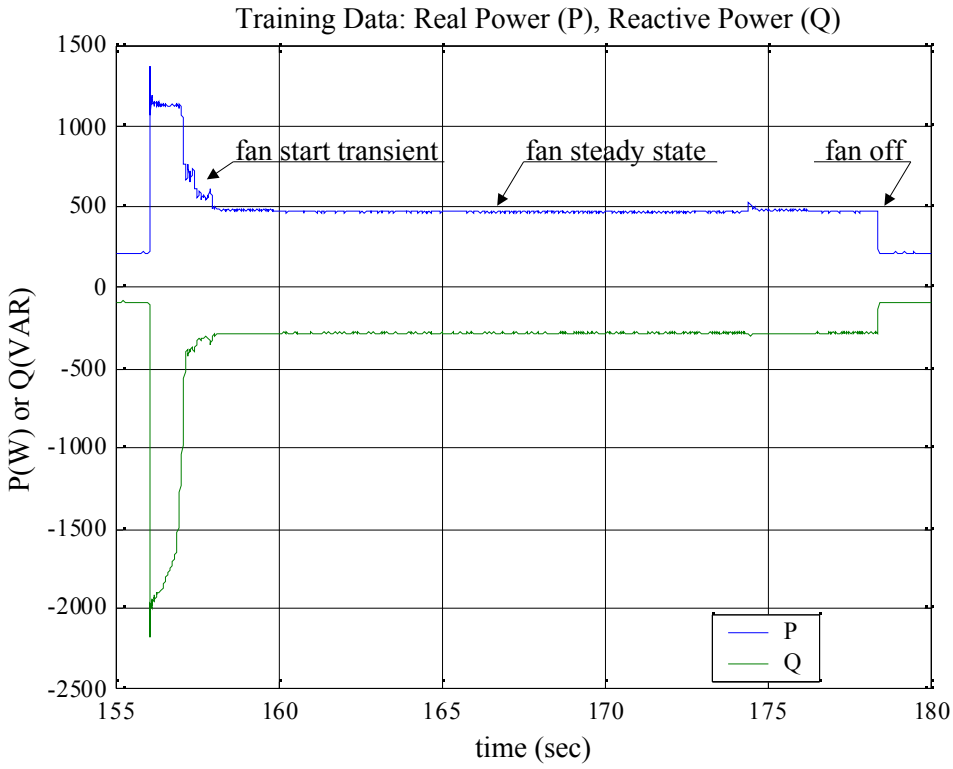


Figure 5. A step test of the cooling tower fan 1

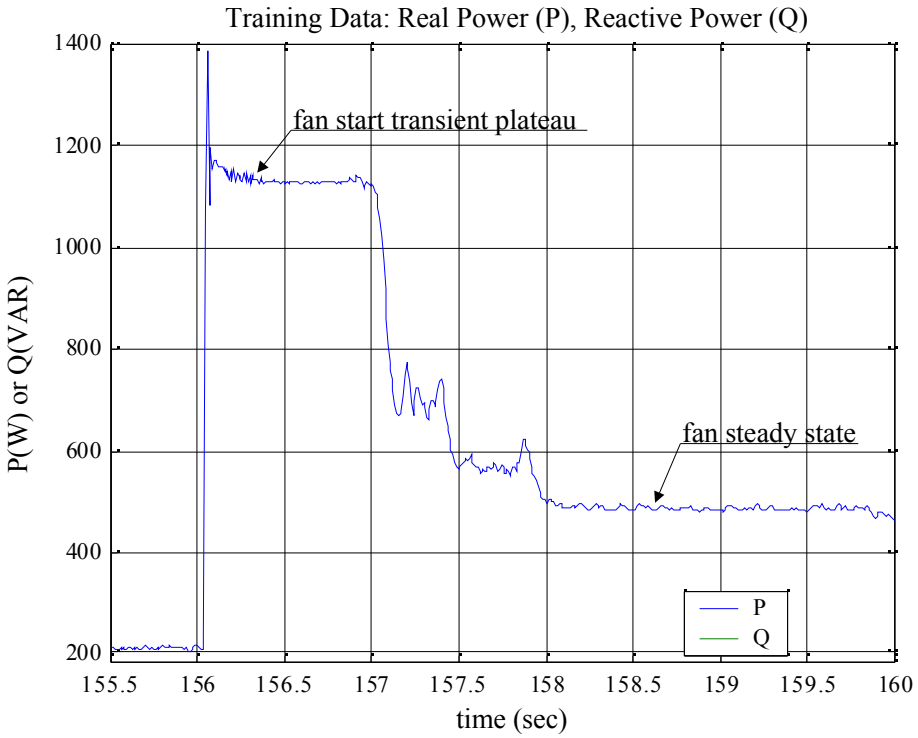


Figure 6. Start transient of the cooling tower fan 1

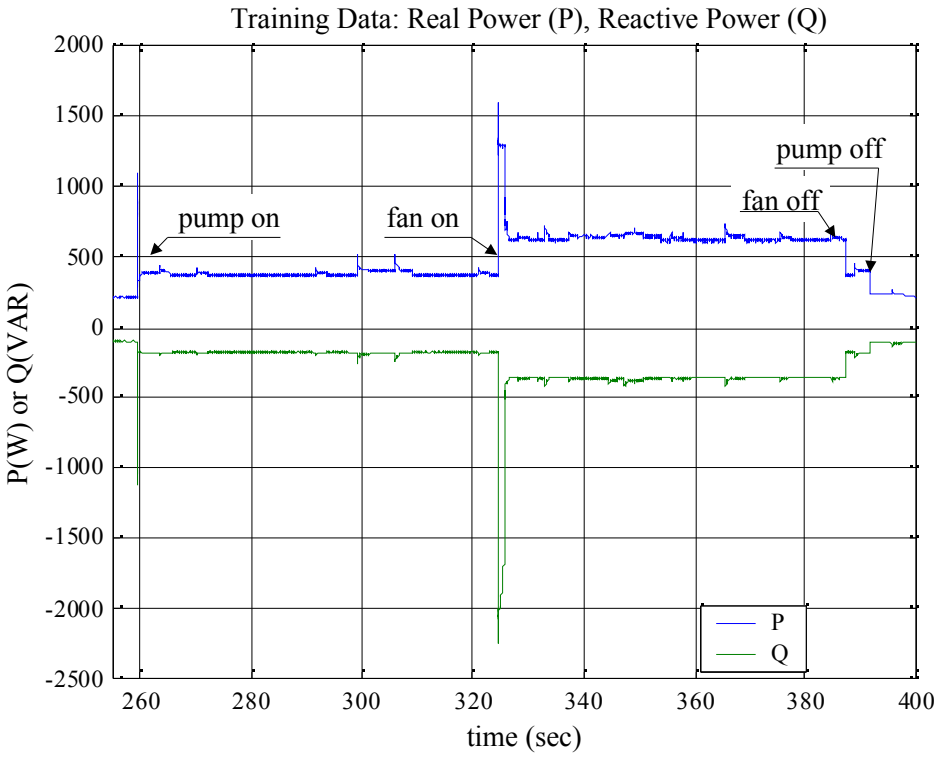


Figure 7. Training data of the cooling tower pump 1 and fan 1

Figure 8 shows the training data of the cooling tower pump 2. Figure 9 shows its step transient. The overall response shape is similar to the pump 1, although specific shapes or values are different from each other. This suggests that NILM data can be used to identify the type of the equipment based on the overall response shape and also can pinpoint a specific piece of equipment if it has numerical data of the equipment's representative response.

Figure 10 and 11 show the training data of the cooling tower fan 2. The overall behavior is similar to the fan 1, though specific values and shapes are different. Small peaks right after the transient plateau are noteworthy.

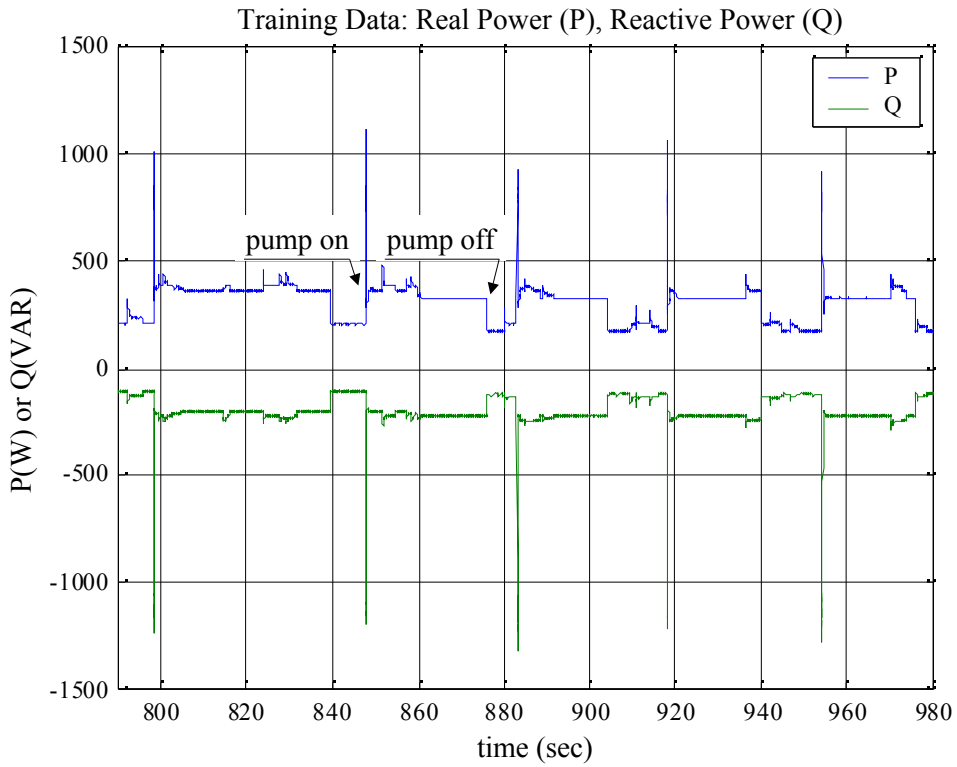


Figure 8. Training data of the cooling tower pump 2

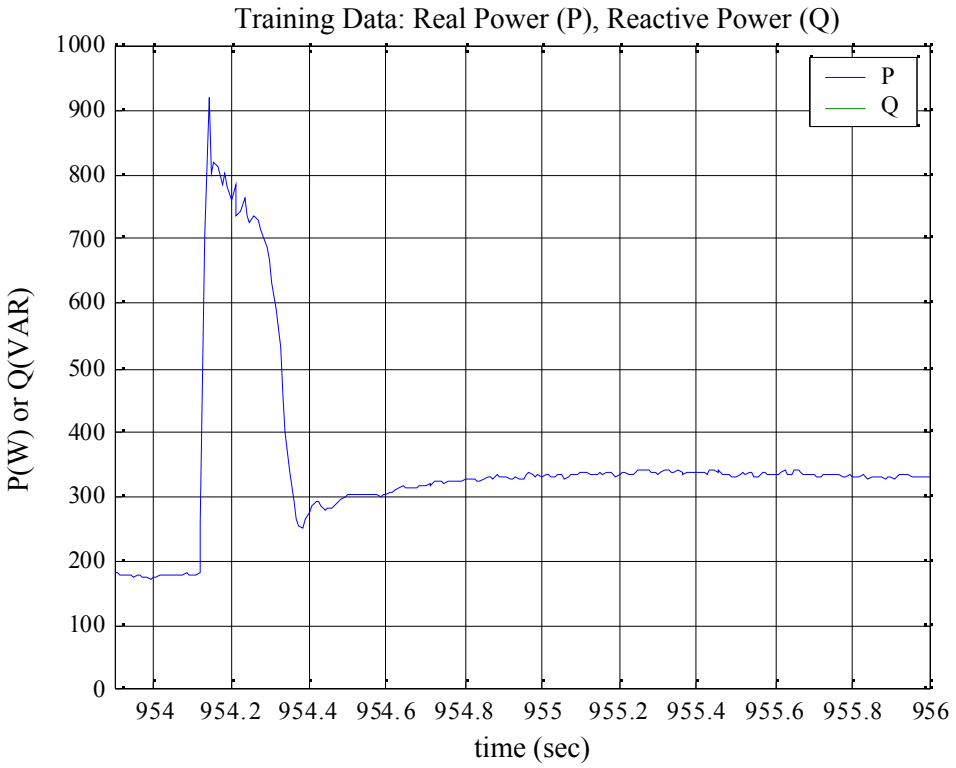


Figure 9. Step transient of the cooling tower pump 2

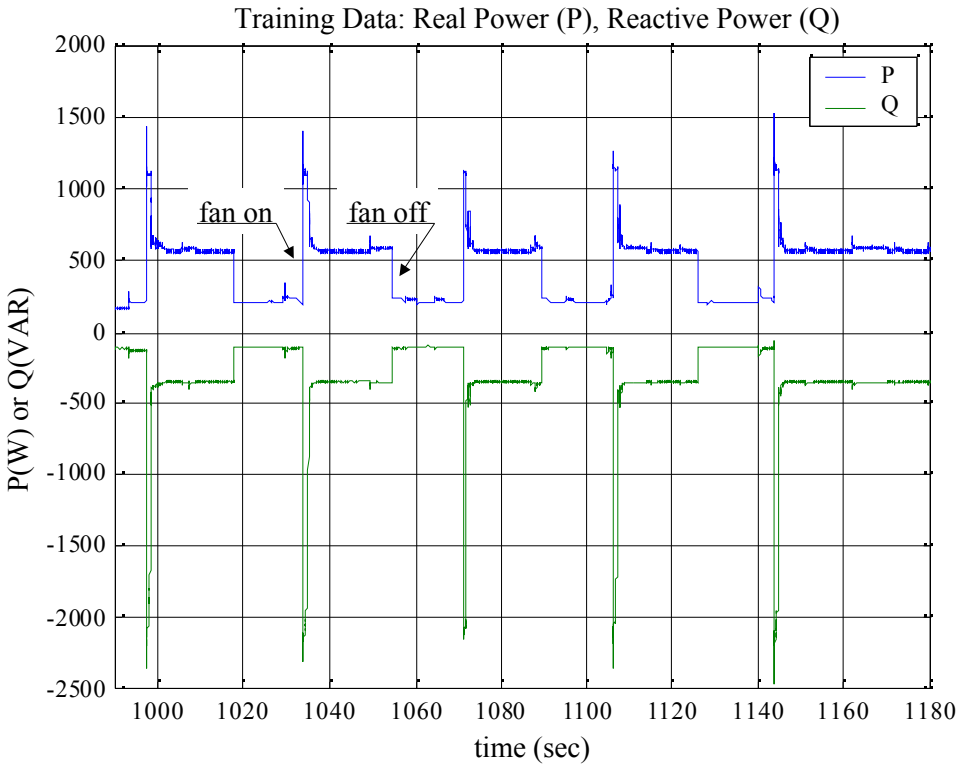


Figure 10. Training data of the cooling tower fan 2

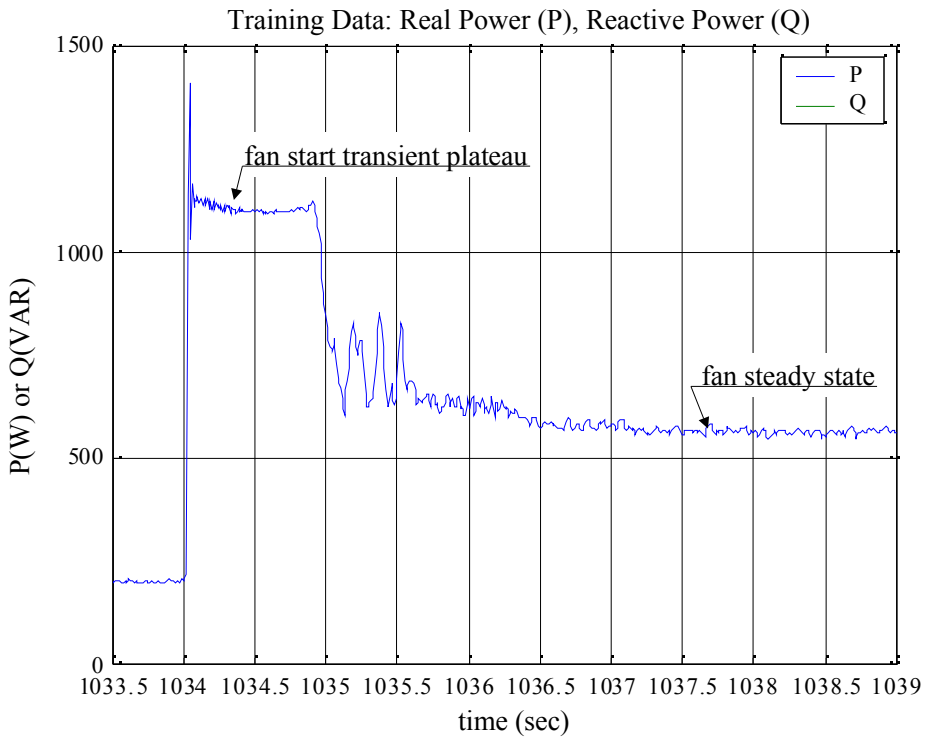


Figure 11. Step transient of the cooling tower fan 2

Chiller Pumps

The building has two chillers and each has a 10 hp centrifugal water pump rated at 450 gal/min. Figure 12 shows the training data for the chiller pump 1. The step transient is shown in Figure 13. The overall behavior is similar to the cooling tower pumps, because they have the same 10 hp capacity and handle the same liquid (water).

Figure 14 and 15 show the training data of the chiller pump 2. Compared to the chiller pump 1, it does not show the oscillations during the transient peak.

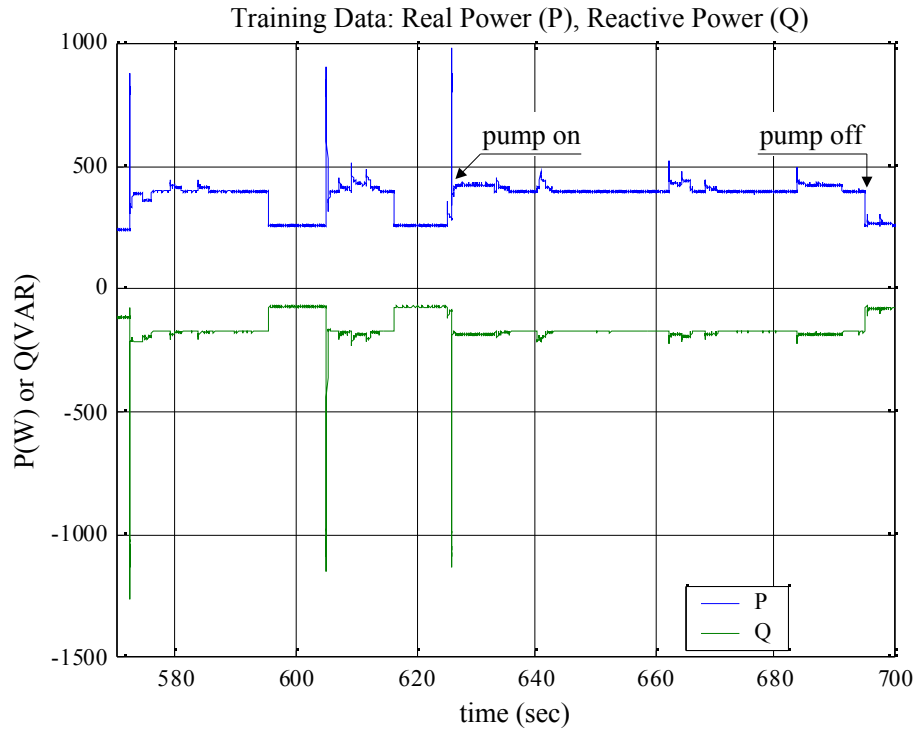


Figure 12. Training data of the chiller pump 1

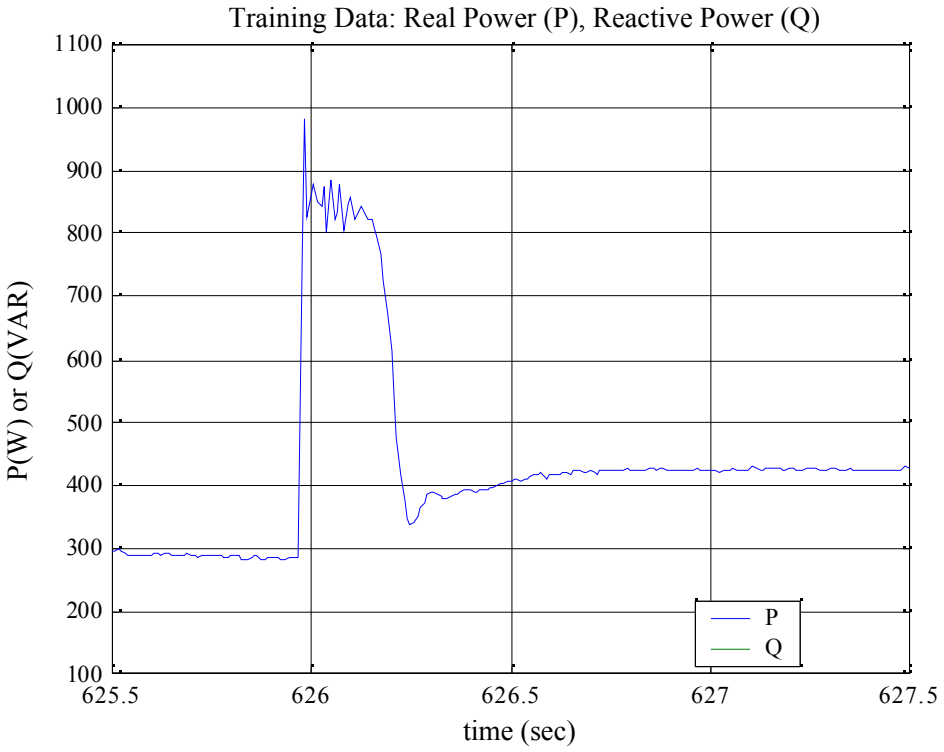


Figure 13. Step transient of the chiller pump 1

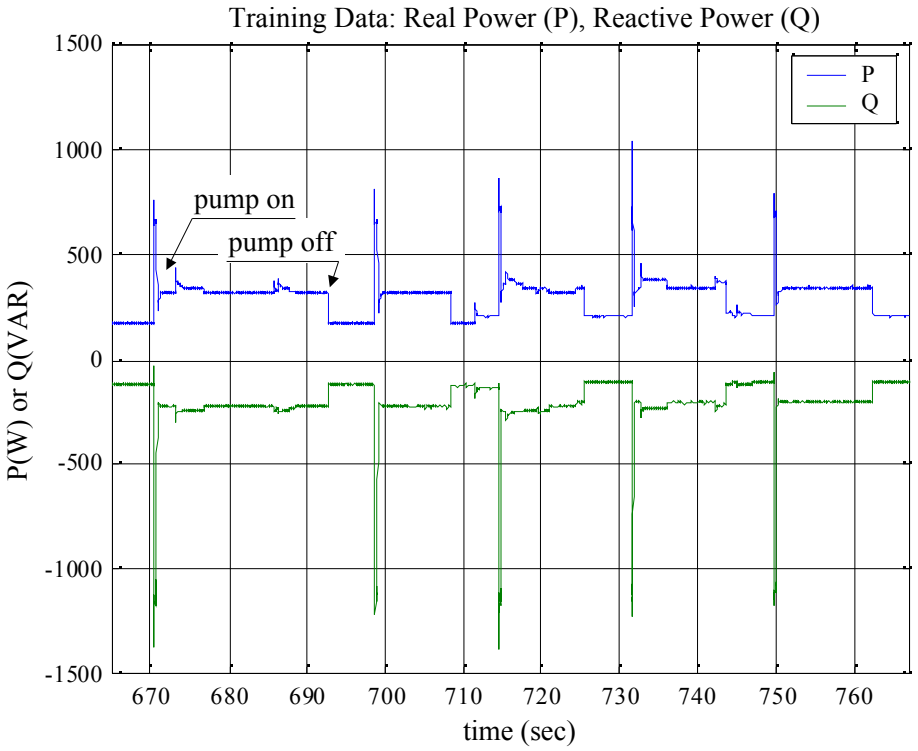


Figure 14. Training data of the chiller pump 2

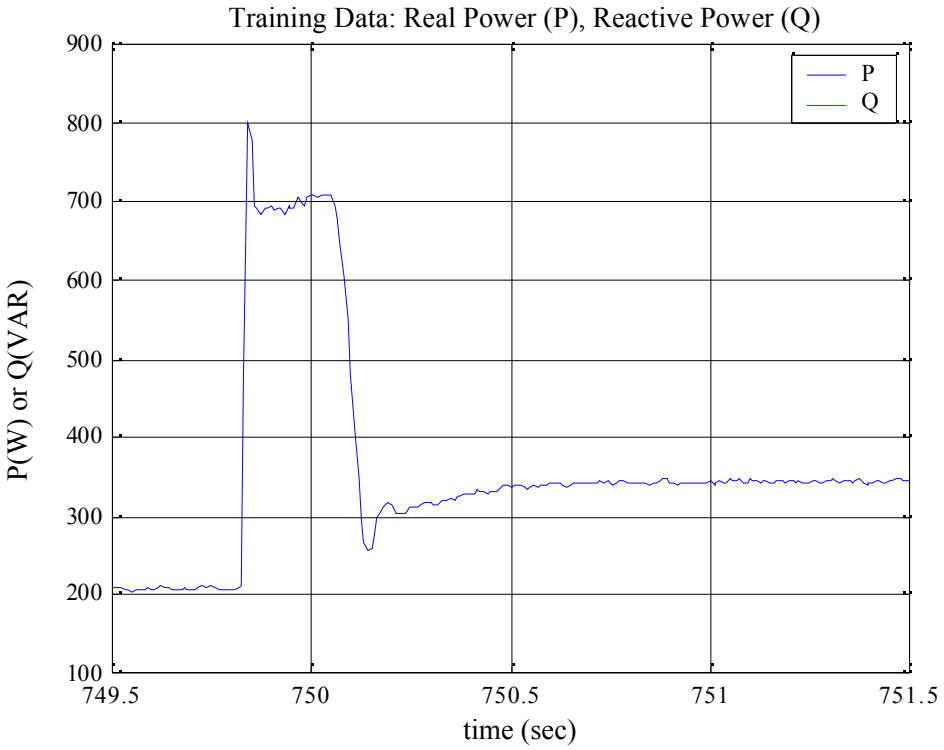


Figure 15. Step transient of the chiller pump 2

Supply and Return Fan

The building has a 100 hp supply fan (SF) and a 75 hp return or exhaust fan (RF). Each fan is equipped with a variable frequency drive (VFD). Normally, each fan is slowly accelerated to its steady state value by giving a ramp velocity command to the VFD. However for training purposes, the VFDs are bypassed and each motor is given a step command to start up. Figure 16 shows the training data of the SF. Two distinct step changes are notable. It is assumed that the motor has a two-stage wye-delta starting circuit. The first step is associated with the wye starter. Each motor phase winding is connected across a line and a neutral, effectively reducing the starting current. After a certain amount of time (or acceleration) each winding is connected across the lines, doubling the line voltage and hence generating a full torque. Figure 17 shows the wye transient, while Figure 18 shows the delta transient. Figure 19 shows the detailed view of the delta transient peak.

It is interesting to note that the SF shows a sustained oscillation at steady state. Figure 20 shows the detailed view of the steady state oscillation. From the figure, the frequency of the oscillation is assumed to be around 3 Hz. Figure 21 shows the power spectral density[3] of the oscillation. The harmonics up to 7^{th} ($3\times7 = 21$ Hz) are clearly visible. Three Hz is equivalent to 180 rev/min, which is close to the typical fan shaft speed of 200 rev/min. Thus it is assumed that the oscillation is related to mechanical imbalance, friction or a flaw of the fan shaft. Although not strong, a 30 Hz harmonic is also observable, which can be contributed either from the 10^{th} harmonic of a 3 Hz oscillation or from the motor output shaft which typically runs at 1800 rpm (30 Hz).

[3] If $x(t)$ is a power signal, let $x_\gamma(t)$ be the truncated version *of* $x(t)$, $x_\gamma(t) = x(t)$, $|t| \le \gamma$ and $x_\gamma(t) = 0$, otherwise. Then the power spectral density $S(\omega)$ is defined as: $S(\omega) = \lim_{\gamma\to\infty} \frac{1}{2\gamma} \left| X_\gamma(\omega) \right|^2$, where $X_\gamma(\omega)$ is the Fourier Transform of $x_\gamma(t)$. The power spectral density is determined from the square of the amplitude spectrum of the truncated power signal, normalized by the length of the spectrum 2γ.

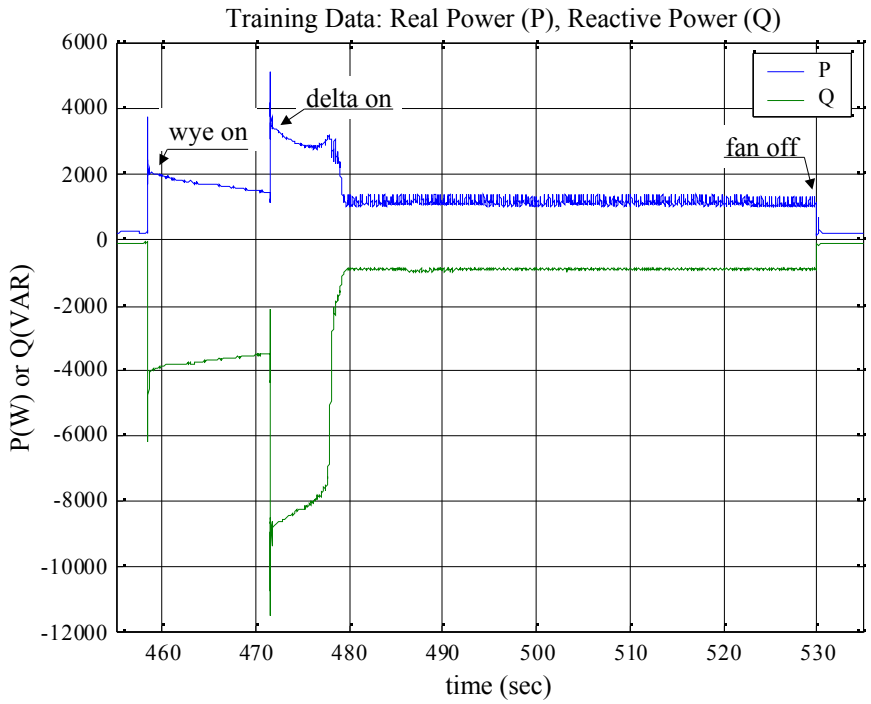


Figure 16. Training data of the supply fan

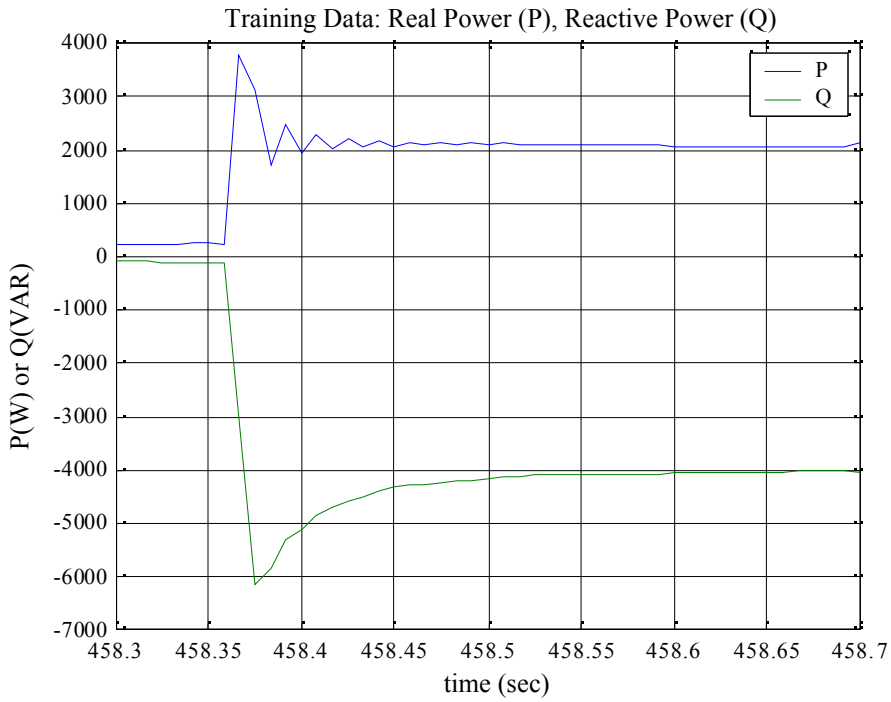


Figure 17. Wye transient of the supply fan

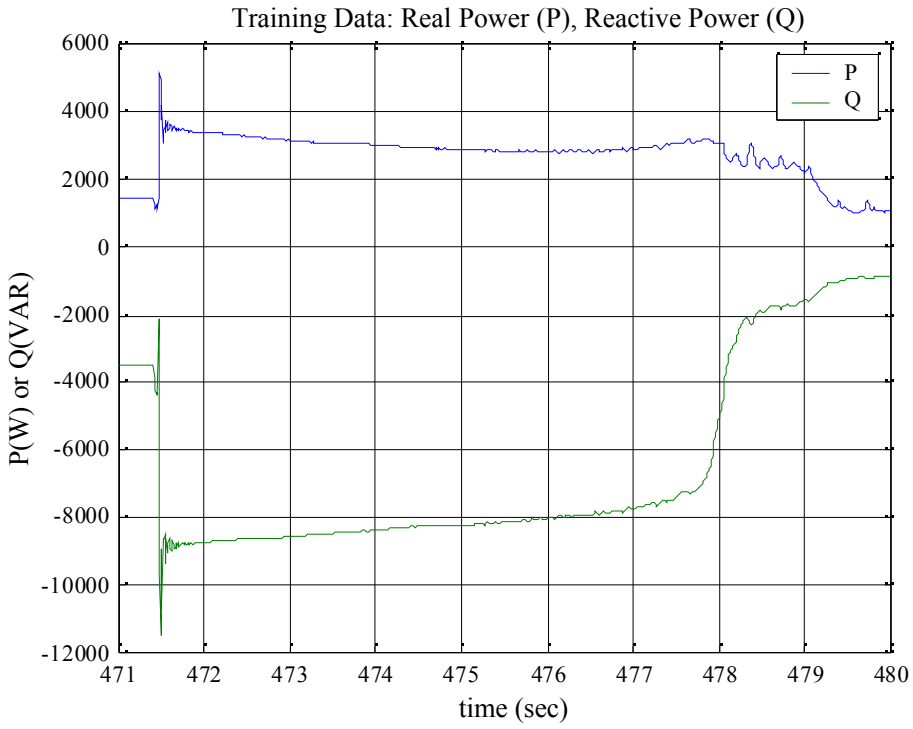


Figure 18. Delta step of the supply fan

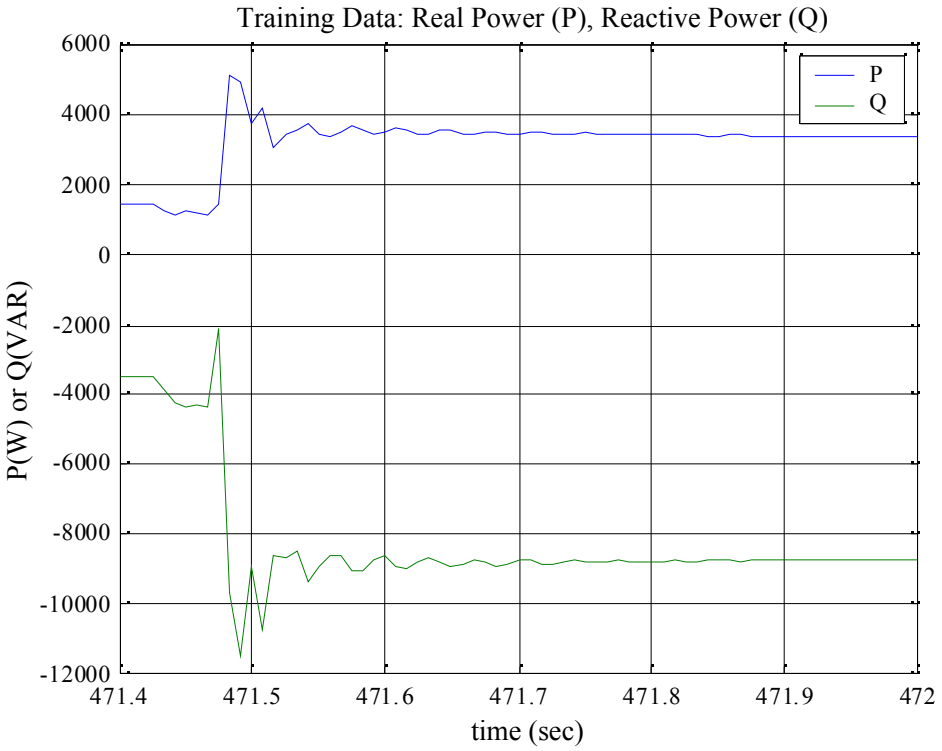


Figure 19. Delta transient of the supply fan

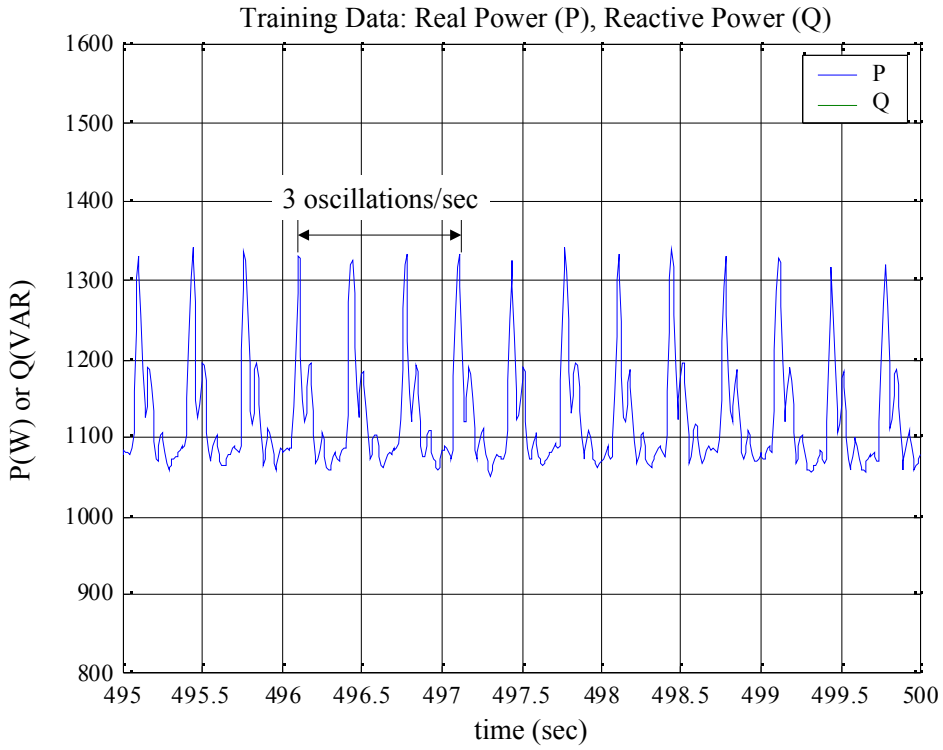


Figure 20. Oscillation of the supply fan during steady state

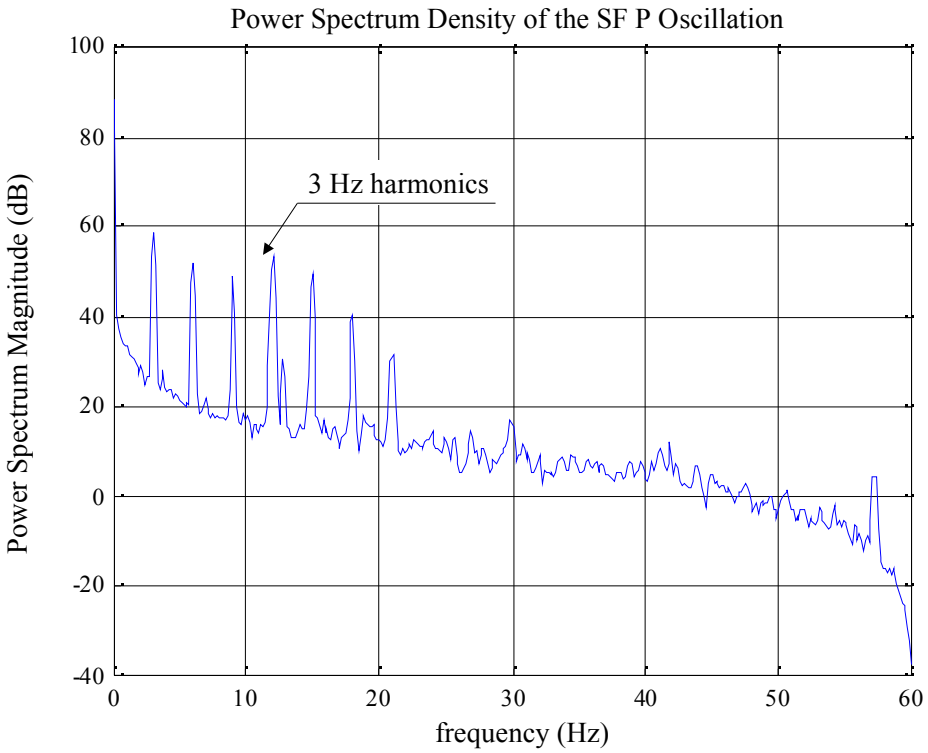


Figure 21. Power Spectrum Density of the supply fan P oscillation

A similar test was also performed on the return fan. Figure 22 shows the training data of the return fan. It appears to have the wye-delta starting circuit as the supply fan. But as Figure 23 shows, it shows no transient peak during the wye start. The delta start behavior, shown in Figure 24, is similar to the SF. Interestingly, the RF also shows the oscillation behavior in its steady state. Figure 25 shows the detailed view of the oscillation. The figure suggests that the oscillation is mainly composed of two parts: one about 2.5 Hz and the other about 11 Hz. The 2.5 Hz component may be related to the fan shaft rotation, while 11 Hz to the motor shaft rotation. The power spectrum density of the steady state P values, shown in Figure 26, revels the harmonics of 2.5 Hz (5.0, 7.5, 10.0, 12.5, 15, 17.5, 20, …) and the harmonics of 11 Hz (22, 33, …).

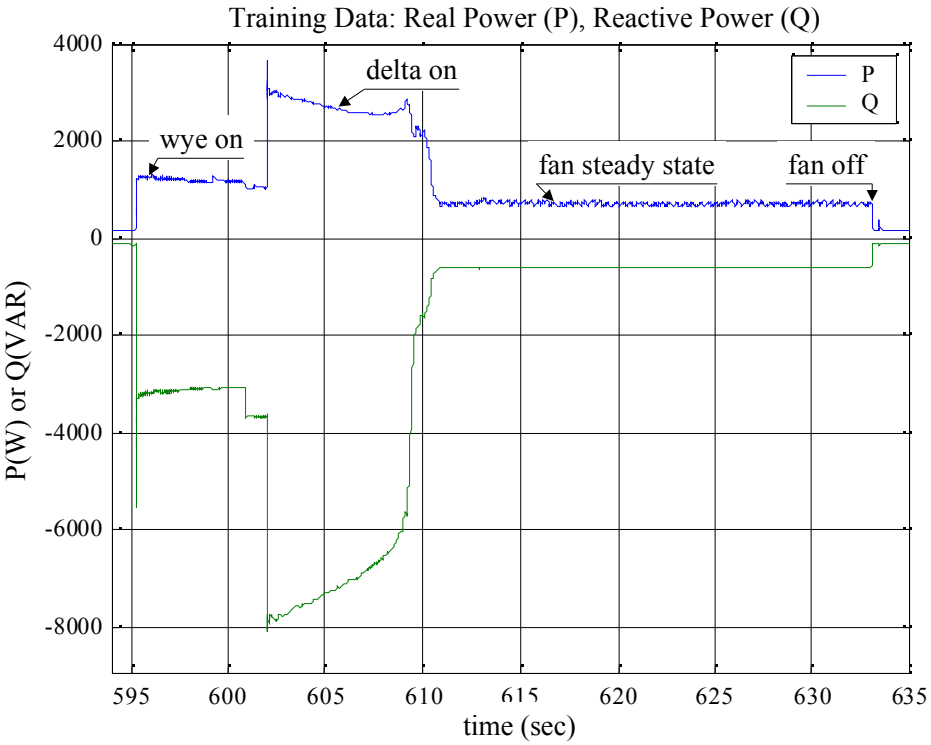


Figure 22. Training data of the return fan

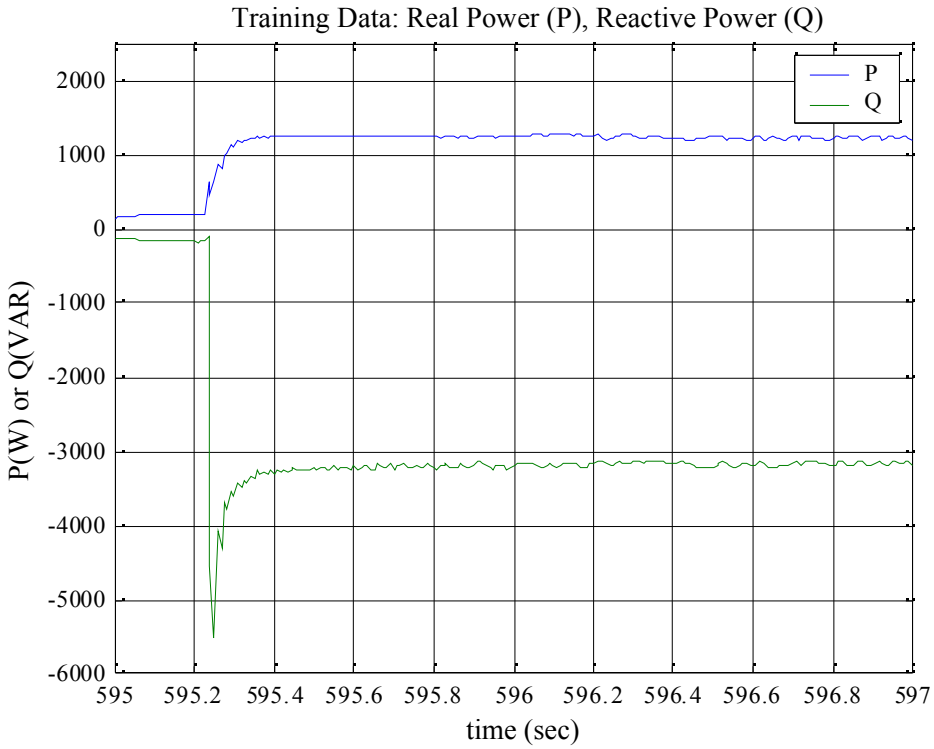


Figure 23. Wye transition of the return fan

Figure 24. Delta transition of the return fan

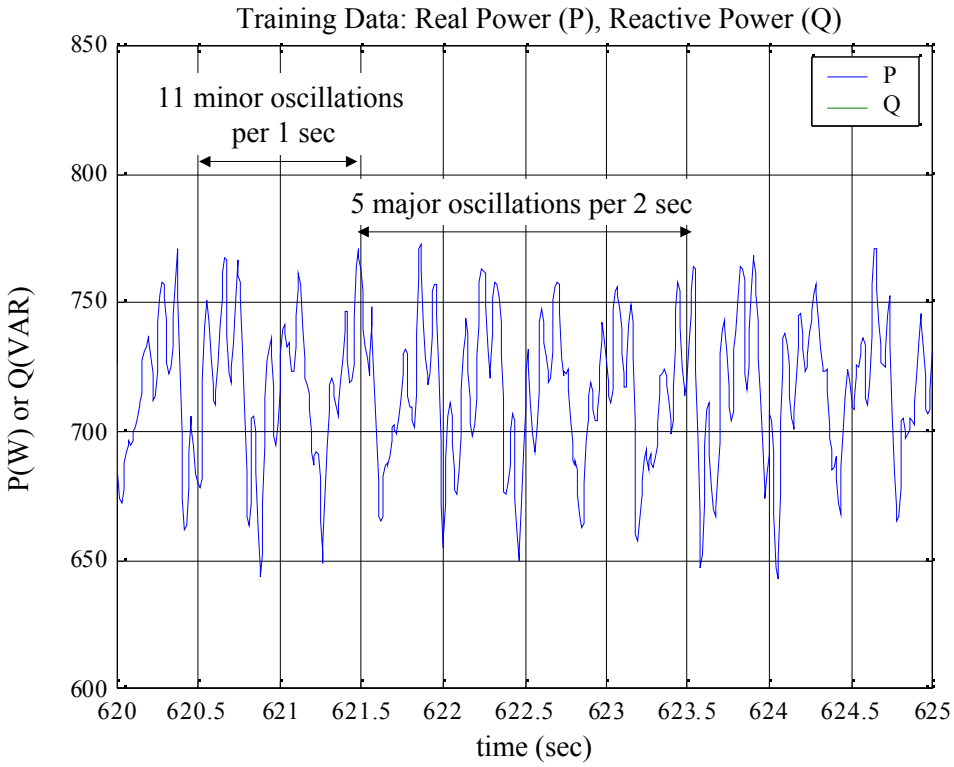


Figure 25. Return fan oscillation in steady state

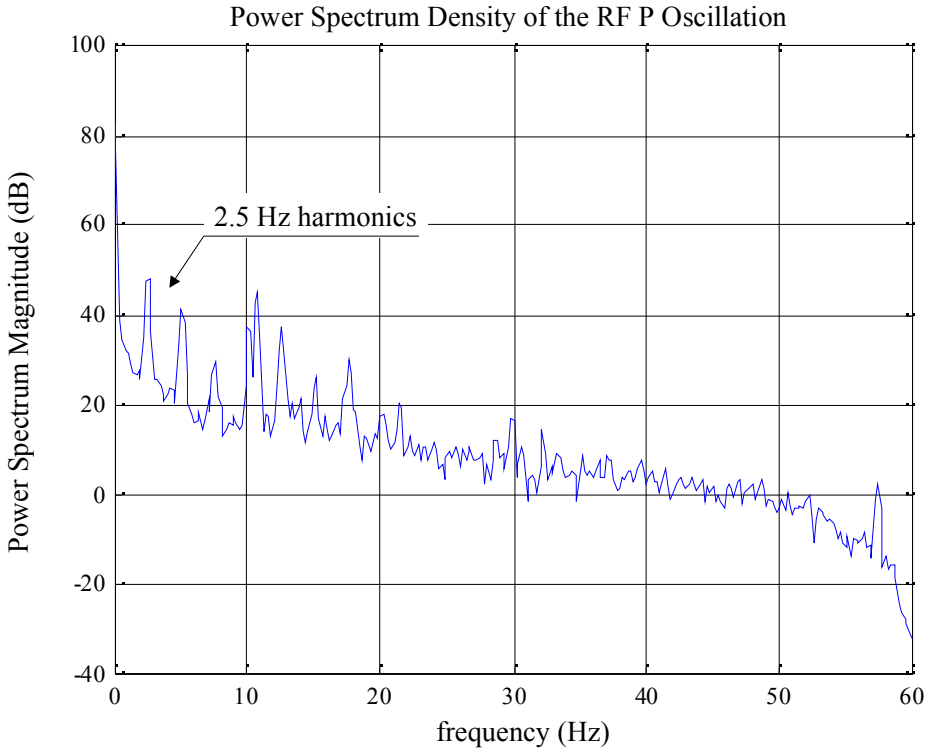


Figure 26. Return fan P power spectrum density

The VFD startup behavior of the supply and return fan was also recorded during the training data collection. Figure 27 shows the VFD start up of the supply and return fan. It is clearly seen that the VFD slowly accelerates the motors toward their steady state values. It is also interesting to observe the increasing amount of oscillation during the acceleration. Figure 28 shows the oscillation behavior in the middle of VFD start up. A frequency component of about 1.65 Hz is clearly visible, along with other high frequency components. Figure 29 reveals a high frequency component around 20 Hz superimposed on the 1.65 Hz component. The low frequency component is suspected to be related to the fan shaft speed. Thus the frequency of the low frequency component should increase as the VFD accelerates. Figure 30 shows the oscillation at the early stage of VFD acceleration. It yields about 1.4 Hz of the low frequency component. In contrast, in the later stage, as shown in Figure 31, 2.3 Hz is clearly visible. Thus we can conclude that the low frequency component is related to the mechanical rotation of the fan shaft. Figure 32 shows the power spectral density of the P value during SF/RF VFD startup. Because of the changing nature of the lowest frequency component, it is not clearly visible compared to the SF or RF only case. It is interesting to observe many high frequency components between 30 and 50 Hz.

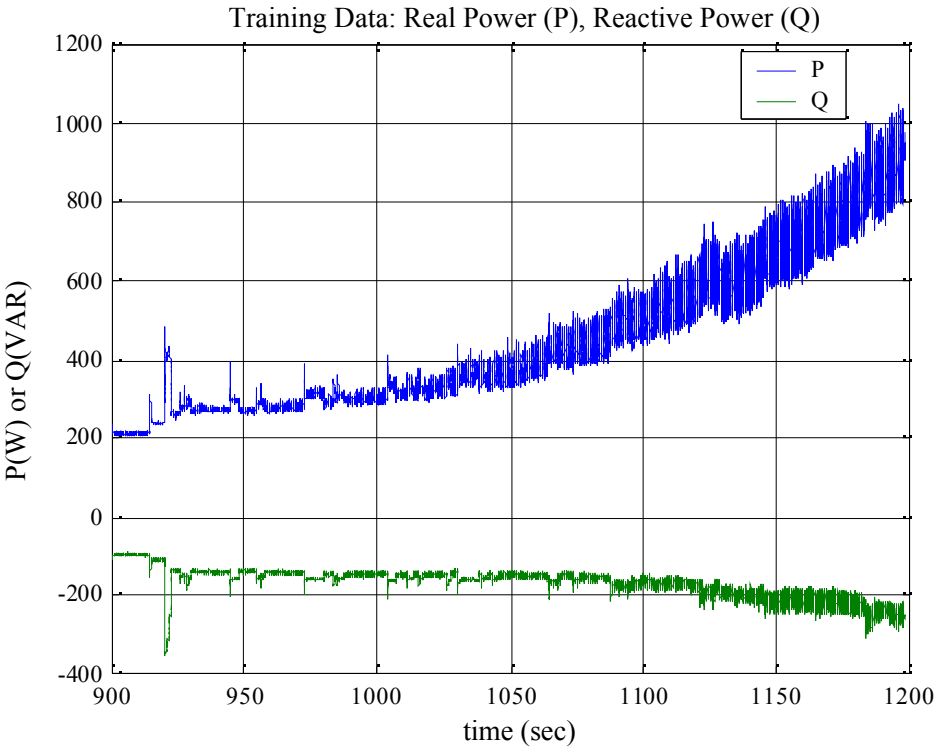


Figure 27. Supply and return fan VFD start up

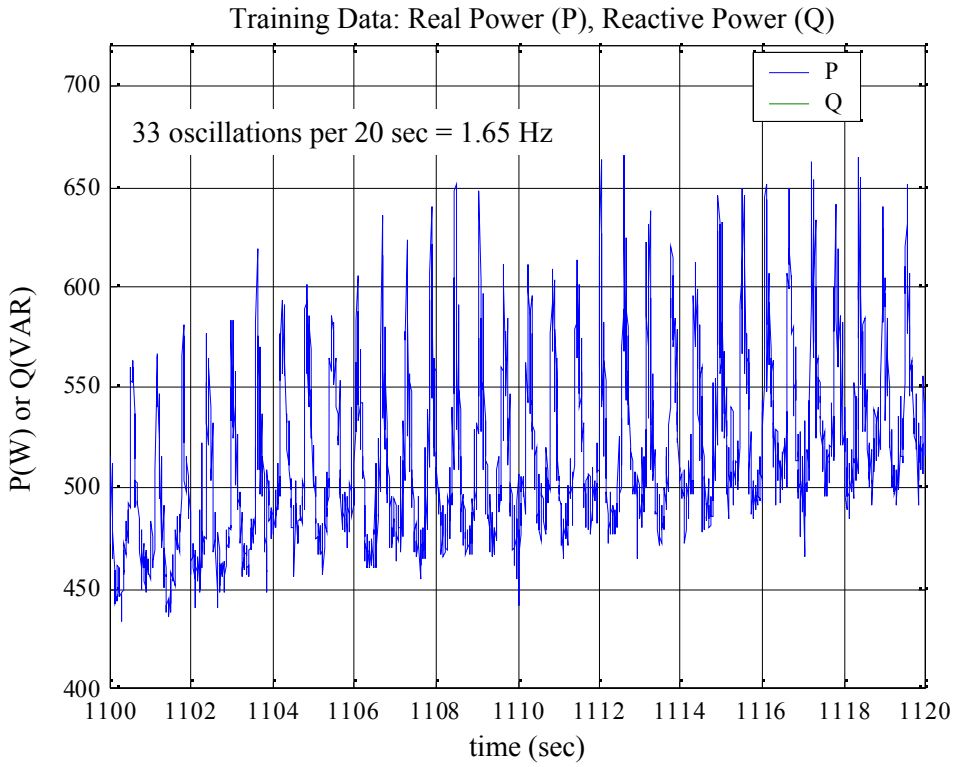


Figure 28. Supply and return fan VFD start up oscillation

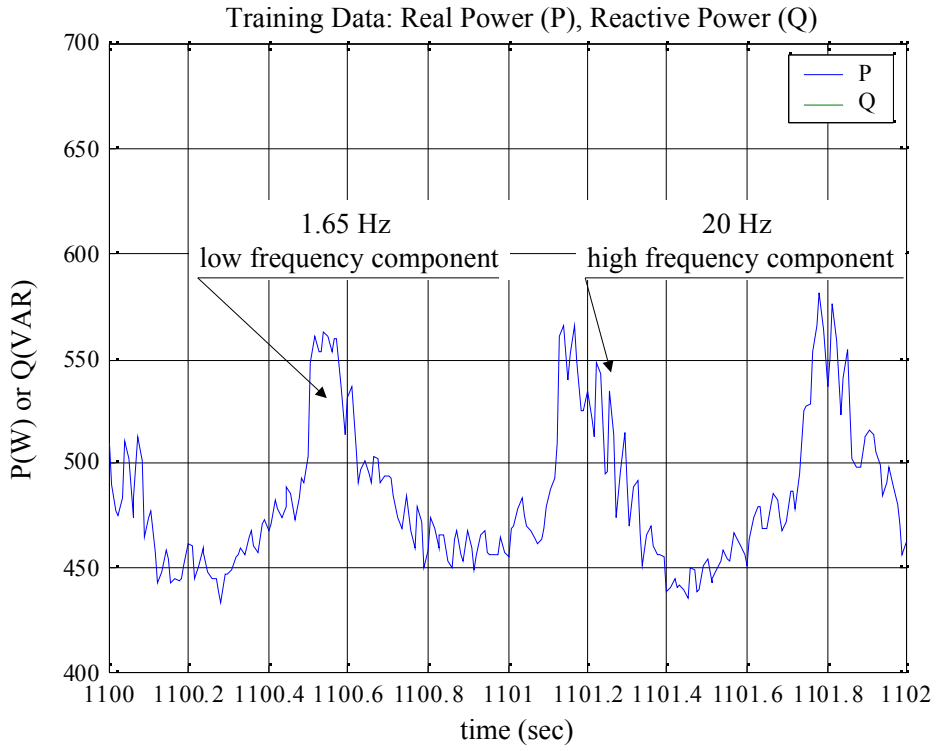


Figure 29. Detailed view of the supply and return fan VFD start up oscillation

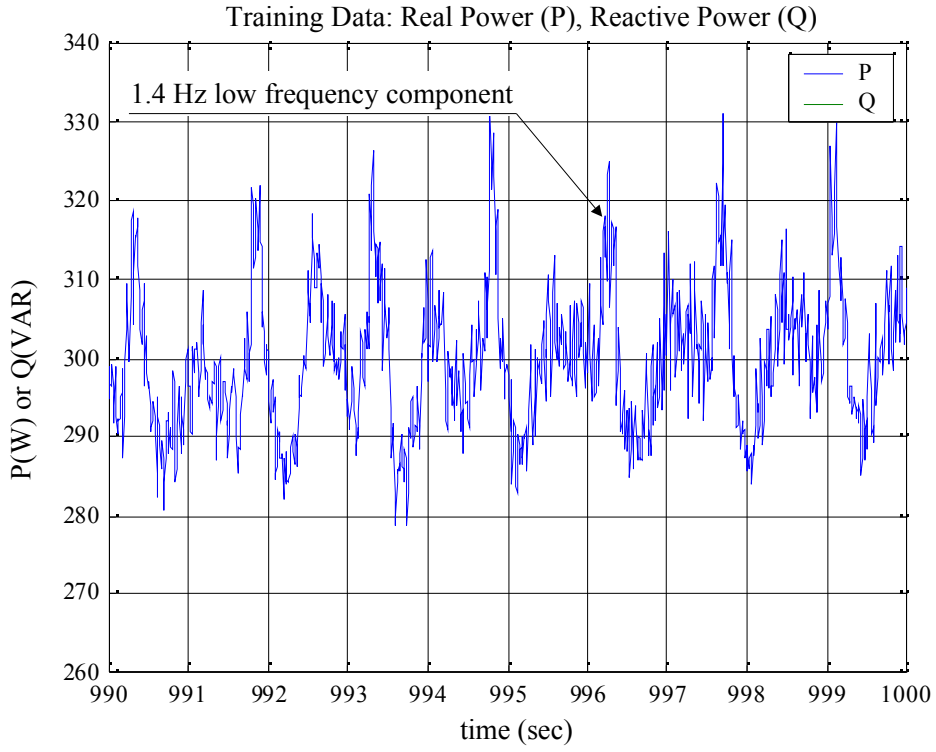


Figure 30. Supply and return fan VFD start up oscillation in its early stage

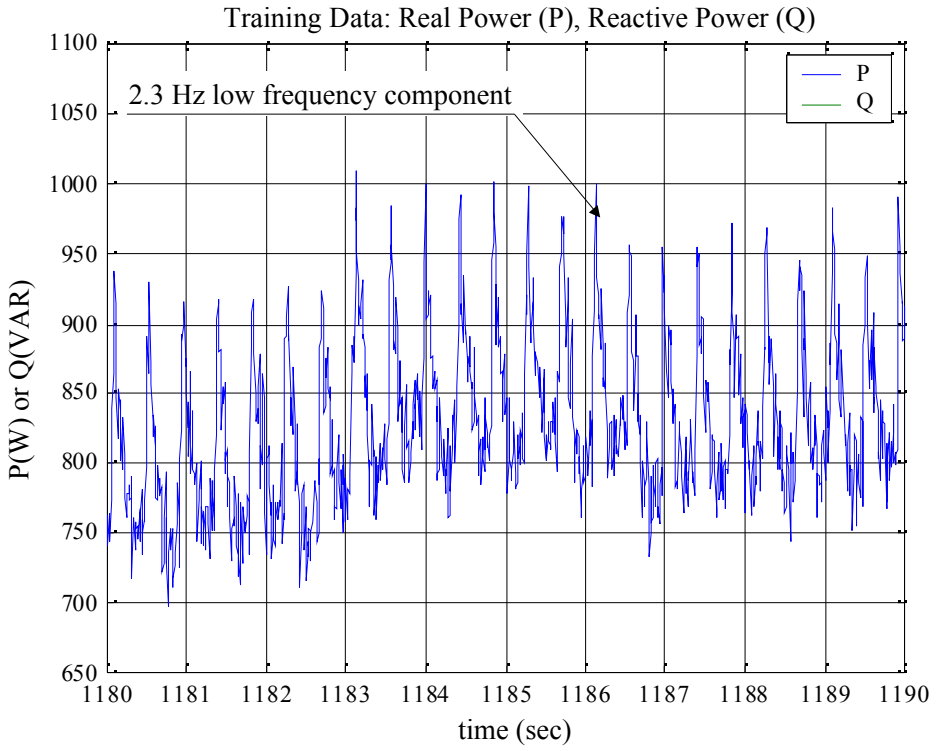


Figure 31. Supply and return fan VFD start up oscillation in its later stage

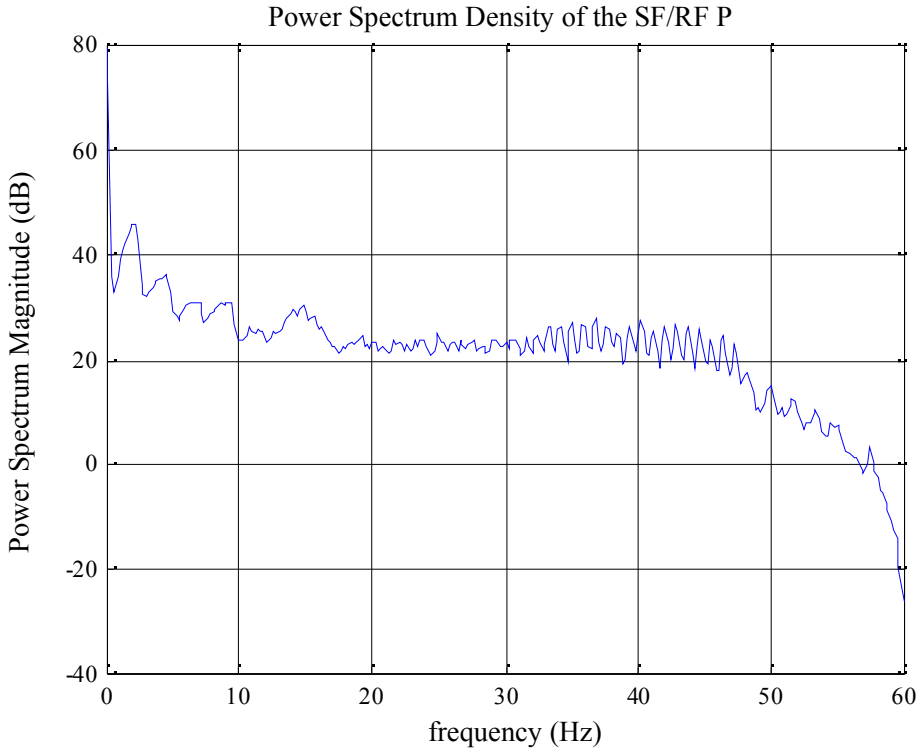


Figure 32. Supply and return fan VFD start up power spectrum density

Steady State Loads

Steady state loads of the HVAC equipments are summarized in Table 2. The average value of the real power P can be used to measure power consumption. Also, the average values of P and Q are used to compute a power factor of equipment, i.e. the ratio of the power consumed to the power drawn by the equipment. Standard deviations are usually 1 to 3 % of the average values. However the supply and the return fan shows higher standard deviations (7.8 % and 5.8 %, respectively). This is assumed to be related to the fan shaft oscillations, as discussed in the previous section. Another interesting observation is that fans usually shows higher standard deviations compared to pumps. This may be related to the nature of the motor itself or the nature of the fluid it delivers. Normally, water provides better damping than air, thus less vibration. It is also surprising to note differences between supposedly identical equipment. For example, CTWR fan 1 and fan 2 show a big difference in their steady state value. A more detailed study is required, if the difference is the reality or an error of the measurement or the training sequence.

Table 2. Steady state loads of HVAC equipments (multiply P and Q by 50)

Name	Rating (HP)	P (W)		Q (VAR)		Power Factor
		Ave	Std	Ave	Std	
CTWR pump 1	10	166.6	1.93	-68.6	1.94	0.925
CTWR pump 2	10	120.0	2.10	-113.7	1.94	0.726
CTWR fan 1	20	251.8	4.99	-183.7	2.26	0.808
CTWR fan 2	20	354.3	7.04	-236.4	2.61	0.832
CHW pump 1	10	180.0	2.22	-74.8	2.03	0.923
CHW pump 2	10	109.4	1.73	-118.0	2.03	0.680
Supply fan	100	924.8	72.3	-771.2	17.7	0.768
Return fan	75	518.4	29.8	-473.7	7.73	0.738

Real Time Plots

With the installation of the DSL line to the mechanical room, the 160 Sansome St. NILM server is online and real time plots can be viewed on a Netscape web browser by typing http://66.120.20.99/. Figure 33 shows the shot of the web page taken at 11:35, 06/11/01, Pacific Time. The page currently shows the real and the reactive power and their third harmonics. Future version will display higher harmonics. Figure 34 show the last 24 hour plot taken at 5:30 following morning. The plot suggests that most of the electricity is consumed during the occupied period. P drops sharply at night. It is interesting to note the leveled nature of the power consumption during day, which is a typical pattern when chillers are not running and the major HVAC loads are fans throughout the day.

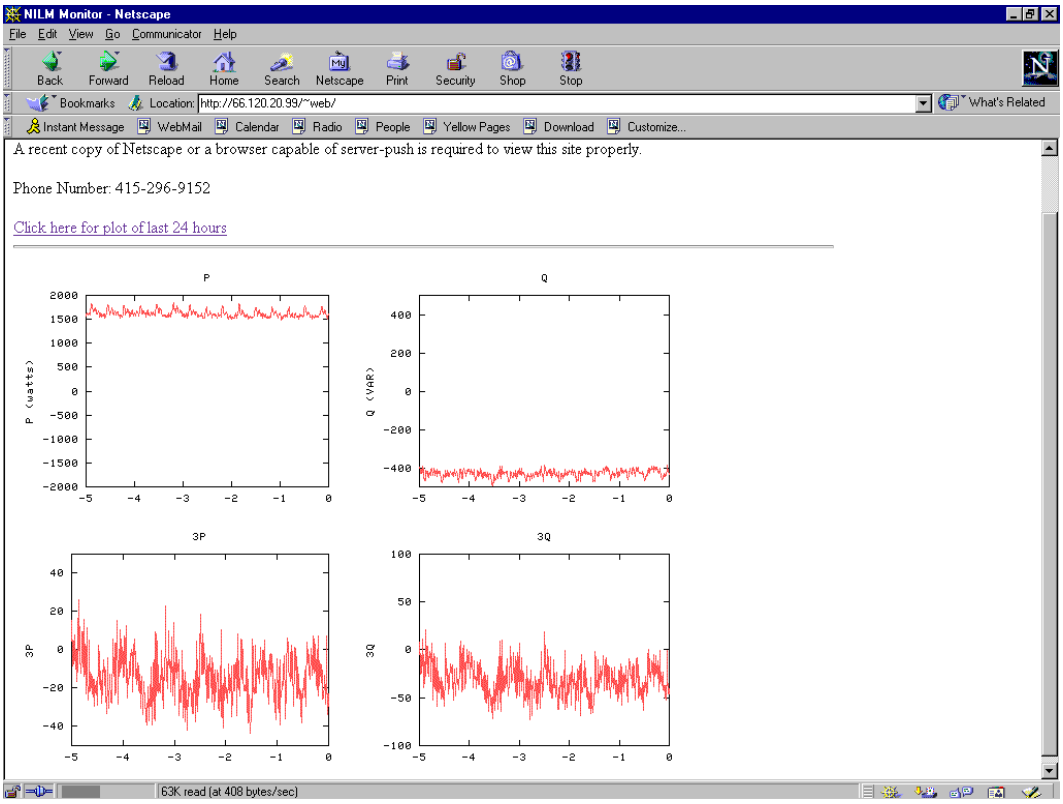


Figure 33. Real time graphs of 160 Sansome St. NILM web page

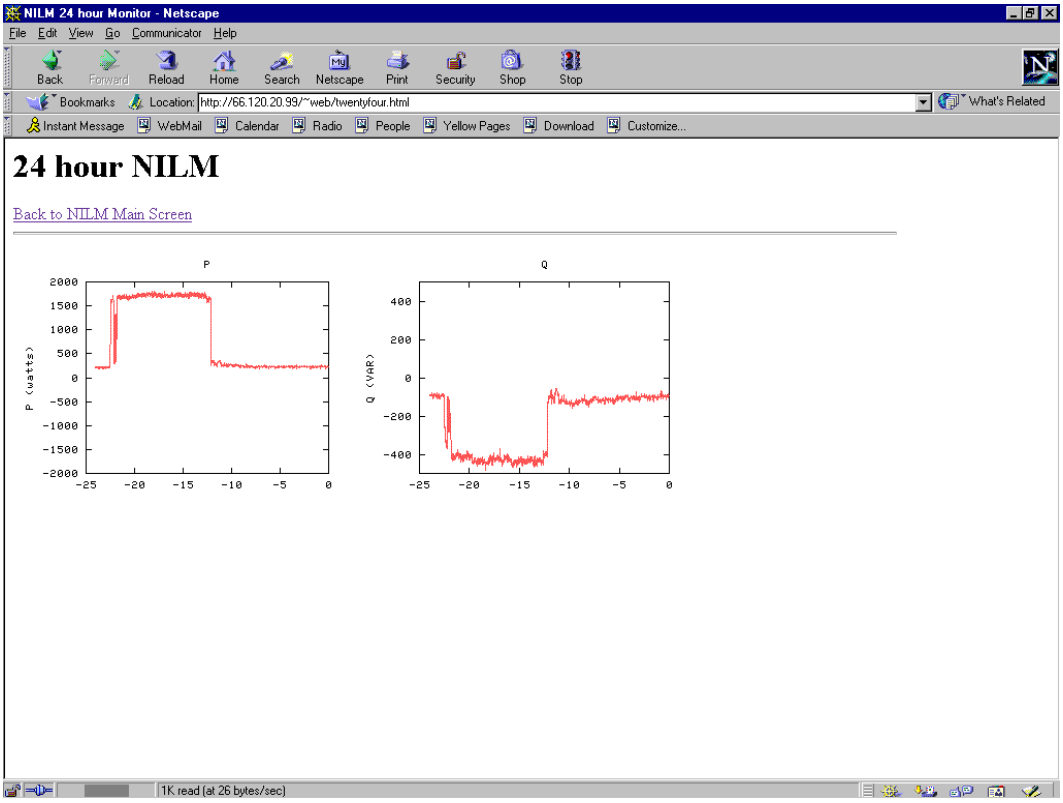


Figure 34. Plot of the last 24 hours

One Day Training Data

For the purpose of the load pattern and the individual load profile identification, training data were collected each day between 7/1/01 and 7/3/01 (Sunday – Tuesday) with a sampling frequency of 1 Hz. Figures 35 and 36 show the data for July 3 and July 2, respectively. They both show the typical weekday load pattern with a single chiller (CH2) running, in addition to the basic fan loads. From the graphs, we note that the major turn-on events occur between 7:40 and 8:20 and the turn-off events around 17:30 each weekday. No significant load was observed on Sunday and hence its graph is not included in this report. In Figures 35 and 36, the time on the horizontal axis greater than 24 (hr) corresponds to the following day. Because they both show a similar overall behavior, we will focus on the Monday data for the detailed load analysis.

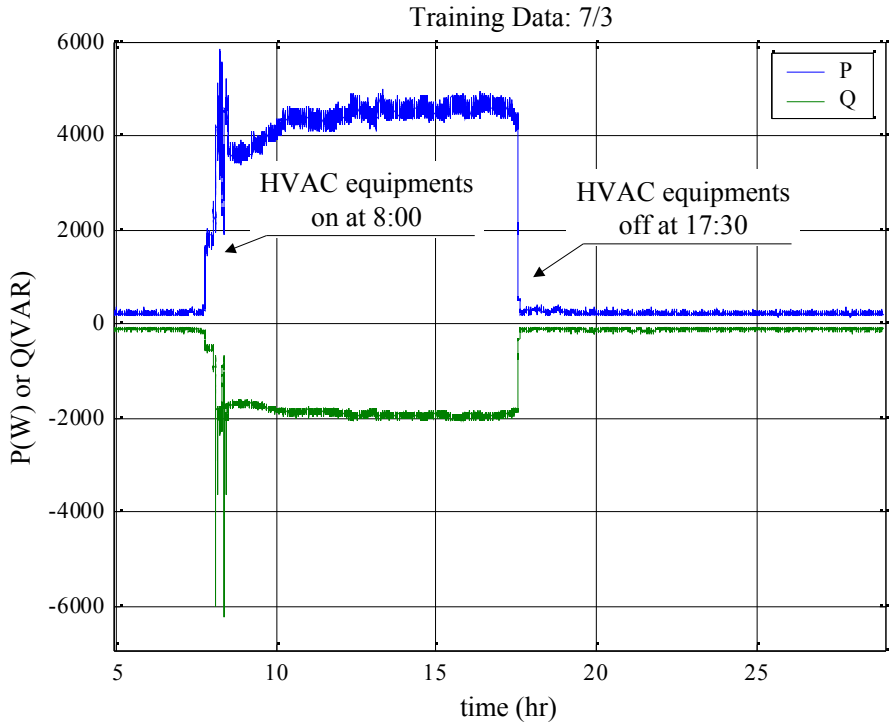


Figure 35. Training data: Tuesday, July 3, 2001

Figure 37 shows the turn-on events of the Monday data. To identify the major events, the submetered data provided by the 160 Sansome building management team was used. Figure 39 and 40 show the submetered data on that day. In the near future, more training data with a higher sampling frequency (120 Hz) will be collected and used to train the Vsection, which is the load transient identification algorithm of the NILM. The current algorithm recognizes a step change in the data stream (P, Q and their harmonics) and calls the identification routine, which computes the proximity to the exemplars (template patterns). Based on the proximity value, the algorithm either decides on a match or rejects the transient. If it is a match, it is posted on the web page with a description tag and the proximity value. Training data are used to produce the exemplars. Chiller training data

need to be recorded, because the chillers were not tested in the initial data recording in March.

The current load identification algorithm works well for the instant loads such as the incandescent bulbs or the rapid start motors. But it is weak at detecting slowly varying transients, such as VFD start up of fans, and a transient with multiple step changes (or spikes). A power supply for a computer typically shows multiple step changes, as each device in a computer is powered up sequentially. A better algorithm needs to be conceived to detect the slow and the multi-step changes. With high quality training data and an upgraded algorithm, the NILM is expected to be able to do load identification autonomously.

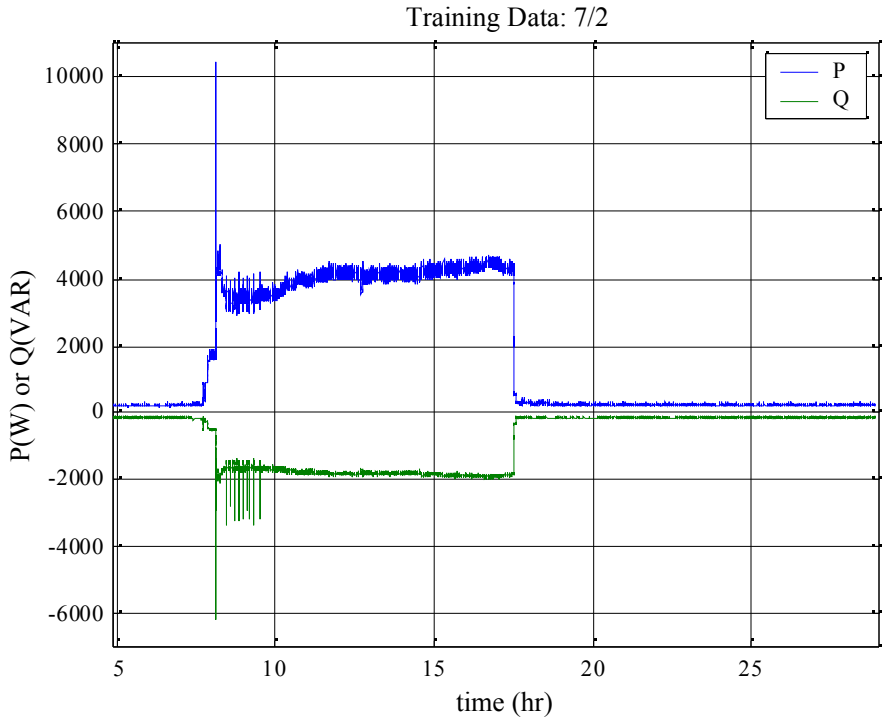


Figure 36. Training data: Monday, July 2, 2001

The first event of the day was the on/off transient of the four pumps in the circulation loop, as evidenced by Figure 39. Each chiller has one pump for chilled water (CHWP) and one for condenser water (CWP). On that day, only the chiller 2 (CH2) was used. However, every pump was turned on momentarily for two minutes and turned off. Twenty minutes later, four pumps were turned on again, but this time, the pumps for the chiller 1 are turned off and the chiller 2 pumps remained on. It is interesting to question why all the pumps are turned on and off before the steady turn-on event and why the chiller 1 pumps are also turned when only the chiller 2 is used. These pump turn on/off events are clearly visible on Figure 37.

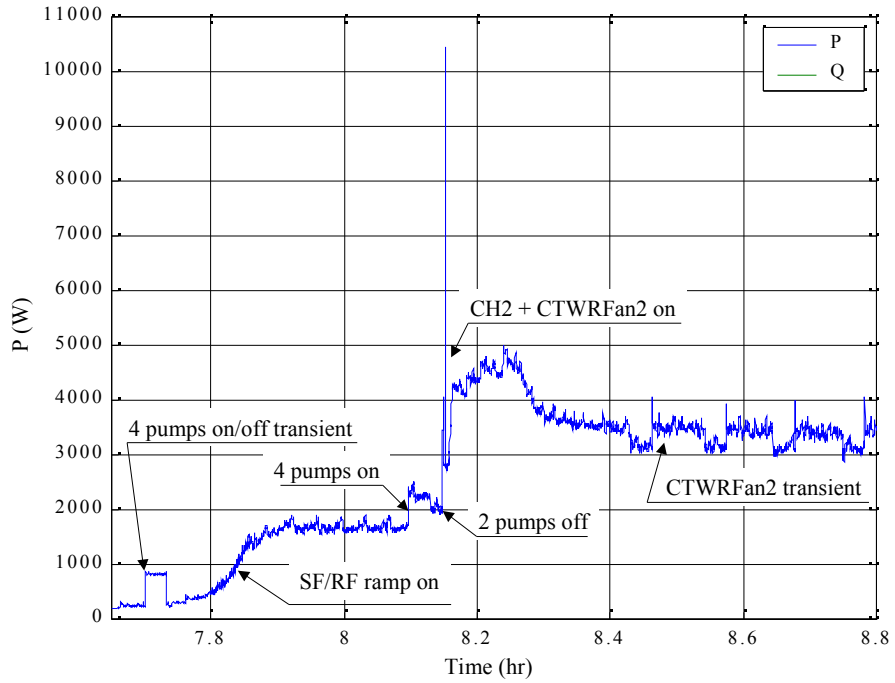


Figure 37. Turn-on events of Monday, July 2, 2001

Shortly after the four pumps' on/off transient, the supply and the return fan (SF/RF) started their accelerations at 7:45. The SF and RF pair is equipped with variable frequency drives and their speeds are slowly ramped up over the eight minute period. A modification in the Vsection algorithm will be required to identify a turn-on event over an extended time period, because the current algorithm is based on detecting a step change in load over a very short time period.

After the pumps for the chiller 2 circulation loop are turned on, chiller 2 and the accompanying cooling tower fan 2 (CTWRFan2) are turned on. It is hard to distinguish the two events when they overlap in time scale. However, if they always happen in a pair with a repeatable transient load pattern, they can be labeled as a single event and will be able to be detected by the Vsection.

It is interesting to note the continued on/off behavior of the cooling tower fan 2 even after the chiller 2 is turned on. The last graph of Figure 40 shows this phenomenon. It is also caught in the NILM training data in Figure 37. CTWRFan2 is first turned on at 8:11, but it experiences the repeated on-off behavior with about a six-minute interval until 9:32. After then, it is stabilized to remain on. It is assumed that a faulty logic in the control loop or a malfunction in the fan unit caused the abnormal behavior, because this type of extended transient behavior is not visible on the July 3 data. The NILM will be able to detect these faults and report to the building HVAC control system to take appropriate actions, with an improved fault detection algorithm and the better communication environment.

The building management team also recognized the problem from the submetered data and probed the cause. The cooling tower fan is controlled by an algorithm that turns on the fan if the supply water temperature exceeds 75 °F and turns it off at 72 °F. Thus it is possible for a cooling tower fan to show oscillatory on/off behavior if the cooling tower load (hence chiller load) is relatively small. The management team is considering a conversion to a VFD drive to avoid the problem. The faulty behavior is clearly visible on the NILM data. Thus with further research, the NILM will be able to do load identifications and fault diagnostics in an environment where no submetered data are available.

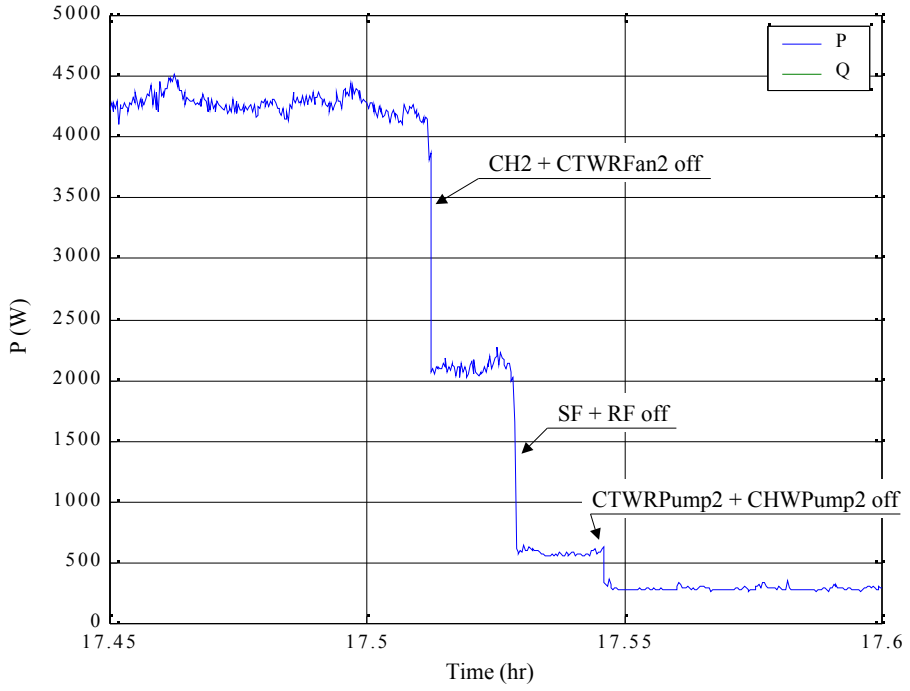


Figure 38. Turn-off events of Monday, July 2, 2001

Turn-off events are usually simpler than turn-on events. They often involve step changes from their steady state load. Figure 38 shows the turn-off events. The events are centered on 17:30 and each event is separated by about a minute. As the turn-on events, the turn-off events occur in pairs. The chiller 2 and the cooling tower fan 2 are turned off first. The supply and the return fan are turned off a minute later. Finally the two pumps for the chiller 2 are turned off. It will be straightforward to detect the off transient, because the Vsection algorithm is good at detecting step transients.

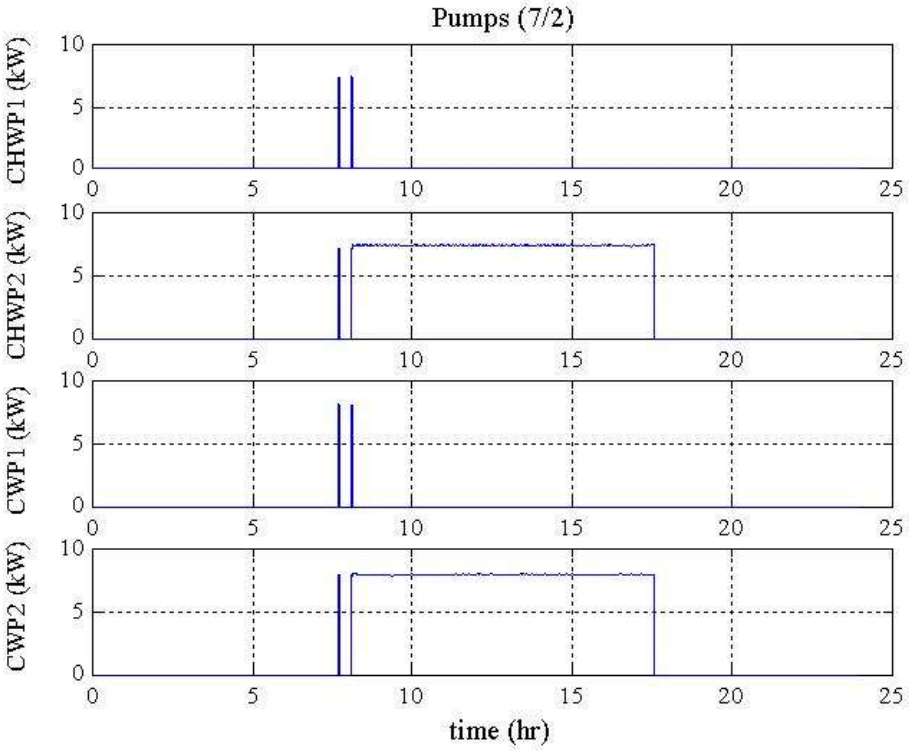


Figure 39. Submetered pump data of Monday, July 2, 2001

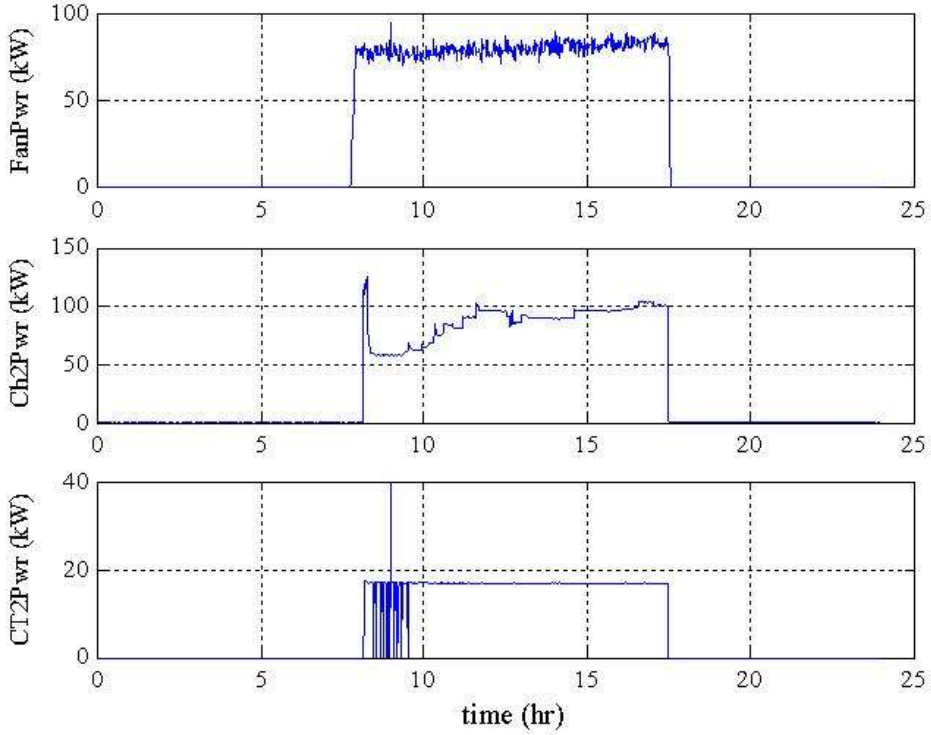


Figure 40. Submetered data of Monday, July 2, 2001

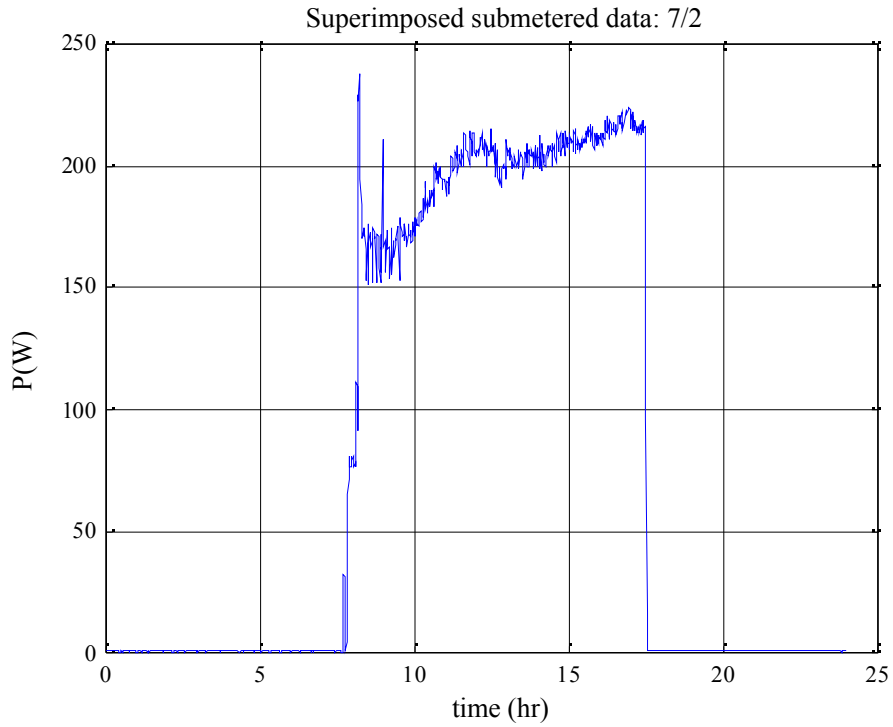


Figure 41. Superimposed submetered data of Monday, July 2, 2001

It is interesting to see if by superimposing the submetered data, one can construct a pattern similar to the NILM data. If so, it also means that the NILM data can be used to dissect the individual load components. Figure 41 shows the plot of the superimposed submetered data. The plot is simply the sum of the individual components in Figure 39 and 40. The overall pattern is strikingly similar to Figure 36, which is the NILM plot of the same day. The glitch in the middle of the CTWRFan2 transient seems to be noise in the submetered data. Subtle differences exist because the sampling frequencies are different (1 Hz for the NILM and 1/60 Hz for the submetered). Also the P values in Figure 36 need to be multiplied by 50 to get the correct scale.

Summary

A NILM system, consisting of CTs, a PT, a signal conditioning box and a host computer, has been installed on the Hong Kong bank building located at 160 Sansome Street, San Francisco. The installation has been successful and the real time plots of P and Q and their harmonics can be seen at http://66.120.20.99 via a web browser.

Training data for each HVAC component, except chillers, were collected to identify load shapes and to prepare the exemplars for load identification. The data shows each device has its own characteristic load shape, for example a cooling tower fan transient is different from a condenser water pump one. The data and the subsequent frequency analysis revealed the (mechanical) shaft harmonics of the supply and the return fan.

One day of training data were collected and compared with the submetered data. The NILM data recorded major on/off events and agrees well with the submetered data. The NILM data can be used to identify the loads. Further refinement in the load detection algorithm is required to detect slowly varying or multi-step change loads. A faulty behavior of the cooling tower fan 2 was observed both from the NILM data and the submetered data. Combined with the load detection algorithm, the NILM can be used for fault detection.

Future Work

There are two main thrusts in the future research: upgrading and improving the NILM system and the optimization of the HVAC equipment operation.

1. NILM system upgrade

    - The current load identification algorithms needs further refinements to recognize slow or multi-step transients. Downsampling (using slow sampling/detection frequency) can be a solution to identify the slow transients. Spacing between the steps may be used to identify the multi-step transients.
    - Frequency component analysis (Fast Fourier Transform, etc.) can be included in the NILM software to identify harmonics and faulty vibrations. Combined with the load identification algorithm, it can be used to constitute a fault detection and diagnostics sub system.
    - Other upgrades, such as providing a communication channel to the main building control system or improving web pages and user interface, will also be necessary.

2. HVAC operation optimization

To achieve the goal of energy saving, it is required to establish the correct control logic of the HVAC equipment and to optimize their operation. Ideally, the control logic and the operation strategy should be based on the correct physical model and the current status of the plant. However, in a real environment, physical models are not perfect and all the environment and process variables required to measure the status of the plant are not available. Thus it is plausible to come up with a simple operational strategy to realize the energy saving based on the measurement of a few simple and easy-to-measure variables. The strategy may have an adaptive and evolving nature to adjust itself and make decisions in real time. The NILM system can assist the decision making by providing the power data, the load identification and the fault detection.

In the immediate future (for the next three to six months), the emphasis will be on collecting more training data at the 160 Sansome building (and other sites), including chillers, to test the robustness of the current load identification algorithm. New algorithms will be tested off line first and, if possible, will be tested on line too. The operations of the chillers and the fan units will be studied to come up with a simple and easy-to-implement operation strategy, which can significantly contribute to energy savings.

CEC Comments on the first submittal of the report

This is an interesting report on a very promising new technology in building automation and diagnostic systems. This CEC reviewer would like to pose the following questions:

1) How do you envision the market development of this product taking place? While there is no end to the intelligence that can be built into the software, the device should aim to satisfy the needs of a target market on its way to becoming a self-sustaining product, and then build on that market niche. This market might be service-type professionals who could be expected to learn how to install and train the NILM, but they would probably not be post-doctoral electrical engineers. So the question is, how much built-in intelligence is enough to get the device into the market and make it a success? How much training will be required of the user?

2) Can the device deliver status information and diagnostics to a facility management system (FMS)? Would this interchange take place by DDE or are other information exchange formats more suitable? Could the software run concurrently with the FMS software on the same machine or would a separate and dedicated computer be required?

3) If NILM is connected to only one phase will it be able to detect single phasing conditions in motors and other equipment? If nominally identical equipment is in close physical proximity, will NILM be able to determine which of them is running? If two motors happen to come on at essentially the same time, but are essentially independent of each other, for example two pieces of package equipment that operate on their own independent controls, does that create an error problem? Are there ways to deal with this? Can NILM determine the contribution of individual equipment to the overall building power factor, or general building power quality?

4) Is it going to be low enough cost that it would be left in most facilities more-or-less forever, or would a temporary installation make sense?

MIT Reply

The MIT team thanks the CEC reviewer for insightful comments and responds as follows:

1. Market for the NILM

MIT agrees that is important to have a vision for market development, even as the tests in California buildings will help shape that market by defining the NILM's performance. MIT finds it useful to consider four options, which are the possible

combinations of centralized and local power monitoring and applications that focus on detection and diagnosis of operating faults and optimization of controls and, separately, estimation of equipment energy use. Please see the following table.

Application	Meter location	
	Central	**Local**
Fault detection and diagnosis and control optimization	Feasible in large buildings with a central HVAC plant, if the NILM is monitoring HVAC equipment only	FDD to be explored at the level of individual packaged units
Estimation of equipment energy consumption and detection of high-level faults (equipment running when not needed)	Feasible if the NILM is monitoring HVAC equipment only	Conventional power meters could be made more intelligent with limited use of NILM technology

MIT considers that fault detection and diagnosis and control optimization from a NILM installed on a motor-control center in a large building to be the most successful application at this time and the market niche most likely to appeal to a potential commercializer. While much remains to be done, work in the Iowa test building and at 160 Sansome St. show that the NILM is capable of producing useful information. The potential savings associated with better operating a large plant are larger than in smaller buildings and would yield the most benefit in return for the cost of the NILM. Also, there are personnel who operate or service such buildings; such personnel should as a rule have the training and interest to use the information provided by the NILM. It appears reasonable at this time to initially target this market.

Fault detection in packaged rather than built-up HVAC equipment is an area under exploration by others. Tabulation of survey data concerning the nature of faults (see Breuker, Rossi and Braun's article in the November 2000 ASHRAE Journal) shows that several classes of faults could potentially be detected by electrical monitoring. Compressor faults account for the largest percentage of total service costs and many of these faults are electrical in nature: short to ground, open windings, and locked rotor. MIT intends to explore this area with the NILM, using a packaged unit at a school that is being monitored as part of the benchmarking project under LBNL. A market for electrical monitoring of packaged units would likely require monitors dedicated to each unit or at most a few units, depending on electrical wiring. The monitor would need to be inexpensive, given the size of the equipment. MIT will try to assess whether electrical monitoring produces useful results before directing attention to the packaging of monitor hardware and software.

Estimation of equipment energy use is an important area and one that deserves a careful assessment. To date, MIT has had less experience with this area than with detection of faults. However, MIT's experience in a fast-food restaurant shows that a NILM attached to the HVAC electrical service panel could detect nearly all on-off events (over 97%) and could successfully classify about 91% of the detected events,

with no additional information. However, the small number of events misclassified (primarily due to the fact that several devices have similar real and reactive power) led to large errors in energy estimation. If, for example, a kitchen exhaust fan is successfully detected as having turned on but the turn-off event is attributed to a fan in a rooftop unit, then the run-time and energy consumption of the exhaust fan will be incorrectly estimated. If this sort of problem can be eliminated, via the use of more information (analysis of start-up transients and of frequency content of power, which is important in tracking belt-driven equipment), then there could be a market for the NILM as a power monitor. This market would likely develop first in larger buildings, such as the LA County Children's Court, where operating personnel have expressed an interest in first-order energy estimation as a means of assessing equipment performance under normal operating conditions and when load curtailment is required.

Finally, there may be cases where one meter is not sufficient. If a motor-control center serves a large number of fans and pumps and some of those loads are small, one meter may not detect all of them. It may be practical, in the future, to embed some of the NILM software in a dedicated submeter, at a cost lower than the current NILM platform. For a fan, for example, the meter could then provide power usage and energy consumption (the main role of a submeter) and also some amount of fault detection (slipping fan belt, damper problems).

2. Interaction with an FMS

Yes, the NILM could interact with an FMS. One controls company has been interested in the NILM under just such an arrangement. The exchange of information could take place in a variety of ways and in principle the NILM software could run on the same computer used by the FMS. At the moment, MIT is using the NILM platform to access data collected by a separate logger with associated current transducers (that provide information at the component level needed to assess the performance of the NILM. Instead of using a separate telephone line for the logger, MIT has slaved the logger to the NILM.

The NILM currently uses a very inexpensive and slow PC as a hardware platform. Refurbished computers that are of little use as office equipment because they are too slow are very suitable. The latest in performance is not needed. All of this argues in favor of a shared platform. However, the NILM software is written to run under Linux and not Windows, for a number of reasons related to data handling and maximizing the performance of inexpensive hardware. Retaining the use of Linux may require a separate computer. Alternatively, a commercializer could rewrite the code to run under Windows.

3. Limitations on detection and classification of loads

Nominally identical equipment in close physical proximity cannot be distinguished in the absence of other information: control signals, thermo-fluid measurements, or some sort of identification signal transmitted over the power lines, as CEC suggested in its comments on report 2.2.1b. It may be possible to distinguish nominally identical equipment in separate parts of a building, due to the influence of the building electrical distribution system on the signal as it is received at a centrally located NILM. In many built-up HVAC plants, equipment identification should be a relatively minor problem, because all three options listed above are reasonable approaches to obtaining additional information For buildings with distributed packaged equipment, the third option may be most viable.

Simultaneous start-up is a problem on two time scales. If the NILM uses software that detects step changes in loads, enough time is required to distinguish two steps. This is on the order of 10-30 seconds. If the NILM uses software, now under evaluation in 160 Sansome St., that analyzes start-up patterns that typically occur over 1-2 seconds, events can be separated by that amount of time and still be distinguished.

The NILM obtains samples at a rate sufficient to assess power quality and can relate changes in power quality to changes in equipment operation (i.e., what's running at a particular time). Other problems can be detected with the same approach. For example, low-voltage cut-out of industrial equipment due to voltage sags caused by start-up of other equipment can be diagnosed.

4. Cost and equipment installation

While it is conceivable that the NILM could be installed temporarily, MIT envisions it as a permanently installed device. Permanent installation is needed to continuously monitor equipment performance. Permanent installation would exclude the NILM from very small buildings but may make sense for larger buildings. The hardware cost is modest – on the order of $1000 for the computer, A/D card, current transducers, voltage tap and cabling. The cost would be reduced if a computer already installed as part of an EMS could be used.

Economics should be assessed as the benefits to be obtained from such a capital investment and, more significantly, the cost to pay attention to the data. The NILM has been tested in a test building in Iowa, where HVAC faults were artificially introduced. While the NILM found these faults, the test did not include an assessment of how often certain faults occur naturally. In the 160 Sansome St. building, the NILM has found several potential problems. For a building of that size, which the management company is committed to understanding and operating

efficiently, the NILM is attractive as an alternative to more-expensive submetering. In a fast-food restaurant MIT monitored with a NILM, the goal was to produce low-cost information that could be quickly assessed by energy-service personnel at a remote location. For such a building, the alternative to the NILM is likely to be no monitoring of electricity usage. MIT has yet to demonstrate that there is sufficient benefit in this class of buildings, as quantified by detection of inefficient equipment operation or equipment faults. Field tests of sufficient duration to detect naturally occurring faults in any class of buildings are required. Tests in California buildings will be useful in this regard.

Project 2.2

Equipment Scheduling and Cycling

Task 2.2.3

Automated Diagnostics Based on Start-Up Signals

Deliverable 2.2.3(a)

Development of a Functioning Centrally Located Electrical-Load Monitor

Submitted to Architectural Energy Corporation

Under the

California Energy Commission's

Public Interest Energy Research (PIER) Program

Energy-Efficient and Affordable

Small Commercial and Residential Buildings

California Energy Commission Contract 400-99-011

K. D. Lee, L. K. Norford and S. B. Leeb

Massachusetts Institute of Technology

May 27, 2003

Development of a Functioning Centrally Located Electrical-Load Monitor

Overview of project

This report documents the development of a functioning centrally located electrical-load monitor. The non-intrusive load monitor (NILM) is a personal-computer based instrument that samples voltage and current at a single location and estimates the electrical energy drawn by individual pieces of equipment. NILM results can also be used for detecting and diagnosing faults.

A Ph.D. dissertation by Kwangduk Douglas Lee constitutes this report. The thesis, titled "Electric Load Information System based on Non-Intrusive Power Monitoring," is included in its entirety. Mr. Lee acknowledges the contributions of Prof. Steven Shaw at Montana State, Fred Smothers and Ted Ludwick, both formerly associated with the test-site office building, and Peter Armstrong, Robert Cox and Chris Laughman at MIT.

The PC-based NILM, running the Linux operating system, uses a data-acquisition card to sample voltage and current at high speed, about 16 kHz. It then Fourier transforms the data to obtain real and reactive power at the fundamental and higher (up to seventh) harmonics. These data are stored at 120 Hz for subsequent analysis. At the beginning of this project, the on-line NILM was capable of detecting loads via analysis of the electrical-power transients observed during start-up of equipment. All other analysis, notably detection of loads via changes in steady-state power and detection of faults, was done off-line, via a combination of algorithms and intervention of an analyst.

Project 2.2, "Equipment Scheduling and Cycling," has focused on developing a more fully automated NILM. The project has had three, yearly phases. The first phase was based on detecting loads from step changes and made use of data collected at the Iowa Energy Center's Energy Research Station. The second phase focused on fully automated detection of equipment operation. This phase developed the algorithms documented in Deliverable 2.2.2c, "Intermediate Report on Automated Diagnostics." The delivered report included examples of naturally occurring faults that the NILM detected, and that would be very difficult to detect with more traditional thermo-fluid fault-detection methods. However, detection of faults must build on automated detection of normal equipment operation. In particular, power oscillations associated

with such faults as a chiller vacuum leak, a misaligned fan shaft, and under-damped feedback-control loops, are detected with the same frequency-analysis techniques used to track variable-speed drives. Accordingly, the work in the third year, initiated intended to focus on automatic detection of faults, instead concentrated on further developing load-tracking algorithms and coding them for on-line use in the NILM. This report will be supplemented by a CD with NILM code. The remainder of this summary highlights the major advances documented in this report and identifies future work.

Use of both steady-state load detection and analysis of start-up transients

This research has developed a load disaggregator that analyzes both step changes in steady-state loads and start-up transients and automatically arbitrates the results in cases where the two methods disagree. The step-change analyzer is significantly more powerful than that used in previous versions of the NILM but is still subject to background electrical noise. The transient detector is relatively immune to white noise but can be fooled by colored noise and is not useful for abrupt shut-down signals. If single-event load detection fails, the signal is sent to a multi-event detector that searches a binary tree of possible combinations of temporally overlapping loads. The step-change analyzer, the multi-event detector and the state estimator, discussed next, make use of a consistent signal-analysis framework that considers load signals to be stochastic events, each associated with a mean and variance.

State estimation

Event detection is susceptible to occasional errors: turn-on and turn-off events will be missed entirely or mislabeled. Such errors can accumulate and reduce the value of the NILM. Further, the NILM needs a starting point when it is first started, so that the status of individual loads can be determined before turn-on or turn-off events are detected. To address these issues, a state estimator was developed to determine the most likely set of loads associated with a given steady power signal.

Estimation of power drawn by variable-speed drive (VSD) loads

Since its inception, the NILM has successfully detected only constant-power loads. VSDs are in common use and have posed significant challenges. Even if the power drawn by a VSD fan were nearly constant over the day, or could be approximated as such with acceptable loss of accuracy, detection of the start-up of a VSD is nearly impossible with either step-change or start-up transient algorithms, due to the very long (order of 10 minutes) start-up period. This report documents a fully functional approach to tracking VSDs that is based on observing the higher harmonics, particularly the fifth and seventh harmonics of real and reactive power that are associated with VSDs. This method cannot separate individual VSD devices but can track their aggregated power use.

Testing of software

NILM algorithms were developed in Matlab and tested off-line with NILM-collected data, primarily from the test office building in San Francisco. The algorithms were converted to C++,

compiled, and embedded in the on-line NILM in the office building. The code was subjected to limited testing, in two phases. First, the disaggregator for constant-power loads was tested by injecting off-line data into the NILM. This procedure was necessary at the time to isolate constant-power loads. The code successfully identified loads and maintained a log of their on-off status, presenting this information via a Web-based interface. Second, the VSD tracking software was loaded and tested. It successfully separated the supply- and return-fan VSD power from the HVAC total.

Future work

Several important steps need to be taken:

- Tracking of variable-power, constant-speed loads. This last remaining load class includes most chillers and those fans still using inlet vanes. Relations between real and reactive power for these loads will be explored. Directly tracking a chiller is preferable to assigning to the chiller that portion of HVAC power not due to constant-power and VSD loads, because chiller power "by subtraction" is subject to errors in following other loads. However, the subtraction approach is possible if the NILM is mounted at the HVAC service entrance.

- Integration of the disaggregator for constant-power loads and the VSD tracker. The code was designed to be integrated. However, most constant-power loads are for pumps and tower fans associated with operation of a chiller. Until the chiller can be detected or is assigned remaining HVAC power, the integration has reduced practical value.

- Testing of the NILM and determination of its commercial value. This work, the most important, requires that NILM load-tracking output be compared with submeters and that a potential commercializer determine whether NILM output, either load tracking or power signatures associated with faults, has value to its customers.

- Automated training of the NILM. The state estimator is a good start here, because it makes a best guess of the loads in operation at the time the NILM is turned on. However, the state estimator and the load disaggregator rely on a knowledge of the characteristics of individual loads, including steady-state real and reactive power and start-up transients. To date, this information has been gathered manually, by turning on and off individual loads several times and analyzing the power data. For commercial applications, the training must be automated to the extent possible and clear instructions must be prepared for remaining, manual steps.

- Automated detection of faults. As a starting point, detection of faults associated with oscillatory power signals should be automated, based on frequency analysis. Automatic fault detection will aid those potential users who have limited time or ability to analyze NILM data.

Table of Contents

Chapter 1

Introduction

1.1 Motivation and background of research

Electric power has been with us for more than a century and surrounds us in the form of AC wall outlets, power supplies for mobile phones and laptops, power sources for the onboard devices of automobiles, etc. It affects every aspect of our daily lives and our daily activities depend on its implicit presence. It is undoubtedly one of the core foundations of the modern industrial and information society. However, this once well-established technology recently faces three major challenges: *deregulation*, *autonomous power systems* and *power quality*.

The purpose of the deregulation of the electric utility industry is to introduce competition to the traditional monopoly [Wollenberg, 1998; Philipson and Willis, 1998; Zaccour, 1998]. Before deregulation, the electric utility was vertically integrated, performing generation, transmission, distribution and retail under one umbrella. It was given exclusive sales rights in specific regions, in return for the acceptance of regulatory supervision and the guarantee of reliable service. After deregulation, those functional units become independent entities with more horizontal relationships and participate in open markets [Lee and Lin, 1998]. The original intent of deregulation was to increase efficiency, lower price and improve, or at least retain, the quality of service [Barkovich and Hawk, 1996]. The deregulation of electric utilities is an inevitable trend, as it was the case in the telecommunication and gas industry, but it can generate many adverse effects along its course [Draper, 1998; Abdoo, 1999; Ilic *et al*., 2001]. The most notable example is the recent California electricity crisis [Budhraja, 2002; Crow, 2002; McNamara, 2002; Sweeney, 2002].

Electricity is traded as a commodity in deregulated markets. Generators sell their capacity on a wholesale market. Distribution companies bid in a spot market for best prices and resell the electricity to their local customers. In an open market, the maximum utilization of existing resources and the acquisition and communication of information become critical for the survival and profitability of the participants and for the prosperity of the market itself [Harris, 2000]. To this end, concerned parties have invested in and developed various real-time transaction, trade and monitoring systems, fueled by advances in computer and communication technologies [Gill *et al*., 2001; McClanahan, 2002]. However, one party is particularly unprepared for these changes– the customers, who are on top of the electricity food chain and the alleged beneficiary of all those reforms [Ilic *et al*., 2001].

Customers need to make educated decisions in their energy purchases, if they have an option to choose among different vendors and if the electricity price changes in real time. The selection of a vendor will depend on the overall quality of service, including price, reliability and customer relationship. After a service provider is selected, the customer

needs a monitoring tool to ensure that the desired level of service is being provided, and to verify whether its energy consumption is optimal. Real-time pricing makes the situation more complicated, because the customer must know how much power is consumed and what loads are being used, in real time, to make prudent purchase decisions. Utilities are moving from monthly metering to interval metering, with metering intervals as short as fifteen minutes [McClanahan, 2002]. It is rather unclear whether the utilities will share metered data in real time, along with current prices and price forecasts, with customers. Even if metered data are available to end users, they will be the aggregated sums of powers consumed by individual loads with no specific detail about the nature of power consumption.

The manager of a commercial or production facility may opt to know the state of electric loads under his or her supervision to make the best energy purchase decision subject to market forces. However, this information is typically unavailable unless a manual inspection is performed or a separate instrument is installed at every single load. Both options are cost prohibitive and difficult to implement. In a deregulated market, customers need a *ubiquitous electric meter*, which can provide them with itemized information in real time to assist their energy purchase decisions.

Various autonomous power systems are emerging on the horizon, gaining more and more significance. They refer to electric systems, independent or semi-independent from the utility grid. Self-sufficient power systems on-board automobiles or ships are good examples. Isolated power systems based on alternative or renewable sources, such as fuel cells and wind energy, are more widely used [Wolk, 1999; Cook, 2002; Walker, 1998; Bull, 2001]. They are sometimes connected to the grid (utility-interactive) as a form of distributed or dispersed generation [Naik *et al*., 1995; Price *et al*., 1999].

They pose difficult challenges such as how to generate and distribute power efficiently within a self-contained system, and how to manage interactions if it is connected to a grid. For example, the generation and distribution of electricity on automobiles becomes more critical, as the auto manufacturers install more power-intensive components for better fuel economy, safety and comfort [Kassakian *et al*., 1996; Kassakian, 2000; Emadi *et al*., 1999]. Upgrade to a 42 V DC system is one of the answers to this inevitable trend [Neubert, 2000]. AC distribution systems are also considered for automobiles [Miller, 1996; Khan, 1999]. Another example is how to connect a fuel cell generator to the utility grid. It should able to be configured either to send power to or receive power from the grid, while serving its local load [Farooque and Maru, 2001]. Those challenges can be met only by the accurate knowledge of the state of its own system, and the network when connected.

The issue of power quality is getting more attention, partly because of the consequences of the first two challenges and partly because of the increased use of sensitive and nonlinear electric loads [Kennedy, 2000; Stones and Collinson, 2001]. The deregulation of the electric utility seems to have degraded, at least temporarily, the reliability and quality of power, leading to several blackouts and brownouts and numerous disruptions and disturbances [Kendall, 2001; Arrillaga et al., 2000]. Autonomous power systems

often experience wiring or ground faults, such as a short circuit between 14 V and 42 V rails on automobiles [Neubert, 2000], and can suffer severely if they are mission-critical systems, e.g. modern naval combat vessels and spacecrafts packed with computers and electronics.

The increased use of semiconductor-based power electronic devices creates severe harmonic distortions in the current and voltage waveforms. Voltage distortions can lead to the malfunctions of equipment and protective circuit elements. The widespread use of information processing equipment exacerbated the situation, but gave rise to the public awareness of the importance of the power quality. Every concerned party desires a type of affordable power quality monitoring tool, possibly operating in real time, to characterize the distortions and disturbances and take necessary actions. Modern power quality monitoring tools can detect and classify disturbances and are network-capable [Wijayakulasooriya *et al*., 2002; Bucci *et al*., 2003]. End users often use the results to specify power-quality-tolerant equipment [Oliver *et al*., 2002]. There also are attempts to add power quality monitoring capability to the smart meters in the deregulated market [Saari *et al*., 1996; Koponen *et al*., 1996].

In author's viewpoint, the prime motive behind these challenges faced by the electric power systems is the fact that we are now entering a new era of information society. The power system infrastructure and practices were established during the first half of the twentieth century, when the electricity became the basis of the energy-centered industrial society. The second half of the last century saw the electricity expanding its role also as a basis of the new information society. Consequently, by the turn of the century, the existing installations and established technologies were inadequate to meet the requirements of the new era, when the electricity is expected to perform the dual role of the medium of both energy and information.

Ironically, possible solutions of the problems caused by the new needs of the information society will require the knowledge of the electric systems they intend to improve. In other words, the *information* of the power systems will be the key in meeting the challenges posed by the fledgling information society.

However, obtaining high quality information economically and reliably is always a difficult objective to achieve. The Non-Intrusive Power Monitor (NIPM) is a novel attempt to collect such information with a minimal physical installation. Raw voltage and current are measured at a single location to yield harmonic power signals. They typically carry the fingerprints of the electric loads present in a system, and their analysis can produce such information as the operational and diagnostic status of the loads. The power signals can also be used for the system identification, parameter estimation and energy consumption optimization study.

This research was mainly sponsored by the California Energy Commission and inaugurated in the summer of 2000, at the height of the California energy crisis. Its main purpose was to develop an intelligent yet affordable monitoring device so that the consumers can make better choices in a deregulated and volatile electricity market. It

focuses on the development of mathematical models and signal processing algorithms to extract information for the customers at the end of the AC distribution system. Its results will directly benefit the developments of a ubiquitous electric meter in a deregulated market, a diagnostic or prognostic tool for mission-critical systems and an intelligent power quality monitor. With minor modifications, the results will also be useful for other sectors, such as DC electrical systems and electric utilities.

1.2 System overview and example power waveforms

The NIPM is typically installed at the main electric service entry of a system, and thus can monitor the powers consumed by the individual loads in the system. The NIPM samples raw voltage and current waveforms and computes harmonic powers from the samples. The harmonic power signals are analyzed by various detection, estimation and classification algorithms to extract desired information, such as the ON/OFF status of the loads, existence and diagnosis of faults, optimality of energy usage, etc. The information will be displayed to users, so that they can know the system state and make appropriate operational and energy purchase decisions. The information may also be forwarded to an electric load control system to adjust the state of the loads following certain guidelines, such as minimizing energy usage while not compromising comfort or productivity. In this thesis, the Electric Load Information System (ELIS) denotes the overall information extraction and handling system. Figure 1-1 shows the overall schematic of the system.

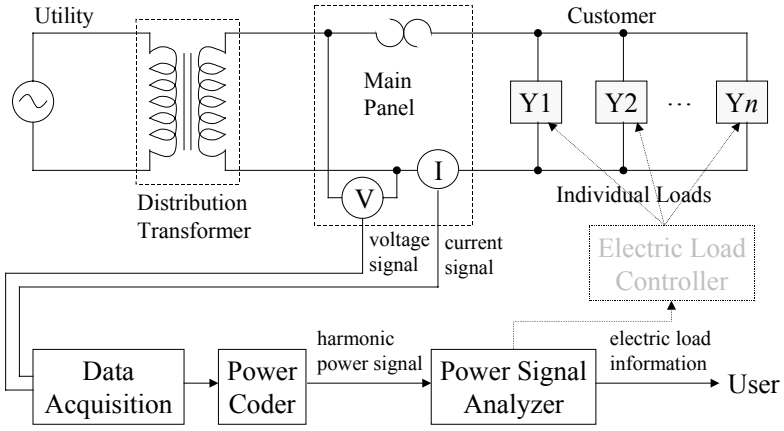


Figure 1-1 Schematic of ELIS based on NIPM.

Figure 1-2 shows a daily power consumption pattern of a commercial building logged by an NIPM. The building has large variable speed drives (VSDs) for its major air-handling unit. The VSDs turn on around 7:15 and turn off at 17:30. These events are clearly visible in the graph. The building also has a large chiller. It turns on slightly after 8:00 and turns off around 17:00. Again, these events are clearly visible in the picture. The chiller is equipped with capacity control. In the morning when the cooling load is minor, its power

consumption is relatively flat. However, after 11:00, as the cooling load increases, the chiller starts to consume more power and its power consumption varies with the varying cooling load. Small spikes in the morning and the late afternoon are turn-on and turn-off events of a single-speed cooling tower fan. Because the fan has no capacity control, it periodically turns on and off when the cooling load is minor. However, once the cooling load exceeds a certain threshold, it operates continuously.

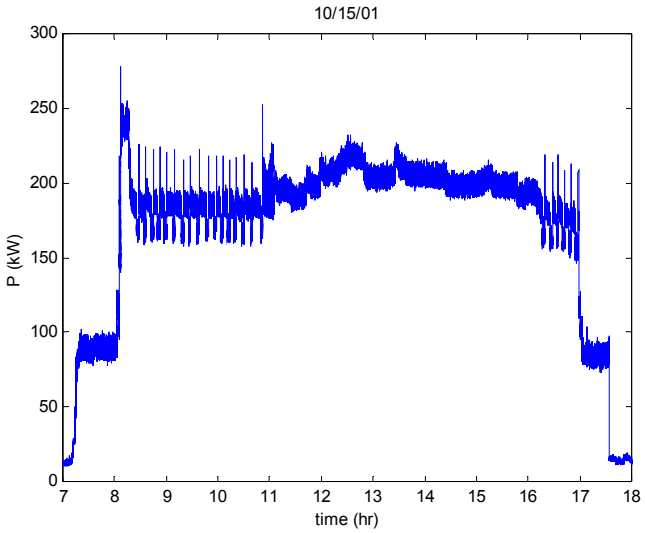


Figure 1-2 Daily power consumption profile of a commercial building.

Figure 1-2 suggests that by analyzing the non-intrusive power waveforms, we can identify load events and thus recognize which loads are on at a given time. The analysis may be performed either *in situ* or offline. The knowledge of load events allows us to infer the behavior of the monitored loads. For example, the knowledge of the frequent on-off events of the cooling tower fan lets us know that the cooling load at that moment is minor. Also, if its on-off frequency is too excessive, we may decide to take an action such as increasing the dead-time zone in its ON/OFF control or add a capacity control.

Figure 1-3 shows another power waveform taken from the same building. On that day, there was a leak in the evaporator chamber of the chiller. As a result, outside air infiltrated the refrigeration cycle. Because the air is lighter and less compressible than the refrigerant, the centrifugal compressor of the chiller sees a sudden decrease in its load when the entrapped air enters its compression chamber, resulting in a rapid decrease in its electric power consumption. The graph shows the onset of this phenomenon and because the air is circulated along with the refrigerant, the power fluctuates almost periodically. It took two hours for the engineer of the building to get to know the situation, by the symptom of lost cooling, and stop the chiller to purge the evaporator manually. But during those two hours, the chiller was subject to an excessive amount of wear and thermal comfort was lost.

This example suggests that certain faults are visible in power waveforms and that the power signals can be used for their detection and possibly for their identification. For

example, a very simple variance estimator may be designed and when the variance estimation exceeds a certain threshold over a prolonged period, the ELIS may decide that there exists a significant fault. Identifying the nature and cause of a fault is much more difficult. It will inevitably involve accumulating the knowledge of the faults and decide the character of a detected fault, using a classification or pattern recognition algorithm.

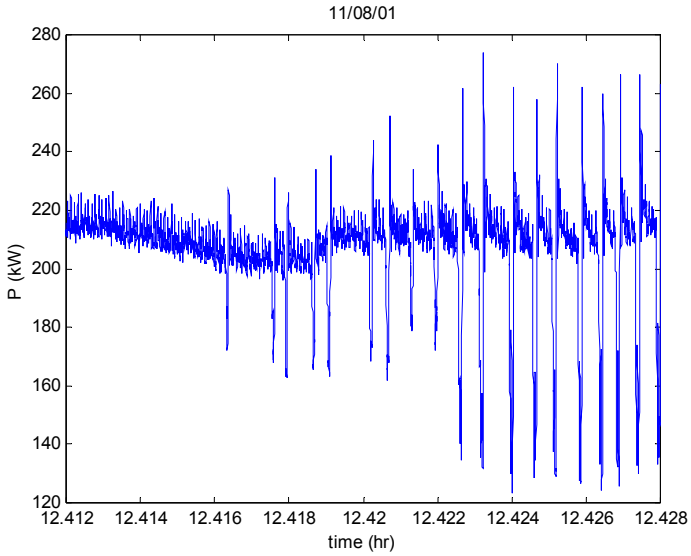


Figure 1-3 Onset of chiller evaporator vacuum leak.

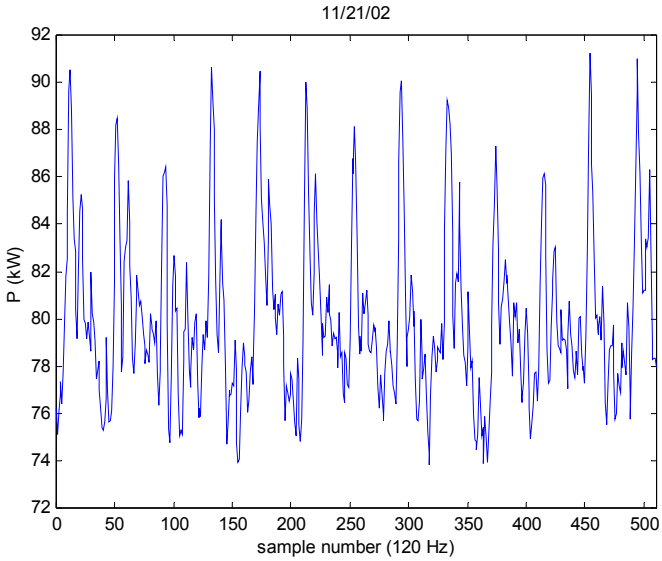


Figure 1-4 Harmonic signal present in power waveform.

Figure 1-4 shows a harmonic signal present in a power waveform. Its periodogram[1] in Figure 1-5 clearly shows the spectral peaks in a harmonic relationship. In power engineering, these harmonic signals are called *interharmonics*, because their frequencies

[1]Periodogram is defined in Chapter 7. It is a natural estimator of the power spectral density (PSD) of signal.

are not the multiples of the fundamental frequency of the voltage waveform, i.e. 60 Hz in the US. Common causes of the interharmonics are ripples in a DC bus between a rectifier and an inverter and periodically changing loads. It usually is necessary to eliminate the interharmonics in analyzing a power signal, because they can interfere with the event detection and classification algorithms, which rely largely on the change of mean detection and the uniqueness of event patterns. On the other hand, an interharmonic may be actively estimated, if it can be associated with the physical status of a load. For example, if the interharmonic is generated by a periodic pulsation of a rotating machine, extracting its parameters, such as amplitude and frequency, can be extremely useful to monitor the machine. The amplitude estimation can be used to quantify the severity of the pulsation, and the frequency estimation can be used to monitor its rotation speed.

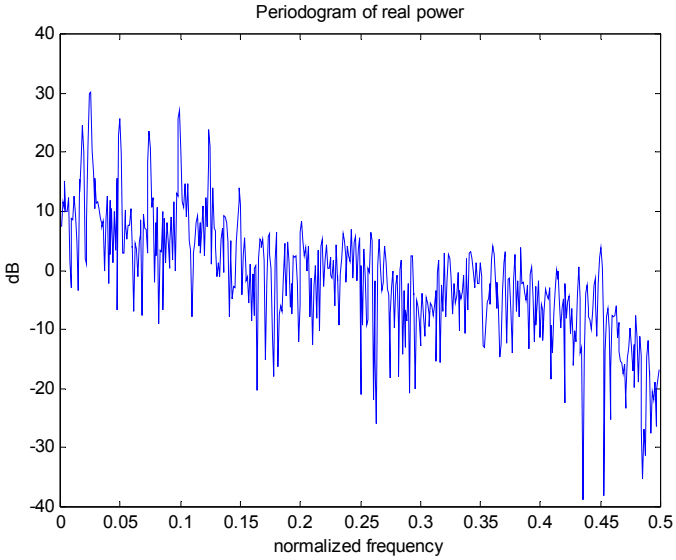


Figure 1-5 Periodogram of harmonic signal.

The example graphs in this section illustrate that a rich amount of information is embedded in non-intrusive power signals and that appropriate analysis techniques will be necessary to extract those information. We certainly want most analyses to be performed automatically by a computer, in real time if possible, although certain higher-level analyses may still have to be performed by human users. Thus, the goal of this thesis is to identify those tasks suitable for computers and to develop various signal processing and information handling algorithms to accomplish them.

1.3 Brief history of non-intrusive power monitoring

The non-intrusive power monitoring was originally called non-intrusive appliance load monitoring (NALM), and later called non-intrusive load monitoring (NILM) to imply that the monitored loads are not necessarily household appliances. The author prefers to call it the NIPM, in the sense that power is the signal we observe, whereas load is the information we have to extract from the observations. It also is in line with the philosophy of this thesis, in which the functionality of lower level non-intrusive power

signal acquisition is separated from that of higher level power signal analysis to extract the electric load information.

The NALM was first pioneered by George Hart in the 1980s [Hart, 1992]. The original goal was to provide electric utilities with a cost-effective monitoring tool, which can collect the energy consumption data of their residential customers with minimal intrusion on their properties. He resolved this problem by monitoring the power at a metering location. However, because the metered power is a lumped sum, the individual power consumed by a load was not readily obtained. He developed a novel solution for this problem based on the fact that a constant load tends to have a unique power consumption level and that its ON and OFF events show up as distinctive step changes in time plots.

He measured the real and reactive powers at a 1 Hz sampling rate and the step changes were recorded and sent to a central processing location via a telephone modem. The archived step change data were plotted on a 2-dimensional (real-reactive) power plane, and classified based on a distance metric from cluster centers. He also developed a load event sequence analysis program based on the Viterbi algorithm to extract the time history of electric loads from the sequence of classified load events. The Electric Power Research Institute (EPRI) developed a prototype residential load monitoring system based on his work [Carmichael, 1990; Drenker and Kader, 1999]. Other residential power monitors are also reported in the literature [Sultanem, 1991; Cole and Albicki, 1998].

However, his algorithm is limited in a sense that it can detect only the isolated ON/OFF events of constant loads. For example, if two separate events happen at a same time, his classifier either fails or misclassifies, making subsequent load event sequence analyses more complicated. Also, some loads have multiple power consumption levels, e.g. a heater with a LO and HI switch, and those levels may follow a certain sequence, e.g. a dishwasher or washing machine. Although he proposed a finite state machine (FSM) model to handle those multi-stage loads, he did not devise any actual classification or sequence analysis algorithm. Another shortcoming is the fact there is a class of loads that do not have fixed power consumption levels- *variable loads*. A light dimmer is an exemplary residential variable load. This issue becomes more prevalent as one tries to monitor commercial or industrial loads, where significant portions are variable loads, such as variable speed drives. His approach also lacks real time analysis capability.

Steven Leeb approached this problem from a slightly different angle, based on his observation that electric loads tend to have unique patterns when they are turned on [Leeb, 1993]. To observe the transient patterns, it was necessary to reduce the time scale. He developed an analog spectral envelope processor, to estimate the real, reactive and harmonic powers at the fundamental frequency (e.g. 60 Hz) of the voltage waveform. He also developed a rudimentary transient detection and pattern classification algorithm and proposed to examine power waveforms at multiple time resolutions to search for desired patterns [Leeb *et al*., July 1995]. He and Leslie Norford investigated the feasibility of using the NILM for commercial buildings [Norford and Leeb, 1996].

Steven Shaw first developed a digital spectral envelope processor by simulating the functionality of the analog processor in the digital domain [Shaw *et al*., 1998]. He later developed an improved processor based on the discrete Fourier transform (DFT) of waveform samples [Shaw, 2000]. Because the spectral envelope estimation was performed on a computer, it also provided an opportunity to do the analysis locally in real time. He developed a transient pattern matching algorithm based on the method of least squares. However, he did not devise a load status analysis routine, which can be linked to his transient event classifier. He also performed motor parameter estimation and diagnosis based on raw current waveform measurements [Shaw and Leeb, 1999].

Dong Luo modified George Hart's algorithm to analyze the building heating, ventilation and air conditioning (HVAC) loads [D. Luo, 2001]. He was able to extract most of their usage history from the archived power measurements. He also used the power signal to detect and diagnose certain faults present in the HVAC systems [Shaw *et al*., 2002].

For the last twenty years, there have been many advances in monitoring electric loads nonintrusively to extract their status information. However, many problems remain unsolved, such as the event overlap, multi-stage load state estimation, variable load tracking, real-time local processing, etc. Also, no practical system has been offered in the market, whether it is residential, commercial or industrial. This thesis is an attempt to address these longstanding issues and lay a foundation for developing a commercial-grade electric load information system based on the non-intrusive power monitoring.

1.4 Objectives and outline of thesis

Traditionally, electricity has been considered as a medium of energy. It is now considered also as a medium of information with the widespread infiltration of computers and communication networks to modern civilization. However, the electric utility network is still used and mostly considered as a medium of energy transmission, although there is an attempt to use the power line for communication [Pavlidou *et al*., 2003; Lin *et al*., 2002; Chen and Chiueh, 2002].

In this thesis, we are interested in extracting information from the power flows. As a medium of energy, electricity interacts with the loads, and during the process, it inadvertently carries information about their nature. We view the electricity as a medium of energy plus embedded information about the characteristics of the loads. The concept of the ELIS based on the NIPM is to extract high quality and quantity information embedded in the electricity with a minimum physical installation and intrusion.

The examples and discussions in the previous sections suggest that there basically exist three types of information that can be offered by the NIPM: *load disaggregation*, *fault diagnosis* and *system analysis*. Load disaggregation is essentially the study of how to divide the total power observed into the powers consumed by individual electric loads. It requires the identification of load events to keep track of the ON/OFF status of each load, so that the total power can be allocated to each load accordingly at a given instant or

interval. Fault diagnosis is to detect the presence of a fault, identify its nature and possibly probe its cause. It can be a simple summary of an observation or a more complete description possibly with a remedy. System analysis is to use the power signal to identify the models of the electric loads and estimate their parameters. The result of the system analysis can be linked to a proper electrical, mechanical or thermal model of the system, for energy usage optimization studies, supervisions of electric load control loops, power management, etc. It will also be useful in designing load disaggregation and fault diagnosis algorithms, by providing system models and parameters.

The common element of these three types of information is the state of the system. The exact knowledge of the status of the electric loads in a system makes the task of load disaggregation almost trivial (when supported by the knowledge of their nominal behavior). Any event or gradual degradation that can not be supported or explained by the legitimate combination of load states can be considered as a fault. Also, detecting and identifying a fault is basically similar to load event detection and identification, and the latter algorithm can be used for the former application with the support of a *fault database*. In addition, any energy usage optimization or control loop supervision implicitly assumes the *in situ* knowledge of the states of the loads being monitored.

This thesis concentrates on state identification, and load disaggregation as its immediate application. The fault diagnosis is not specifically pursued here, but many models and algorithms developed for the state identification and load disaggregation will be useful for diagnosis with minor modification. System analysis is performed only at the level to provide necessary models and parameters for the load identification routines.

The main objective of this thesis is to develop fundamental models and signal processing and information extraction algorithms to benefit many possible future applications based on the NIPM, rather than to develop a complete application package for a specific sector. It is the author's belief that commercial-grade package developments should be performed by business entities. This thesis is intended as a cornerstone in developing numerous future applications, such as a ubiquitous electric meter in a deregulated and volatile utility market, a diagnostic tool for end users or autonomous power systems, a power quality monitor for various entities, etc.

Chapter 2 explains the concept of current harmonic powers, their estimation method and implementation issues. It also illustrates their advantage in the efficient conservation of the information present in a raw current waveform. The following three chapters are dedicated to developing algorithms to identify the status of constant loads. The power signal generated by an active constant load is modeled as a Gaussian random process, and this model is the key in developing various load status identification algorithms.

Chapter 3 deals with the simplest case when load events are isolated. A new classifier is introduced which can take into account both step change observation and transient pattern. The former is analyzed by a maximum-likelihood (ML) classifier and the latter is categorized by a least squares (LS) error criterion. Chapter 4 handles the case when multiple load events happen at a same time. It generalizes the approach in the previous

chapter by searching a possible event space for event candidates and by developing an LS error criterion that can account for multiple patterns at different time indices. Chapter 5 is an endeavor to identify load states without relying on the occurrence of load events. It develops a state model derived from the combination of individual Gaussian random process models, and performs a search to extract state candidates based on the steady-state power observations. Then, an ML classifier selects the best state candidate. The state estimation without load events is useful for the initial state estimation when the program is first started and for resetting the event classifier when it fails.

Chapter 6 develops an algorithm to keep track of the power consumed by a set of VSDs, using the correlations between the fundamental powers and higher harmonic powers. VSDs are the most visible variable loads in commercial or industrial settings, but no method has been available to estimate their power consumption from the non-intrusive power signals, unless an NIPM is solely dedicated to a VSD. Chapter 7 explains how to estimate or eliminate the harmonic signals present in power waveforms. It is a sister chapter to the previous one, because the VSDs tend to generate harmonic signals. It introduces three different estimation methods based on: periodogram, autocorrelation and fractional pitch estimation. One elimination method is presented, based on an autoregressive signal model and white filtering from its parameters. Chapter 8 shows test results, comes up with conclusion and discusses future work.

Chapter 2

Current Harmonic Powers: Mathematical Theory and Practices

This chapter discusses the issue of how to define powers when a current waveform has spectral components other than the fundamental frequency of its corresponding voltage waveform. Definitions of current harmonic powers are proposed, which characterize the harmonic components present in a current waveform with respect to a reference fundamental voltage waveform. The method to obtain them from the current waveform samples via the discrete Fourier transform (DFT) is introduced and its implementation issues are discussed. There has been no consensus on the definition of powers when voltage or current waveforms contain harmonic elements. The proposed current harmonic power is not a complete solution to this problem, rather it is a convenient tool chosen to characterize and analyze the electric loads in the most effective and efficient manner. This chapter provides a basis for the discussions in the following chapters, because it is the current harmonic power we compute from the voltage and current samples.

2.1 Introduction

Since the adoption of AC electric power system over DC system, mainly because of its low transmission cost, harmonics have always been present in power systems. One of the main reasons is the versatility of the DC motor, which requires a type of rectifier to convert AC to DC [Yacamini, 1994]. Only in recent years, have AC motors achieved a comparable level of controllability, ironically at the cost of increased harmonics.

Harmonics refer to spectral components present in a voltage or current waveform, whose frequencies are integer multiples of the fundamental frequency (i.e. 60 Hz in the US) of the voltage waveform. Harmonic currents are created by non-linear loads, such as variable speed drives (VSDs), electronic ballasts for fluorescent lighting, switching power supplies, rectifiers, etc. These harmonic currents are injected to the utility grid and interact with the impedance of the system to create harmonic voltages. Harmonics can cause such damage to utilities and end users as overheating of transformers and motors, inadvertent tripping of relays, incorrect reading of watt-hour meters, etc. [Kennedy, 2000; Driesen *et al*., 1998]. Distribution companies try to minimize the voltage harmonics present in grids and set certain limits on the harmonic currents that can be injected to the grids by their customers [IEEE Standard 519-1992].

In a deregulated electricity market, power quality becomes more important and the harmonics represent a significant portion of the overall power quality problem[2]. A historical account of the harmonics in power engineering can be found in Emanuel [2000]. Current status and future directions of the power system harmonic analysis are reviewed in Xu [2003].

[2] Or, a product, if the power quality is considered as a service to be delivered to customers.

Definitions of the fundamental powers, such as real, reactive and apparent powers, are based on the assumption of a single-frequency sinusoidal voltage waveform and linear loads, which in turn create a monochromatic current waveform with the same frequency but with a possible phase shift. How to define powers when harmonics are present has been debated over the years, but no consensus has been reached yet. The practical definitions of powers under the presence of nonsinusoidal waveforms and unbalanced loads are *recommended* by the IEEE Working Group on Nonsinusoidal Situations [1996]. However, the implication and usefulness of the various power terms proposed in the definitions are rather dubious. The physical meanings of those powers are partially explained in Emanuel [1990].

We adopt the term *harmonic powers* to represent various types of possible powers in harmonic situations. Certainly, the harmonic powers need to be defined and obtained from the harmonic analysis of both the voltage and the current waveform, with a certain amount of cross-correlation between the two waveforms. However, this approach can lead to an explosion of the number of the harmonic power terms. For example, if there are N harmonic terms each in the voltage or the current waveform, N^2 harmonic power terms can possibly be defined. The meaning of those terms is rather unclear, and the practicality and rational of computing those unwieldy components are questionable.

In this chapter, we concentrate on the characterization of the harmonic content of a current waveform, by defining a power-like quantity at each harmonic frequency. We may name it a *current harmonic power*, implying that only the current waveform is used in computation. This approach can be justified by the fact that the voltage distortion (ratio of the sum of harmonic voltage amplitudes to a fundamental voltage amplitude) is usually less than 5 %, whereas the current distortion (similarly defined) can be as high as 150 % [IEEE Working Group, 1996]. Thus, end users will find that the current harmonic power is a practical and effective solution in characterizing and monitoring their electric loads to combat the power quality problem.

Because utilities are more concerned about the quality of the voltage waveform they deliver to their customers, they will find these definitions unsatisfactory. However, the same technique can be extended to characterize the harmonic content of a voltage waveform. The distribution companies can fully characterize the quality of the electricity they market, by the estimation of the current harmonic powers, augmented by an additional harmonic analysis of the voltage waveform.

This chapter starts with the review of the concept of the fundamental powers to obtain a useful insight in deriving the current harmonic powers. Then, the current harmonic powers are formally defined and the relationship with the recommended definitions of the harmonic powers is discussed. The method to compute the current harmonic powers from the samples of a current waveform via the DFT is also proposed and its implementation issues are discussed. The chapter also shows a few examples of the current harmonic power estimations from raw waveforms. Their usefulness in characterizing and monitoring electric loads is also illustrated.

2.2 Fundamental powers

When an electric load of impedance $\mathbf{Z} = R + jX = Z(\cos\theta + j\sin\theta)$ is excited by a sinusoidal voltage waveform with a period T (e.g. 1/60 sec), it draws a sinusoidal current with the same period and with a phase shift determined by its impedance. Using phasors,

$$\mathbf{I} = \mathbf{YV} = \frac{1}{\mathbf{Z}}\mathbf{V}, \tag{2.1}$$

where $\mathbf{Y}$ is the admittance of the load. Thus, when the voltage waveform is given by $v(t) = \mathrm{Re}\{Ve^{j\frac{2\pi}{T}t}\} = V\cos\frac{2\pi}{T}t$, the current lags the voltage by θ, i.e.

$$i(t) = \mathrm{Re}\{Ie^{-j\theta}e^{j\frac{2\pi}{T}t}\} = I\cos(\tfrac{2\pi}{T}t - \theta). \tag{2.2}$$

Note that V and I are peak amplitudes. They are related to the root-mean squared (RMS) values via,

$$V = \sqrt{2}V_{rms}, \quad I = \sqrt{2}I_{rms}. \tag{2.3}$$

Two power terms are associated with the load. The first term $P = {I_{rms}}^2 R$ is called a *real* or *active* power, and is the power consumed by the resistive component of the load. The second term $Q = {I_{rms}}^2 X$ is called a *reactive* power, and is the power periodically stored and released by the reactive component of the load. The two power terms can be combined into a single complex quantity, called an *apparent* power:

$$\mathbf{A} = P + jQ = V_{rms}I_{rms}(\cos\theta + j\sin\theta). \tag{2.4}$$

We observe that the angle of the apparent power is the same as the impedance angle of the load. Figure 2-1 summarizes these relationships.

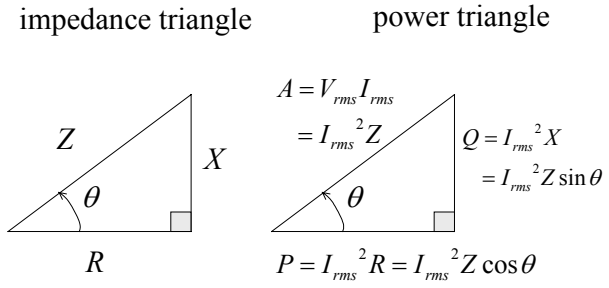


Figure 2-1 Impedance and power triangles.

The real power can be obtained by taking the product of the voltage and the current waveform and computing its one cycle average:

$$\begin{aligned} P &= \frac{1}{T}\int_t^{t+T} v(\tau)i(\tau)d\tau = \frac{1}{T}\int_t^{t+T} V\cos\tfrac{2\pi}{T}\tau\, I\left(\cos\theta\cos\tfrac{2\pi}{T}\tau + \sin\theta\sin\tfrac{2\pi}{T}\tau\right)d\tau \\ &= \frac{1}{T}\int_t^{t+T} VI\cos\theta\frac{1+\cos 2\tfrac{2\pi}{T}\tau}{2}d\tau + \frac{1}{T}\int_t^{t+T} VI\sin\theta\frac{\sin 2\tfrac{2\pi}{T}\tau}{2}d\tau \\ &= \frac{VI}{2}\cos\theta = V_{rms}I_{rms}\cos\theta = {I_{rms}}^2 R. \end{aligned} \tag{2.5}$$

The real power is the time average of the product of the voltage waveform and the in-phase component of the current waveform. This implies that the reactive power will be the average product of the voltage waveform and the 90° out-of-phase, or quadrature, component of the current waveform. Mathematically, it is obtained by shifting the voltage waveform by a quarter cycle before integration:

$$\begin{aligned} Q &= \frac{1}{T}\int_t^{t+T} v(\tau-\tfrac{T}{4})i(\tau)d\tau = \frac{1}{T}\int_t^{t+T} V\sin\tfrac{2\pi}{T}\tau\, I\left(\cos\theta\cos\tfrac{2\pi}{T}\tau + \sin\theta\sin\tfrac{2\pi}{T}\tau\right)d\tau \\ &= \frac{1}{T}\int_t^{t+T} VI\cos\theta\frac{\sin 2\tfrac{2\pi}{T}\tau}{2}d\tau + \frac{1}{T}\int_t^{t+T} VI\sin\theta\frac{1-\cos 2\tfrac{2\pi}{T}\tau}{2}d\tau \\ &= \frac{VI}{2}\sin\theta = V_{rms}I_{rms}\sin\theta = {I_{rms}}^2 X. \end{aligned} \tag{2.6}$$

Here, the shifted voltage waveform does not exist in reality. Rather, it serves as a fictitious reference waveform for mathematical convenience. The fundamental powers can also be interpreted as the correlations between the current waveform and the shifted voltage waveforms.

2.3 Current harmonic powers

The concept of the real, reactive and apparent powers, introduced in the previous section, works well when a load is linear, i.e. it produces a sinusoidal current with a single frequency when excited by a sinusoidal voltage with the same frequency. However, when the load is not linear, the current waveform is not monochromatic even when it is excited by a perfectly clean sinusoid. Rather, the current contains the harmonic components of the fundamental frequency of the voltage waveform. Figure 2-2 illustrates the concept of linear and nonlinear loads.

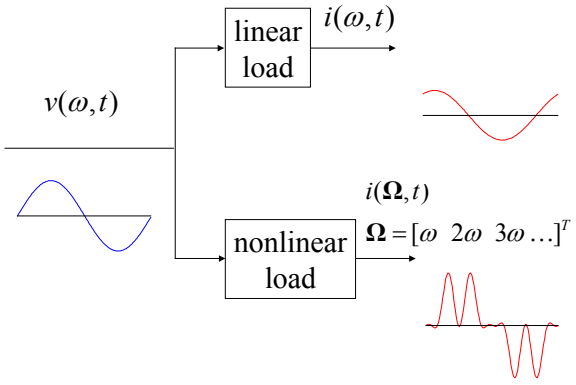


Figure 2-2 Linear and nonlinear loads.

How to define powers in the presence of harmonic currents and how to associate them with meaningful explanations are still unresolved challenging issues. However, the definitions of the fundamental powers in the previous section suggest that we may use a properly shifted harmonic voltage waveform as a *reference* to compute a power at a harmonic frequency. We define the *k*th *current harmonic powers* by the following formulae:

$$\begin{aligned} P_k &= \frac{1}{T}\int_t^{t+T} v(k\tau)i(\tau)d\tau = \frac{1}{T}\int_t^{t+T} V\cos(k\tfrac{2\pi}{T}\tau)i(\tau)d\tau \\ Q_k &= \frac{1}{T}\int_t^{t+T} v(k\tau-\tfrac{T}{4})i(\tau)d\tau = \frac{1}{T}\int_t^{t+T} V\sin(k\tfrac{2\pi}{T}\tau)i(\tau)d\tau \qquad k=1,2,\ldots \end{aligned} \tag{2.7}$$

Note that $P_1 = P$ and $Q_1 = Q$ and hence that the definition is consistent at the fundamental frequency. Thus, the *k*th current harmonic powers are obtained by computing the correlations between the current and the properly shifted *k*th reference waveforms. We may call P_k as the *k*th real power and Q_k as the *k*th reactive power. However, they do not have clear physical meanings except when $k = 1$. Rather, they quantify the magnitude of the *k*th harmonic frequency component present in the current waveform. The phase information is hidden in the relative breakup of P_k and Q_k.

If we want to quantify the *k*th harmonic component of the current waveform without the phase information, we may use the *k*th apparent power A_k defined as:

$$A_k = \sqrt{P_k^{\,2} + Q_k^{\,2}}\,. \tag{2.8}$$

A closer examination of (2.7) reveals that we only need the fundamental voltage waveform (period, magnitude and phase) to compute the current harmonic powers. Thus, the current harmonic powers can still be defined and computed as long as the fundamental frequency waveform can be extracted, even from a (harmonically) distorted voltage waveform.

Because the amount of voltage harmonic distortion is typically very small at end users' sites and because they are more concerned about the current harmonic distortion created by their electric loads, the current harmonic power is a very effective and possibly cost efficient tool to characterize the electric loads and monitor the power quality at their premises. Emanuel *et al*. [1993] conducted a survey of harmonic voltages and currents at several representative customer sites.

Utilities will find the definition of the current harmonic powers less satisfactory, because they are required to supply a clean sinusoidal voltage waveform and try to maintain the voltage harmonic distortion below a certain limit. Thus, they need a separate tool to characterize the harmonic component of a voltage waveform, in order to monitor constantly the *harmonic pollution* and take appropriate actions when necessary. To characterize the harmonic content of a voltage waveform, we may replace $i(\tau)$ with $v(\tau)$ in the definition of the kth current harmonic powers and call the resulting quantities the kth *voltage harmonic intensities*. A survey of the harmonic voltages and currents at several distribution substations can be found at Emanuel *et al*. [1991].

It will be a good idea to review the definitions of powers under harmonic situations, as recommended by the IEEE Working Group [1996], and explore the relationships between the accepted definitions of powers and the current harmonic powers proposed in this chapter. The recommended definitions start by representing the voltage and current waveforms as the sums of harmonic sinusoids.

$$\begin{aligned} v(t) &= V_0 + \sqrt{2}\sum_{h=1}^{\infty} V_h \sin(h\omega t + \alpha_h) \\ i(t) &= I_0 + \sqrt{2}\sum_{h=1}^{\infty} I_h \sin(h\omega t + \beta_h). \end{aligned} \tag{2.9}$$

Note that V_h and I_h are RMS quantities. The RMS voltage V and current I are defined as:

$$\begin{aligned} V^2 &= \sum_{h=0}^{\infty} {V_h}^2 = {V_1}^2 + {V_H}^2, \quad {V_H}^2 = \sum_{h=0,\, h\neq 1}^{\infty} {V_h}^2 \\ I^2 &= \sum_{h=0}^{\infty} {I_h}^2 = {I_1}^2 + {I_H}^2, \quad {I_H}^2 = \sum_{h=0,\, h\neq 1}^{\infty} {I_h}^2. \end{aligned} \tag{2.10}$$

The apparent power S is defined as:

$$S^2 = (VI)^2 = {S_1}^2 + {S_N}^2. \tag{2.11}$$

$S_1 = V_1 I_1$ is called the *fundamental apparent power*. The square of the non-fundamental apparent power S_N has three components:

$${S_N}^2 = (V_1 I_H)^2 + (V_H I_1)^2 + (V_H I_H)^2. \tag{2.12}$$

The first term is called the *current distortion power* and is usually a dominant term. The second term is called the *voltage distortion power*, while the third term is called the *harmonic apparent power*.

Because a current harmonic power is obtained as a correlation between a reference waveform obtained from the fundamental voltage waveform and the current waveform, we recognize that the current harmonic power roughly belongs to the realm of the current distortion power. The difference is that the former represents the spectral component of a current waveform at each harmonic frequency, while the latter quantifies the aggregate amount of the current harmonics in an RMS sense.

2.4 Method to compute current harmonic powers

Once we define the current harmonic powers, the next issue is how to compute them from the measurements of raw voltage and current waveforms, possibly in real time. Their definitions suggest that, once we extract a cycle of the fundamental voltage waveform from the voltage measurement, we can create reference waveforms and thus the current harmonic powers can be obtained by performing integrations on the products of the references and the current.

Leeb developed an analog current harmonic power estimator, composed of a phase-locked-loop to extract the fundamental voltage waveform, waveform synthesizers to create reference waveforms, multipliers and low-pass filters to simulate integration [Leeb, 1993; Leeb *et al*., 1995]. This process of reference waveform generation, multiplication and integration can also be performed by software using sampled waveforms [Shaw *et al*., 1998]. However, one can easily imagine that this is a computationally intensive process.

In this section, we develop a more efficient current harmonic power estimation routine based on sampled waveforms, using the relationship between the continuous-time Fourier series (CTFS) and the discrete Fourier transform (DFT) via the sampling theory. Using the DFT to find the CTFS harmonic power terms is discussed in Z. Luo *et al*. [2002] and El-Habrouk and Darwish [2001]. Other techniques are also reported in the literature, such as the one based on neural network [Moreno-Eguilaz *et al*., 2000; Osowski, 1992] and the one based on singular value decomposition [Osowski, 1994]. Rechka *et al*. [2002] conducted a comparative study of the harmonic detection algorithms.

Consider a voltage signal $v(t) = v_1(t) + v_H(t)$ and a current signal $i(t)$ observed for $0 \le t < T$, where $v_1(t) = V\cos(\frac{2\pi}{T}t)$ is the fundamental voltage waveform and $v_H(t)$ is the additive harmonic voltage distortion. The definitions of the current harmonic powers for the given interval are restated:

$$P_k = \frac{1}{T}\int_0^T V\cos(k\tfrac{2\pi}{T}\tau)i(\tau)d\tau, \quad Q_k = \frac{1}{T}\int_0^T V\sin(k\tfrac{2\pi}{T}\tau)i(\tau)d\tau \qquad k = 1, 2, \ldots \qquad (2.13)$$

Note that the powers are defined for a given interval (period), and in general will have different values at different time instances.

Next, we consider a periodic extension of the current waveform defined as:

$$\tilde{i}(t) = \tilde{i}(t + rT) \quad r = 0, \pm 1, \pm 2, \ldots, \qquad \tilde{i}(t) = i(t) \quad 0 \le t \le T. \tag{2.14}$$

Then, $\tilde{i}(t)$ can be represented in terms of the CTFS:

$$\tilde{i}(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(k \tfrac{2\pi}{T} t) + \sum_{k=1}^{\infty} b_k \sin(k \tfrac{2\pi}{T} t), \tag{2.15}$$

where the Fourier series coefficients a_k and b_k are given by the following formulae:

$$a_k = \frac{2}{T}\int_0^T i(\tau)\cos(k \tfrac{2\pi}{T}\tau)d\tau, \quad b_k = \frac{2}{T}\int_0^T i(\tau)\sin(k \tfrac{2\pi}{T}\tau)d\tau \qquad k \ge 1. \tag{2.16}$$

We note that the current harmonic powers are simply the scaled versions of the Fourier series coefficients, i.e.

$$P_k = \frac{V}{2} a_k, \quad Q_k = \frac{V}{2} b_k. \tag{2.17}$$

Thus, the current harmonic powers of a given interval can be obtained by finding the Fourier series coefficients of the periodic extension of the current waveform with an appropriate scaling determined by the amplitude of the voltage waveform.

In terms of the continuous-time complex Fourier series [Oppenheim *et al*., 1998]:

$$\tilde{i}(t) = \sum_{k=-\infty}^{\infty} \hat{i}_k e^{jk\frac{2\pi}{T}t}, \quad \hat{i}_k = \frac{1}{T}\int_0^T \tilde{i}(t) e^{-jk\frac{2\pi}{T}t} dt. \tag{2.18}$$

Because the current waveform is real, a_k and b_k are also real, and it can be shown that:

$$a_k = \mathrm{Re}(2\hat{i}_k), \quad b_k = -\mathrm{Im}(2\hat{i}_k) \qquad k \ge 1. \tag{2.19}$$

There also exists a continuous-time Fourier transform (CTFT) of the current signal:

$$I(j\Omega) = \int_{-\infty}^{\infty} i(t) e^{-j\Omega t} dt. \tag{2.20}$$

If the transform is evaluated at the harmonic frequencies, i.e. $\Omega = k\frac{2\pi}{T}$, considering the fact that $i(t)$ is time-limited,

$$I(jk\tfrac{2\pi}{T}) = \int_0^T i(t)e^{-jk\frac{2\pi}{T}t}dt = T\hat{i}_k. \text{ Or, } \hat{i}_k = \frac{1}{T}I(jk\tfrac{2\pi}{T}). \quad (2.21)$$

Thus, the Fourier series coefficients are simply the scaled versions of the Fourier transform evaluated at the harmonic frequencies.

$$a_k = \mathrm{Re}(2\hat{i}_k) = \mathrm{Re}\left(\frac{2}{T}I(jk\tfrac{2\pi}{T})\right), \quad b_k = -\mathrm{Im}(2\hat{i}_k) = -\mathrm{Im}\left(\frac{2}{T}I(jk\tfrac{2\pi}{T})\right) \qquad k \geq 1. \quad (2.22)$$

The current harmonic powers can also be represented in terms of the Fourier transform:

$$P_k = \frac{V}{2}a_k = \frac{V}{T}\mathrm{Re}\big(I(jk\tfrac{2\pi}{T})\big), \quad Q_k = \frac{V}{2}b_k = -\frac{V}{T}\mathrm{Im}\big(I(jk\tfrac{2\pi}{T})\big). \quad (2.23)$$

If the Fourier transform of a continuous current signal can be computed, its current harmonic powers can be obtained simply by evaluating the transform at the harmonic frequencies with the scale factors. However, due to the difficulty in performing the CTFT, the signal is typically sampled and the DFT is performed instead. Thus, we need to find a relationship between the CTFT and the DFT, to obtain the current harmonic powers from the DFT of the current signal samples.

Consider a continuous signal $x_c(t)$ and its sampled version $x[n]$ with a sampling period T_s (sec), i.e. $x[n] = x_c(nT_s)$. The discrete-time Fourier transform (DTFT) of $x[n]$ is defined as:

$$X(j\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}. \quad (2.24)$$

Note that $X(j\omega)$ is periodic with 2π, i.e. $X(j\omega) = X(j(\omega+2\pi))$. It is well known from the sampling theory that $X(j\omega)$ is a scaled and aliased version of the CTFT of $x_c(t)$ [Oppenheim *et al*., 1999]:

$$X(j\omega) = \frac{1}{T_s}\sum_{k=-\infty}^{\infty} X_c\left(j\frac{\omega - 2\pi k}{T_s}\right) \qquad \omega = \Omega T_s. \quad (2.25)$$

It should be noted that ω is in [rad], whereas Ω is in [rad/sec]. If $X_c(j\Omega)$ is band-limited to $\frac{1}{2}\Omega_s$, where $\Omega_s = \frac{2\pi}{T_s}$ (rad/sec), no aliasing occurs and $X(j\omega)$ is simply,

$$X(j\omega) = \frac{1}{T_s}X_c\left(j\frac{\omega}{T_s}\right). \quad (2.26)$$

If N points are sampled for a given interval T, with a sampling period T_s, $T = NT_s$ and $x[n]$ is an N-point sequence. The discrete Fourier series (DFS) is obtained from the N-periodic version of $x[n]$ and the DFT $X[k]$ is defined as a frequency-limited version of the DFS. As a result,

$$X[k] = \begin{cases} \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} & 0 \le k \le N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.27}$$

$X[k]$ can also be interpreted as the samples of the DTFT. Combining the relationship between the DTFT and the CTFT,

$$\begin{aligned} X[k] &= X(j\omega)\big|_{\omega=\frac{2\pi}{N}k} \qquad 0 \le k \le N-1 \\ &= \frac{1}{T_s} X_c\left(j\frac{\omega}{T_s}\right)\Bigg|_{\omega=\frac{2\pi}{N}k} = \frac{1}{T_s} X_c\left(j\frac{2\pi k}{NT_s}\right) = \frac{1}{T_s} X_c\left(j\frac{2\pi k}{T}\right). \end{aligned} \tag{2.28}$$

Thus, the CTFT evaluated at a harmonic frequency is simply the scaled version of the DFT:

$$X_c\left(j\frac{2\pi k}{T}\right) = T_s X[k]. \tag{2.29}$$

Now, the current harmonic powers can be represented in terms of the DFT $I[k]$ of the sampled current waveform:

$$\begin{aligned} P_k &= \frac{V}{2} a_k = \frac{V}{T}\mathrm{Re}\left(I(jk\tfrac{2\pi}{T})\right) = \frac{V}{T}\mathrm{Re}\left(T_s I[k]\right) = \frac{V}{N}\mathrm{Re}\left(I[k]\right) \\ Q_k &= \frac{V}{2} b_k = -\frac{V}{T}\mathrm{Im}\left(I(jk\tfrac{2\pi}{T})\right) = -\frac{V}{T}\mathrm{Im}\left(T_s I[k]\right) = -\frac{V}{N}\mathrm{Im}\left(I[k]\right). \end{aligned} \tag{2.30}$$

We note the similarity in the definitions of the current harmonic powers in terms of the CTFT and the DFT, i.e. the real or imaginary part of the transform is multiplied by the peak voltage and divided by a corresponding interval.

An interesting case occurs when the fundamental voltage waveform is not necessarily locked to $V\cos(\frac{2\pi}{T}t)$ for $0 \le t < T$, i.e. when there is a time shift in $v_1(t)$. Consider a sampled waveform $v_1[n]$ for $0 \le n \le N-1$ and assume that shifting $v_1[n]$ by n_0 results in a locked cosine waveform:

$$v_1[n-n_0] = \hat{v}_1[n] = V\cos\tfrac{2\pi}{N}n \qquad n_0 \le n \le n_0 + N - 1. \tag{2.31}$$

A time-shift corresponds to a phase-shift in the DFT:

$$\hat{V}_1[k] = \begin{cases} V_1[k]e^{-j\frac{2\pi}{N}kn_0} & 0 \le k \le \lfloor \frac{N}{2} \rfloor \\ V_1[k]e^{-j\frac{2\pi}{N}(k-N)n_0} & \lfloor \frac{N}{2} \rfloor + 1 \le k \le N-1. \end{cases} \tag{2.32}$$

If we evaluate the transforms at $k = 1$,

$$V_1[1] = \hat{V}_1[1]e^{j\frac{2\pi}{N}n_0}. \tag{2.33}$$

Because $\hat{V}_1[1]$ is a positive constant ($\hat{V}_1[k]$ is the DFT of a pure cosine), the argument of $V_1[1]$ corresponds to the additional phase angle introduced by the phase-shift of the cosine waveform.

Because the fundamental voltage waveform $v_1[n] = V\cos\frac{2\pi}{N}(n + n_0)$, we want to take into account the phase shift in computing the DFT of the current waveform:

$$I'[k] = \sum_{n=0}^{N-1} i[n]e^{-j\frac{2\pi}{N}k(n+n_0)} = e^{-j\frac{2\pi}{N}kn_0}\sum_{n=0}^{N-1} i[n]e^{-j\frac{2\pi}{N}kn} = e^{j\phi k}I[k], \tag{2.34}$$

where $I[k]$ is the regular DFT of the current waveform and $\phi = -\frac{2\pi}{N}n_0 = \arg(V_1^*[1])$ (* denotes complex conjugate). However, $V_1[1]$ is equal to $V[1]$, due to the independence of the harmonic components, when the voltage signal is band-limited to $\frac{\pi}{T_s}$ (rad/sec). Thus, $I'[k]$ is a phase-shifted version of $I[k]$, with the phase shift obtained from the DFT of the voltage waveform, and must be used in place of $I[k]$ in computing the current harmonic powers.

The above method of using the phase angle of the first harmonic component of the voltage DFT is convenient because it does not require the sequence of samples to start at the zeroth angle position of the voltage cosine waveform ($V\cos 0$). Sampling can start at any point as long as it samples N equally spaced points per cycle and one of the samples is taken at the zeroth angle. Any time shift will be taken care of by the phase of $V[1]$.

In the implementation point of view, it is an inefficient process because we have to perform the N-point DFT on the voltage samples to obtain a single scalar output. However, it is robust because $V[1]$ is immune to other harmonic components that might be present in the voltage waveform. If this approach is not taken, a certain type of phase-locked-loop algorithm needs to be applied to the voltage waveform samples. One can easily imagine that such an algorithm will perform poorly when there exist other harmonic components and noise. Where resources are limited, a single-bin DFT output routine, such as the Goertzel algorithm, can be used to compute $V[1]$ only, instead of computing the whole N-point DFT [Jacobsen and Lyons; Oppenheim *et al*., 1999].

2.5 Implementation issues

The previous section shows that the current harmonic powers for a given interval can be computed efficiently by taking N samples of a current waveform and taking their N-point DFT. However, a few conditions, imposed in deriving the relationship between the CTFS terms and the DFT bins, need to be observed by any current harmonic power estimator based on the DFT.

The first condition is that both the voltage and current waveforms need to be band-limited to the Nyquist frequency of $\frac{1}{2T_s}$ Hz. This can be achieved easily by low-pass filtering the raw waveforms either before or after sampling. However, because the filtering can not be ideal, certain amounts of attenuation and aliasing are inevitable. The distortion introduced by the low-pass filtering is usually negligible.

The second condition is that N equally spaced samples need to be taken for a period. If the phase rotation method based on $V[1]$ is used, sampling can start at any point as long as the zeroth angle position of the fundamental cosine waveform is included in the sampled sequence. If not, sampling should start at the zeroth angle position. At first, this seems to be a trivial requirement, but in reality, it is the most difficult condition to meet in implementing a real-time high-precision current harmonic power estimator.

The difficulty stems from the fact that the generator of a utility grid and the clock of a data acquisition (A/D) card can not be synchronized. Although the nominal period of the AC voltage is 1/60 sec in the United States, there usually are slight variations over time. In addition, it is common to observe interruptions or transients in the voltage waveform, the discontinuations of the sinusoid voltage waveform for short durations. Thus, even if we start sampling at the correct zeroth angle position, as time goes by, the sampler may obtain an inconsistent number of samples per cycle (e.g. $N - 1$ or $N + 1$ samples per cycle) and the first sample of a cycle may not necessarily be at the correct zeroth angle position. Thus, the sampler somehow has to check the shape of the fundamental voltage waveform continually to adjust its sampling period and sampling start point accordingly to satisfy the condition.

This is an example of a rather common transmitter-receiver synchronization problem, and is related to the issues of non-integer delay, sampling rate conversion and interpolation [Laakso, *et al*., 1996]. In digital communication, a receiver must be synchronized to incoming symbols. It can be accomplished either by an analog feedback or feedforward control of the phase and interval of a sampling clock, or by an autonomous sampler followed by a fractional-delay filter to estimate symbols at correct sampling intervals [Gardner, 1993; Erup *et al*., 1993; Bingham, 1988].

Shaw [2000] approached this problem by first sampling the raw voltage and current waveforms at a higher sampling rate (e.g. 16 kHz) and then filtering and resampling them at a lower rate to obtain N (e.g. 128) samples consistently per cycle. The period and phase of the voltage waveform can be estimated by finding adjacent zero-crossings of the

filtered voltage samples. New samples are obtained by interpolating old samples at estimated sampling locations.

However, one can easily imagine that this filtering and resampling process will not perfectly reconstruct the original bandlimited waveforms and thus certain amounts of errors (both in magnitude and location) will be introduced to the new samples. The magnitude errors will eventually introduce certain biases or random fluctuations to the current harmonic power estimations. We may treat them as noise injected by a non-ideal power estimation process.

The effect of the location errors is rather subtle. Because the process resamples the waveform starting from an approximate zeroth angle position, the zero-crossing estimation determines the accuracy of the new sample locations. The uncertainty of the exact zero-crossing location will result in a shift or bias in the locations of the resampled data points. The actual accuracy depends on the type of zero-crossing estimation algorithm used and the nature of signal. However, it is reasonable to assume that the position uncertainty is on the order of one sample interval (after resampling). If we denote the unknown bias by n_0 and assume that $v[n-n_0] = \hat{v}[n] = V\cos\frac{2\pi}{N}n$, the bias introduces a phase-shift at each DFT bin, following (2.32). Note that $n_0 \sim O(1)$ and that n_0 is a fraction in general.

Thus, the bias in the sampling position results in the bias in the phase of a current harmonic power estimation. The phase bias increases with increasing harmonic number and becomes a maximum when the harmonic number reaches $\frac{N}{2}$. The *k*th current harmonic power will have a phase bias or uncertainty on the order of $\frac{2\pi}{N}k$ (rad). With *N* = 128, this corresponds to 2.81 (deg) for the fundamental powers and 19.69 (deg) for the seventh harmonic powers. The phase bias can become quite significant even at modest harmonic numbers. Also, if the position bias is not consistent, the resulting phase bias of a current harmonic power can move from one extreme to another, creating large fluctuations in its magnitude. The phase bias can be reduced by increasing the sampling rate (larger *N*) and by using a more accurate zero-crossing estimation algorithm.

Recursive estimation techniques can be employed to better estimate voltage waveform parameters and hence to reduce the sampling position bias. The Kalman filter is one of the most well known tools to estimate the magnitude, phase and period of the voltage waveform in a power system [Beides and Heydt, 1991; Dash *et al*., 2000]. However, these techniques may not work properly when there is a significant amount of harmonic distortion in the voltage waveform. A more robust and efficient fundamental voltage waveform estimation technique needs to be investigated in future.

Where the phase bias is a concern, one may use the apparent powers to characterize the harmonic content of a current waveform, instead of using the separate real and reactive powers. The *k*th apparent power is the magnitude of the *k*th complex power, and thus is free from the precarious time-shift of the reference voltage waveform. However, the

dimension of the current harmonic powers is reduced by half, which can reduce their resolving capability in characterizing and analyzing electric loads.

A high degree of parallelism is also required between the voltage and the current samples, such that two waveforms are sampled at the same time instances. The A/D card should be able to sample the voltage and current simultaneously using parallel channels. The filtering and resampling, if used, need to be done with care to ensure a common sampling location for both the voltage and the current sample at a given time.

In this manner, N samples are taken from a cycle of current and N current harmonic power terms are computed using the DFT. It can be conceived as an N-point windowing operation on a stream of discrete current waveform samples. The window is used to extract a cycle of the signal, and the window is shifted to the next cycle. If there is no overlap between the windows, the current harmonic powers are computed at every cycle and their update rate is $1/T$ Hz (e.g. 60 Hz in the US). However, if we shift the window by less than a cycle, creating an overlap between successive windows, the powers are generated faster than $1/T$ Hz. For example, if we shift the window by $N/2$ points, the powers are computed at every half cycle and their rate is $2/T$ Hz. Shifting the window by less than N points can be useful, as it provides a means to interpret between the otherwise $1/T$ Hz power samples. It can be conceived as an interpolation.

So far, we have assumed a single-phase power system. But the majority of industrial and commercial customers use three-phase power systems. Defining the current harmonic powers in the three-phase situation is a rather complicated problem, in part because there is no established standard in defining three-phase powers [IEEE Working Group– survey, 1996]. This is an unresolved problem and will be investigated in future. Until then, we will monitor only a single phase out of a three-phase system and compute its current harmonic powers. When total three-phase (fundamental) powers are necessary, we will simply multiply the single-phase powers by three, assuming balanced loads.

2.6 Examples

Measurements were taken at the main HVAC room of a commercial building at 160 Sansome St. San Francisco, CA. The HVAC loads are served by a 480 V three-phase power system. The voltage was measured across the A phase and the neutral, and the current samples were taken at the A phase.

Figure 2-3 shows the raw voltage and current waveforms over three cycles. At the time of measurement, no HVAC load was on. However, the HVAC panel also serves a number of small lighting and plug loads. We may call them collectively the *background load*. We note that the voltage waveform is a fairly clean sinusoid, whereas the current waveform is slightly distorted. Thus, we expect the presence of higher harmonics.

Figure 2-4 shows the current harmonic powers (of single phase) computed from the waveforms in Figure 2-3. Three cycle averages up to the sixteenth harmonic are reported

in the figure, although $(\frac{N}{2}-1)$ harmonics are available when N samples are taken per cycle. We observe that the fundamental powers are dominant terms. However, there also exist substantial higher harmonic powers, especially the fifth and seventh components.

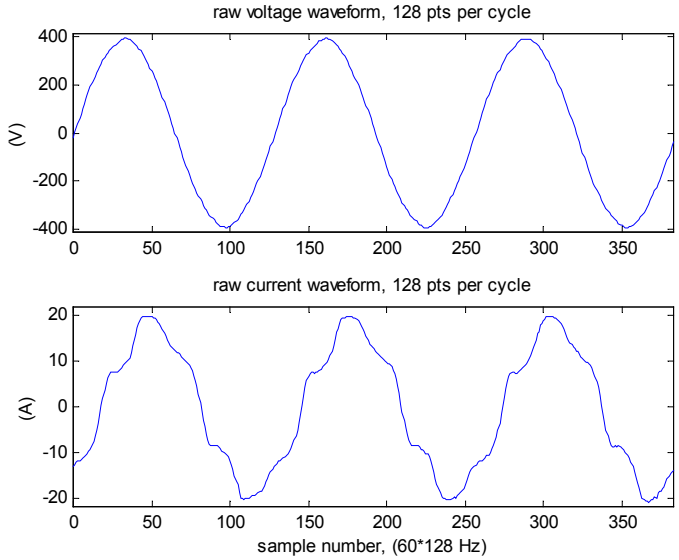


Figure 2-3 Voltage and current waveforms of background load.

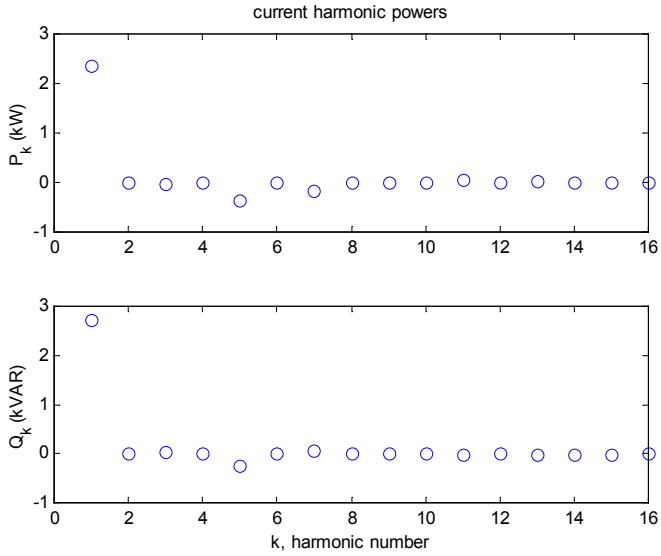


Figure 2-4 Current harmonic powers of background load.

Figure 2-5 shows the waveforms when the 100 hp (air) supply fan is turned on in addition to the background load. The capacity (speed) of the fan is controlled by a VSD and thus the resulting current waveform is highly nonlinear. We note that the voltage waveform is relatively clean. Figure 2-6 shows the resulting current harmonic powers. Fundamental powers are still significant terms, but a few higher harmonic powers, especially the fifth and seventh, are on the order of same magnitude.

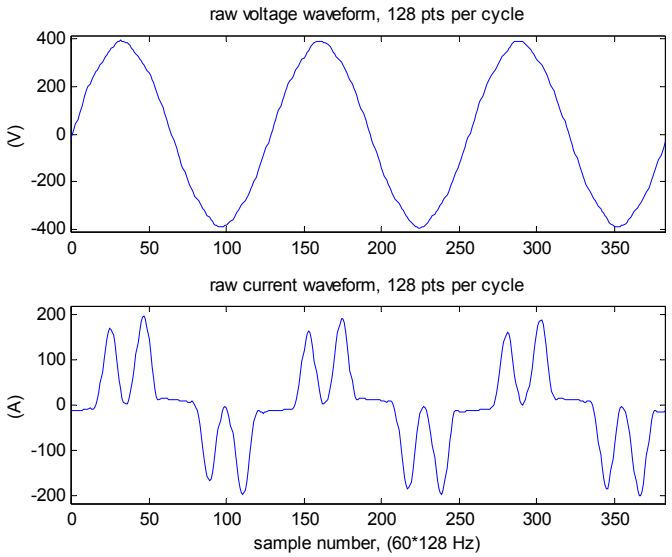


Figure 2-5 Voltage and current waveforms of supply fan plus background load.

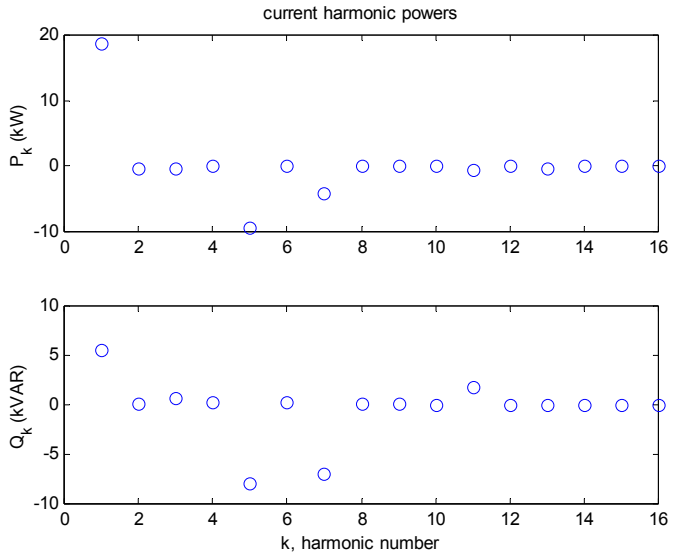


Figure 2-6 Current harmonic powers of supply fan plus background load.

An interesting insight can be obtained by looking at the current harmonic power estimation from the viewpoint of signal compression. We remember that the current harmonic powers are nothing more than a spectral characterization of the current waveform based on the DFT. Thus, the signal energy[3] should be conserved before and

[3] For a deterministic N-point sequence $x[n]$, its signal energy E is defined as,

$$E = \sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2 .$$

The latter equation is commonly referred to as Parseval's relation for the DFT.

after the transform following Parseval's relation [Oppenheim *et al*., 1999]. If we conserve a few significant odd harmonic powers, both real and reactive, and get rid of the rest, we can achieve a significant amount of signal compression while conserving most of the information originally present in the current waveform.

For example, from Figure 2-6, if we retain odd harmonic powers up to the seventh, we achieve the signal compression ratio of 93.75 %. Yet, we retain 99.29 % of the original signal energy. If we retain up to the thirteenth odd harmonic powers, the signal compression ratio drops to 89.06 %. However, the signal energy conservation ratio increases to 99.92 %. There is a tradeoff in choosing the number of harmonic powers to keep. Retaining more harmonic powers is obviously better to transmit more information, but it will increase the data handling and computation cost in subsequent analyses.

Thus, we can conceive the computation of the current harmonic powers from raw current waveform samples as a type of *coding* process. The information present in the signal is converted from one form to another, and the format is chosen so that the efficiency in transmission and storage can be maximized. The current harmonic powers are the codes chosen to represent the spectral components of the current signal, and the coding is very efficient in the sense that only a few symbols are necessary to carry the essential information present in the signal. Thus, we may call the overall process of estimating the current harmonic powers from the raw waveform samples as *power coding*, and the estimator as *power coder* or *pocoder*, in so much as voice coders are called *vocoders* in speech processing.

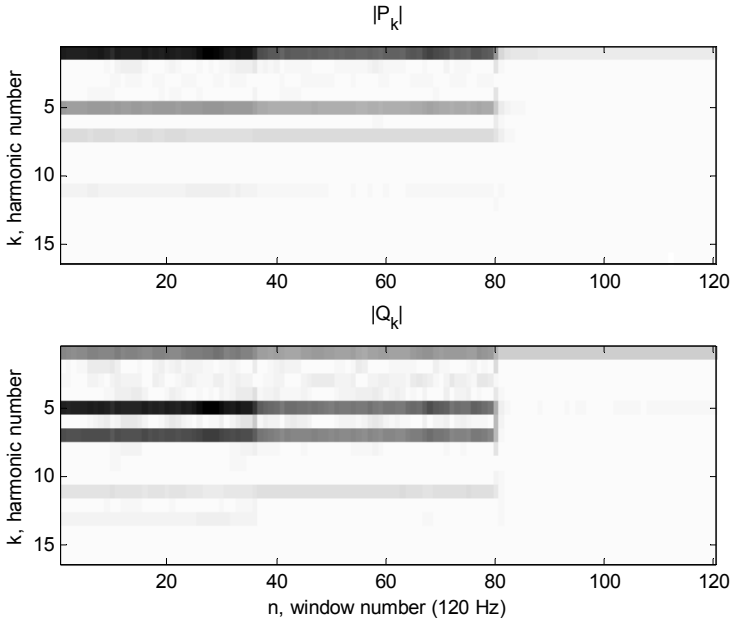


Figure 2-7 Magnitudes of real and reactive harmonic powers over successive windows.

Figure 2-7 shows the magnitudes of P_k and Q_k in gray scale. Larger magnitudes are indicated by darker pixels. Window size N is 128, and the window is shifted by $N/2$. Thus, the power coder generates harmonic powers at 120 Hz. Initially, both the supply

fan and the return fan (75 hp) are on. At n = 36, the return fan is turned off. We note that the magnitudes of the major odd harmonic powers are reduced at the return fan turn-off event. We also observe that Q_{13} is mostly generated by the return fan, because it becomes negligible after n = 36. This phenomenon is more clearly visible in Figure 2-8, where the intensities of P_k and Q_k are plotted in a logarithmic scale.

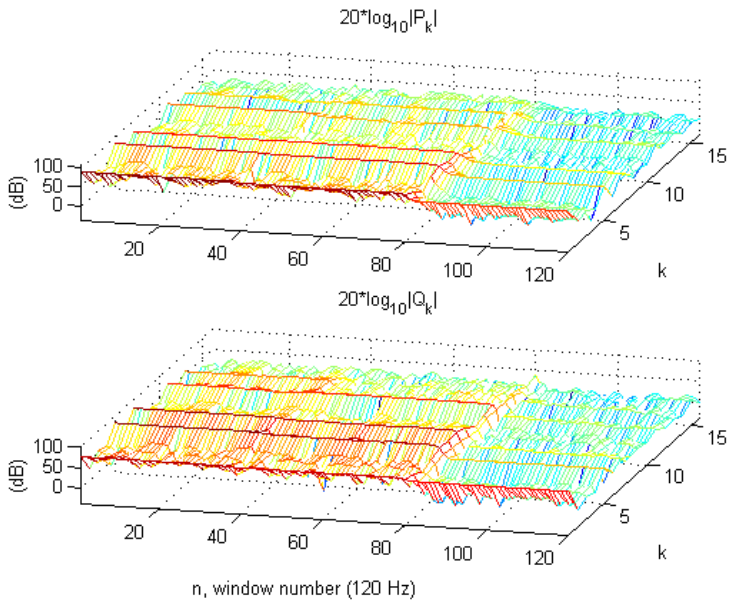


Figure 2-8 Intensities of real and reactive harmonic powers over successive windows.

This example illustrates the fact that an electric load generates a set of characteristic harmonic powers when it is on, and that its characteristic may be visible even when other loads are on, which add up their own characteristic harmonic powers to our observations. Figure 2-8 hints that by monitoring the magnitude of Q_{13}, we can make a relatively sound judgment whether or not the return fan is on, because the supply fan does not generate a significant amount of Q_{13}.

The supply fan also turns off at n = 80. This event is indicated by large reductions in the higher harmonic powers. After then, only the background load is on. The figures in this section illustrate that electric loads generate characteristic harmonic powers and the observation at a given time is their superposition. Thus, by analyzing the composition of the harmonic powers and monitoring their changes, we should be able to extract the information about the operational status of the loads.

2.7 Chapter summary

The current harmonic powers were defined in this chapter, and a DFT based method to compute them was also explained. Their usefulness in characterizing electric loads and monitoring their status was also illustrated through the examples.

How to define powers when harmonics present is a much debated yet still unresolved issue. The current harmonic power is not a panacea to this annoying ailment, rather it is a convenient and computationally-efficient (partial) remedy based on the fact that the majority of the distortion is present in the current waveform. As the estimation of the spectral components of a current waveform, the current harmonic power can be used to quantify current harmonic distortions and to extract the information about the attributes and status of the electric loads. The voltage harmonic intensity was also defined, which can be used to estimate voltage harmonic distortions in a complimentary manner. Its computation can be performed similarly using the DFT.

The current harmonic power computation method based on the DFT of the current waveform samples is a very efficient tool compared to other approaches based on the reference waveform generation and integration. A few conditions were identified to estimate the current harmonic powers accurately via the DFT. However, obtaining equally spaced samples starting at the zeroth angle position of a reference waveform is the most difficult problem due to the generator-sampler synchronization problem. The filtering and resampling approach was discussed and the phase bias of a current harmonic power due to the position bias was also discussed. The apparent powers may be used where the phase bias is a major concern.

An intriguing insight was obtained by treating the estimation of the current harmonic powers from raw current waveform samples as a coding process. The current harmonic powers are the codes we chose to carry the essential information present in a waveform. The coding is very effective (high signal energy conservation) and efficient (high signal compression) in the sense that only a few symbols are necessary to characterize the original signal. For example, we are able to reconstruct the most of the original current waveform with only a few odd harmonic power terms. Also, the examples show that analyzing those symbols retained after the coding can yield valuable information about the character and status of the monitored loads.

Chapter 3

Single Load Event Identification using Steady-state Load Classification and Transient Pattern Recognition

When a non-intrusive power monitor is installed at the main electric service entry of a site, the most basic and important information rendered by the monitor is the operational status of the loads and the subsequent disaggregation of the total power observations into individual power consumptions. The traditional method of load status identification analyzes the sequence of steady-state power changes at a central location. However, the method performs poorly when there are similar loads and does not provide the capability of real-time load disaggregation.

In this chapter, a new approach is proposed: load identification using both *steady-state powers* and *transient patterns*. The transient pattern of a load is mostly unique and repeatable and typically provides more discernibility than its steady-state power consumption. Utilizing two complementary features of a load results in a more accurate and robust load classification algorithm. A decision is made locally based on a single observation, which provides real-time load disaggregation capability.

3.1 Introduction

This chapter presents two load disaggregation methods, *steady-state load classification* and *transient pattern recognition*, and their synthesis. The method of steady-state load classification was developed more than a decade ago and has been used widely including a prototype residential application [Hart, 1992; Drenker and Kader, 1999]. However, the method does not perform well when two loads are close to each other in the steady-state mapping space. The method also suffers when two load events happen sufficiently close in time so that they can appear as a single event.

The problem of load proximity is related to the nature of loads and will not be solved completely, although improvements can be made. The problem of temporal overlap of events can be solved partially by increasing the time resolution of power waveforms. However, increasing the time resolution introduces another problem of how to deal with the transient responses of the loads. Traditionally, the transient responses are regarded as nuisances in measuring the steady-state powers and eliminated as much as possible by an adequate filtering and downsampling [Karl *et al*., 1992; D. Luo *et al*., 2002].

On the other hand, the study shows that the transient response of a load is relatively *unique* and *repeatable*, which implies that the *transient patterns* can be used to identify the loads [Norford and Leeb, 1996]. This approach nullifies the question of how to eliminate the transient responses and provides an additional resolving power when similar loads exist.

By utilizing two distinctive features of a load, steady-state power and transient pattern, in a complementary manner, the task of load identification becomes more viable. Also, due to the improvement in the identification accuracy, the load classification decision can be made locally after a single observation. Compared to the conventional method, characterized by the local collection, central processing and off-line disaggregation, the proposed method collects and processes data locally and has real-time load disaggregation capability.

This chapter first develops the method of steady-state load classification. The overall structure of the method is first presented, followed by the development of its individual components. Then, its performance is evaluated, which naturally proves the necessity to utilize the transient patterns to compensate its limitation.

The second half of the chapter is devoted to the development of a transient pattern recognition method. It first introduces the problem of pattern recognition in conjunction with the electric load identification. It then develops the overall structure and the details of the method. It also discusses the issue of how to coordinate the outputs from two inherently different load classifiers. The overall load disaggregator with two parallel classifiers is tested with the data under various conditions.

3.2 Load classification based on state space mapping

The method of steady-state load classification relies on the fact that constant loads consume relatively fixed amounts of the real and reactive powers and that they tend to occupy distinctive locations (clusters) once mapped to the complex power plane (real vs. reactive power). The steady-state load classification method works reasonably well at residential sites, where the majority of loads are constant and small in size, and thus where measuring the step heights are relatively easy and the mappings can be conducted with fidelity [Carmichael, 1990].

Two essential attributes of the method are measuring the step heights, both real and reactive, when a load is turned on or off, and identifying the event via mapping the measurements on the complex power plane. Traditionally, the mapping has been performed on the complex power plane whose axes represent the real (P) and the reactive (Q) power. However, higher harmonic powers (P_3, Q_3, etc.) can also be introduced when certain conditions are met, as additional dimensions to identify the steady state of a load, and consequently the mapping can be performed in the complex harmonic power *space* whose axes represent the fundamental and harmonic powers. The complex power space is defined as the state space of the electric loads.

The problem of step-height measurement is essentially how to define the steady state of a load and detect its change. The steady state of a load can be defined as the state of the load when it consumes a relatively constant amount of power (real and reactive) after it is turned on. A steady-state event detector should be able to handle various transients and

measure the step heights of an event by computing the differences in the steady-state powers before and after the event.

The state-space mapping is composed of two parts. The first part is the construction of load classes in the complex power space. Conventionally, no prior information about the load classes was given and the mapping algorithm grouped the step-height measurements into load *clusters*. Alternatively, the information about the load classes can be collected *a priori* from a set of test data and the mapping can be conducted against the pre-constructed load classes.

The second part is to classify the step heights of a steady-state event among the load clusters or classes. Obviously the design of a classifier depends on how much the prior information is known, and to less extent how step heights are measured.

However, the biggest problem in designing a steady-state load classifier is the possible spatial proximity of the load classes in the complex power space. This creates an ambiguity when a steady-state event is mapped to a point in the vicinity of those classes. The influence of the cluster proximity can be reduced partially by applying the techniques such as changing bases of observation vectors using the Karhunen-Loève expansion [Fukunaga, 1972].

Figure 3-1 shows the scatter plot of the steady-state powers of four 10-hp water pumps in an office building at 160 Sansome St. San Francisco, CA. Each point of the plot is an eight-point average of the 120 Hz data. The graph clearly shows the overlap between the chilled water pump 1 (CHWP1) and the CHWP2 and, to a less extent, between the CHWP1 and the cooling tower pump 2 (CTWP2). Thus it will be very difficult to identify a steady-state event that happens to fall in the vicinity of these three clusters.

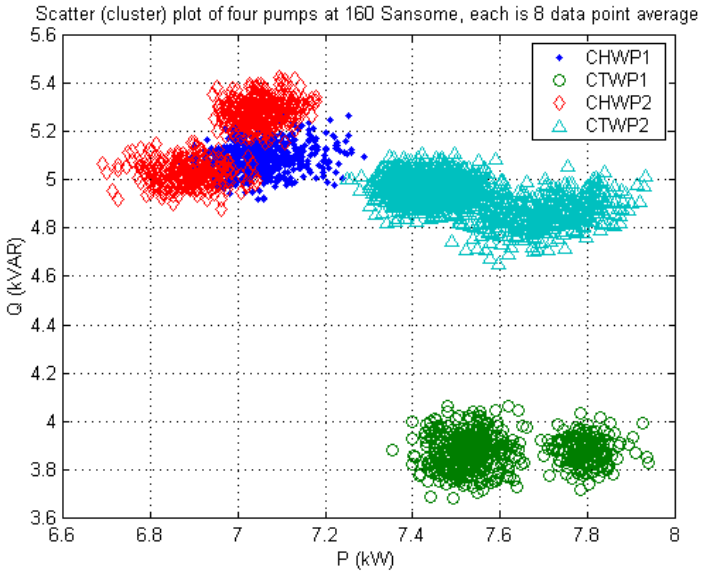


Figure 3-1 Scatter plot of the steady-state powers of four 10-hp water pumps.

The figure also reveals an interesting fact that a load cluster can be composed of several distinctive agglomerates. The CTWP1 cluster, for example, is composed of two clearly separated agglomerates. It is because, at each different ON event, the power consumption of a load can settle on a slightly different mean and furthermore the mean value can change over time.

In some cases, increasing the dimension of the state space provides an additional resolving power to the steady-state load classifier. The mapping is conducted in a harmonic hyperspace spanned by P, Q, P_3, Q_3, … Each basis is orthogonal to each other. This technique is useful in resolving a conflict, when two loads have overlapping clusters in the P-Q plane but they have different positions on a harmonic axis [Laughman *et al*., 2003].

The method of steady-state load classification should be able to measure the step heights, construct the load classes and have a classification rule to identify an event. The design should be flexible enough to select an arbitrary dimension for the harmonic state space.

The power coder generates a stream of harmonic power vectors. Currently, a power vector has a length of eight with odd harmonic powers up to the seventh ([P Q P_3 Q_3 P_5 Q_5 P_7 Q_7]). It is desirable to reduce the number of the vectors, as long as the power vectors after the reduction still contain enough information to identify the steady-state events. Figure 3-2 shows the block diagram of the steady-state load classification.

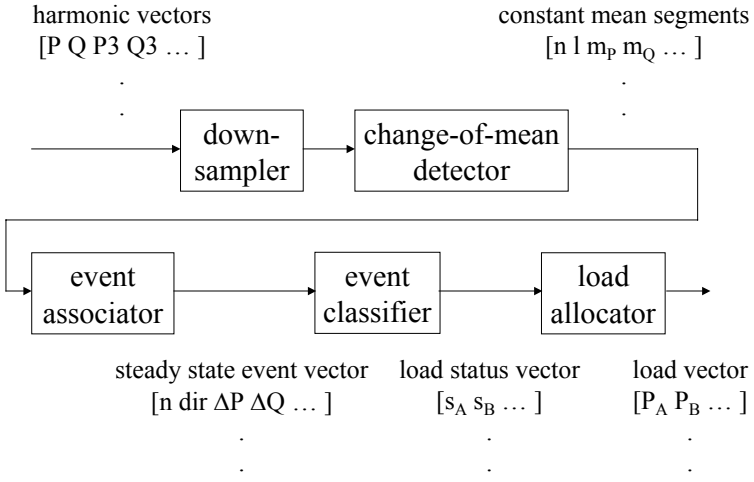


Figure 3-2 Block diagram of steady-state load classification.

The reduction of the harmonic vectors can be achieved easily by downsampling. The downsampling can be followed (or preceded) by a filtering or it can perform its own filtering. For example, if a d-point downsampling is achieved by taking a static average of d data points, the downsampler has a built-in d-point averaging filter. On the other hand, a separate filter can be used to smoothen the output from a downsampler. A variety

of filters can be considered and their selection is important especially when the transients need to be filtered out [Karl *et al*., 1992].

The change-of-mean detector analyzes the stream of power vectors and groups them as a series of constant-mean segments. The output is the stream of constant-mean segments, each of which contains the index (or time), length, mean estimations, variance estimations, etc. It basically extracts sufficient information to recognize the steady-state events, from the stream of power vectors in the form of the constant-mean segments.

The event associator analyzes the constant-mean segments and associates a series of them with a steady-state event. An OFF event is easily identified with a pair of segments, the second of which has a smaller mean compared to the first segment, in the condition that the difference is larger than a prescribed threshold. The identification of an ON event is more involved. It first has to recognize the beginning of a transient and make appropriate decisions for the subsequent constant-mean segments.

If all the loads to be monitored have a similar time scale in their transient response, the event associator and the change-of-mean detector can be combined into a single steady-state event detector preceded by an adequate downsampling and filtering to eliminate the transients. However, where the loads have heterogeneous time scales, an independent event association is necessary, because each load has a different time span in its transient response to be accounted for.

Upon the successful association of a steady-state event, the event associator generates a steady-state event vector, which contains the index, direction (ON/OFF), step heights, etc. The step heights are simply the differences of the mean estimations between the segment before the event and the segment where a steady state is reached.

The event classifier identifies the steady-state event vector with a load event by mapping the step-heights in the state space. A metric is computed for each load class and the load class with the most favorable metric is selected. If no class has a favorable metric, the event is either unclassified or dismissed. The event classifier outputs a load status vector, i.e. the ON/OFF statuses of all the loads monitored in a single vector form.

The load allocator distributes the total power observed to the individual loads based on the load status vector. Any amount corresponding to an unclassified event is added to an *unclassified* account. The leftover is assigned to the background load.

The general structure of the steady-state load classifier was explained and its five major components were introduced in this section. For the load classifier to be developed in this chapter, a downsampler based on the *d*-point static averaging is selected because it is easy to implement and analyze, and performs a decent amount of filtering. A general change-of-mean/variance detector is developed in Appendix A, which we will employ as the change-of-mean detector for the steady-state load classification.

3.3 Event associator

The change-of-mean detector outputs a time-ordered list of constant mean/variance segments from the analysis of its input signal. The event associator analyzes the outputs from the change-of-mean detector, and associates a part of them with a steady-state event when certain conditions are met. It will be helpful to examine typical outputs of the change-of-mean detector when there are steady-state events.

Figure 3-3 shows the turn-on and turn-off events of a 10-hp single-speed water pump. The original data were taken at 120 Hz and are downsampled by taking the 8-point static averages. The change-of-mean detector is applied to the P stream with the maximum number of mean change $k_{max} = 3$, minimum segment length $p = 3$ and number of transient points $q = 1$. The window-shift L is 45 and the nominal window-overlap is $p + q$.

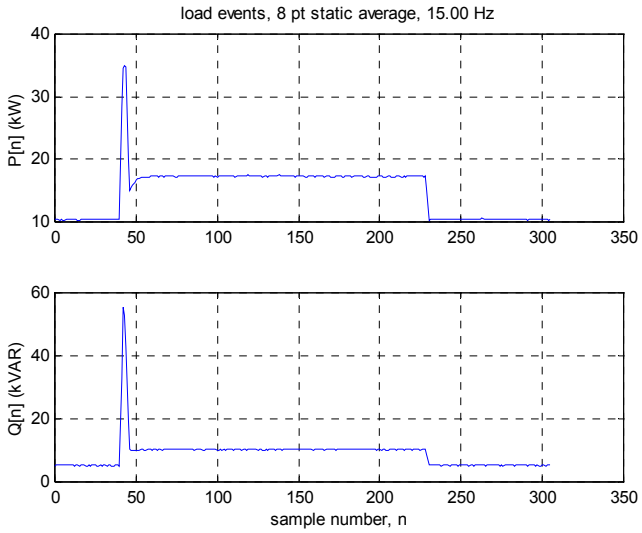


Figure 3-3 Turn-on and turn-off events of a 10-hp water pump.

Figure 3-4 shows the mean estimations at the start locations of the corresponding segments. The graph shows the characteristic of the detector, i.e. it detects more segments where there is more activity and less where there is less activity. It identified four distinct segments where there is an ON event. It also identified two segments where there is an OFF event. In between, the segments typically have a full length of 48 points, implying the detector found a single segment only for a given window.

The ON event is composed of four sections: *pre-event* (n = 37, length = 4), *plateau* (n = 42, length = 3), *transition* (n = 46, length = 7) and *steady state* (n = 54, length = 35). There can be many association strategies to identify an ON event from this type of assortments. The simplest is *idling*: after identifying a sharp step increase, the associator waits a certain amount of time until it declares a steady state. The idling time can be different when monitored loads are of different size. The nominal idling time for a load (or load group) can be recorded in a load database. Because a load with a longer turn-on

response time tends to have a higher plateau step-height, an appropriate idling time can be selected from the database based on the step-height measurement.

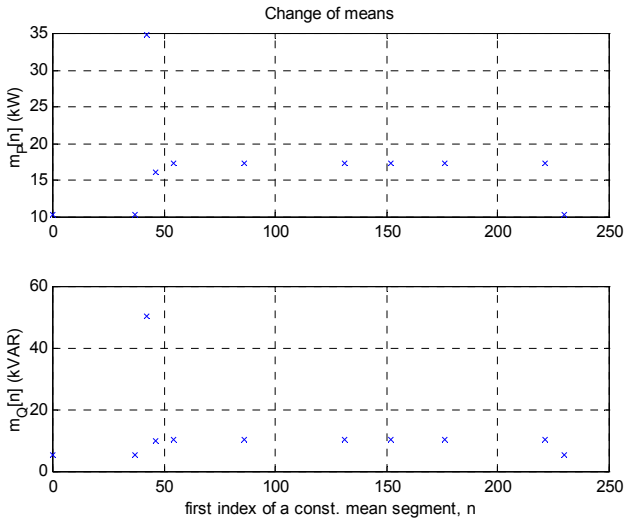


Figure 3-4 Change-of-mean segments of the signal in Figure 3-3.

However, the aforementioned idling strategy does not fully utilize the information provided by the change-of-mean detector. For example, the associator may actively pursue a series of segments, starting with a large mean and moderate variance plateau segment, followed by a reduced mean and increased variance transition segment, and anticipating a steady-state segment with a substantially reduced variance. On the other hand, the associator may perform a crude pattern recognition using the information and identify, at least, the type or group of a load.

The idling strategy will be used in this chapter, due to its simplicity. Because it takes several segments to complete an ON event, an associator flag is set after an initial step change is observed and until a steady state is reached. Once the associator decides that a steady state has been reached, it computes the step heights between the steady-state segment and the pre-event segment and generates a steady state event. The flag is released.

The OFF event which occurred at $n = 230$ can be easily associated with the sudden drop of the average value (of P) compared to the previous segment. Thus an OFF event is generated when the average of a current segment is reduced by more than a preset threshold, compared to the average of a previous segment. The threshold can be selected easily as a fraction of the minimum of the steady-state powers recorded in the load database. An OFF event is generated only when the associator flag is not set, because a large negative step change can occur as a during a turn-on response.

The event associator analyzes the outputs from the change-of-mean/variance detector, locates steady-state events and extracts relevant information for the event classifier. It generates messages with the index, direction (up/down), step heights, segment lengths before and after the event and total P value after the event. Table 3-1 shows the steady-

state events identified by the associator for the constant mean/variance segments of Figure 3-4.

Table 3-1 Steady-state events generated from the segments of Figure 3-4.

index	direction	ΔP (kW)	ΔQ (kVAR)	length-	length+	P (kW)
54	up	6.8624	5.0133	4	35	17.287
230	down	-6.9047	-5.0128	8	39	10.410

3.4 Steady-state event classifier

The steady-state events created by the event associator need to be examined by the event classifier to determine whether they belong to certain load classes. The event classification requires first the construction of load classes and then the design of the classification procedure or criterion itself.

In the case of a residential load monitoring, it usually is difficult to collect and analyze training data manually. Thus, it is a common practice to program the system to collect observations for a certain period and group them as *clusters*. After the training period, an event is classified against these load clusters. The clusters contain all the necessary information to classify an event, but they are not associated with the physical loads. The actual load identification occurs when the load clusters are matched against the *load library* [Drenker and Kader, 1999]. In this approach, the load cluster acts as an intermediate load class, with its name unknown until it is connected to the load library.

For commercial buildings, it is generally possible and makes sense to collect training data first and compile *a priori* information to construct the load classes. From the training data, pertinent information, such as the mean and variance of the steady-state power of a load, can be extracted to constitute a load class along with its name and nominal rating.

In this chapter, we assume *a priori* knowledge of the loads monitored and the existence of the load classes, and concentrate on the development of a classification method. Depending on the circumstances, the load classes will be created explicitly from a set of separate training data or implicitly via the clustering approach.

Suppose $\Delta\mathbf{x}$ is a steady-state event vector. Its components are the step changes in the harmonic powers and its length is equal to the dimension of the harmonic power space used for mapping ($\Delta\mathbf{x} = [\Delta P\ \Delta Q\ \ldots\ ]^T$). Consider the following multiple hypothesis test:

$$\begin{aligned} &H_1 : \Delta\mathbf{x} \text{ belongs to load } 1 \\ &H_2 : \Delta\mathbf{x} \text{ belongs to load } 2 \\ &\vdots \\ &H_m : \Delta\mathbf{x} \text{ belongs to load } m, \end{aligned} \tag{3.1}$$

where m is the number of loads being monitored. The direction of an event (ON or OFF) can be easily identified from the sign of ΔP.

The minimum distance classifier is widely used due to its simplicity. It computes the Euclidean distances from $\Delta\mathbf{x}$ to the class centers and selects a minimum. However, because each class has a different statistical distribution property, the distance is normalized by its variances. For example, on a 2-dimensional plane, the following normalized distance is typically used:

$$d(\Delta\mathbf{x}, \boldsymbol{\mu}_i) = \sqrt{\left(\frac{\Delta x_1 - \mu_{i,1}}{\sigma_{i,1}}\right)^2 + \left(\frac{\Delta x_2 - \mu_{i,2}}{\sigma_{i,2}}\right)^2}, \tag{3.2}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the mean and the covariance of class i, respectively. Furthermore, the event vector needs to be rotated first by an angle which the major axis of the ellipsis of the cluster i forms with the horizontal axis [Palomera-Arias and Norford, 2001]. The aforementioned procedure is essentially a *whitening* transform, which is composed of the Karhunen-Loève (KL) expansion and the normalization [Fukunaga, 1972].

However, the whitening transform is cumbersome in the sense that the eigenvalues and eigenvectors of a covariance matrix have to be found to construct the transformation matrix. Instead, the following simple conditional maximum-likelihood (ML) rule is proposed for (3.1):

$$\hat{H}_k = \arg\max_{H_i} p(\Delta\mathbf{x} \mid H_i). \tag{3.3}$$

The ML classifier uses the conditional PDF of $\Delta\mathbf{x}$ given H_i as the test statistic. To evaluate the conditional PDF, the probability distribution of $\Delta\mathbf{x}$ given H_i should be known. From the observations, a normal distribution is assumed, i.e.

$$p(\Delta\mathbf{x} \mid H_i) = \frac{1}{(2\pi)^{\frac{N}{2}} \det^{\frac{1}{2}}(\boldsymbol{\Lambda}_i)} \exp\left(-\frac{1}{2}(\Delta\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\Delta\mathbf{x} - \boldsymbol{\mu}_i)\right), \tag{3.4}$$

where N is the length of $\Delta\mathbf{x}$ and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$ are the mean vector and the covariance matrix of load class i, respectively. For simplicity, it is assumed that $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$ are deterministic and known at the time of classification.

To simplify the test, a new test statistic ξ_i is defined as

$$\begin{aligned} \xi_i &= -2 \ln p(\Delta\mathbf{x} \mid H_i) - N \ln 2\pi \\ &= (\Delta\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\Delta\mathbf{x} - \boldsymbol{\mu}_i) + \ln\left(\det(\boldsymbol{\Lambda}_i)\right). \end{aligned} \tag{3.5}$$

With the new test statistic, the hypothesis test simply looks for the minimum of ξ_i

$$\hat{H}_k = \arg\min_{H_i} \xi_i, \quad \xi_i < \gamma, \tag{3.6}$$

where γ is the threshold of the test. If there is no ξ_i that satisfies the threshold requirement, the event is unclassified.

The new test (3.6) is easy to implement and utilizes the covariance matrix as it is, unlike the minimum distance classifier that needs a whitening transform. Also, it can be used for any steady-state event vector with an arbitrary dimension, because the test is not bounded by the (planar) geometrical notion of a load class distribution.

One subtlety in implementing the test is the fact that the event vector $\Delta\mathbf{x}$ and the statistical property of a class ($\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$) should be compatible, i.e. they should be able to form a *probability space* from which the samples of $\Delta\mathbf{x}$ can be drawn.

To maintain statistical compatibility, the power vector $\mathbf{x}[n] = [P\ Q\ \ldots.\ ]^T$ is considered as a random process, which is indexed by the discrete-time n and is generated at a power coding rate. The probability space of load class i is defined at this coding rate and its random variable is the load vector $\mathbf{s}_i[n]$, i.e. the power vector when only load i is ON. As explained before, the load vector is modeled by a Gaussian (normal) distribution, i.e. $\mathbf{s}_i[n] \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$. Furthermore, we assume that two different time samples of a load vector are statistically independent, i.e. $\mathbf{s}_i[n] = \boldsymbol{\mu}_i + \mathbf{w}_i[n]$, where $\mathbf{w}_i[n]$ is zero-mean white Gaussian noise (WGN). Also, it is reasonable to assume that two load vectors from two different classes are independent from each other.

If k loads are on at time n, the power vector $\mathbf{x}[n]$ is simply the sum of those individual load vectors

$$\mathbf{x}[n] = \sum_{i=1}^{k} \mathbf{s}_i[n]. \tag{3.7}$$

Because any linear combination of independent Gaussian random vectors is also Gaussian [Papoulis and Pillai, 2002], $\mathbf{x}[n]$ is Gaussian with

$$\begin{aligned} &\text{mean: } \mathbf{m}_\mathbf{x} = E\big(\mathbf{x}[n]\big) = \sum_{i=1}^{k} E\big(\mathbf{s}_i[n]\big) = \sum_{i=1}^{k} \boldsymbol{\mu}_i \\ &\text{covariance: } \boldsymbol{\Lambda}_\mathbf{x} = \sum_{i=1}^{k} \boldsymbol{\Lambda}_i . \end{aligned} \tag{3.8}$$

Also, two different time samples of $\mathbf{x}[n]$ are statistically independent.

Suppose $\bar{\mathbf{x}}$ is an L point average of $\mathbf{x}[n]$ under a load status, i.e. $\bar{\mathbf{x}} = \frac{1}{L}\sum_{n=1}^{L}\mathbf{x}[n]$. Because different time samples of $\mathbf{x}[n]$ are statistically independent, it can be shown that its time average is also Gaussian, i.e. $\bar{\mathbf{x}} \sim N(\mathbf{m}_{\bar{\mathbf{x}}}, \mathbf{\Lambda}_{\bar{\mathbf{x}}})$, where $\mathbf{m}_{\bar{\mathbf{x}}} = \mathbf{m}_{\mathbf{x}}$ and $\mathbf{\Lambda}_{\bar{\mathbf{x}}} = \frac{1}{L}\mathbf{\Lambda}_{\mathbf{x}}$ [Papoulis and Pillai, 2002]. Thus an L-point average of $\mathbf{x}[n]$ has the same mean as $\mathbf{x}[n]$, but with a covariance matrix reduced by $1/L$.

Now, suppose that a steady-state event vector $\Delta\mathbf{x}$ is computed from segment j and segment $j-1$, as the difference of two segment-averages:

$$\Delta\mathbf{x} = \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{j-1}. \tag{3.9}$$

Each average is computed with L_j and $L_{j\text{-}1}$ samples, respectively. If k loads are on at segment $j-1$, $\mathbf{x}_{j-1}[n] \sim N\left(\mathbf{m}_{\mathbf{x}_{j-1}} = \sum_{i=1}^{k}\boldsymbol{\mu}_i\,,\ \mathbf{\Lambda}_{\mathbf{x}_{j-1}} = \sum_{i=1}^{k}\mathbf{\Lambda}_i\right)$. If load m ($m \notin \{1, 2, \ldots, k\}$) is on, at segment j, $\mathbf{x}_j[n] \sim N\left(\mathbf{m}_{\mathbf{x}_j} = \boldsymbol{\mu}_m + \sum_{i=1}^{k}\boldsymbol{\mu}_i\,,\ \mathbf{\Lambda}_{\mathbf{x}_j} = \mathbf{\Lambda}_m + \sum_{i=1}^{k}\mathbf{\Lambda}_i\right)$. Now it is straightforward to show that $\Delta\mathbf{x}$ is Gaussian with $\mathbf{m}_{\Delta\mathbf{x}} = \mathbf{m}_{\mathbf{x}_j} - \mathbf{m}_{\mathbf{x}_{j-1}} = \boldsymbol{\mu}_m$ and $\mathbf{\Lambda}_{\Delta\mathbf{x}} = \mathbf{\Lambda}_{\bar{\mathbf{x}}_j} + \mathbf{\Lambda}_{\bar{\mathbf{x}}_{j-1}} = \frac{1}{L_j}\mathbf{\Lambda}_{\mathbf{x}_j} + \frac{1}{L_{j-1}}\mathbf{\Lambda}_{\mathbf{x}_{j-1}}$.

When the power vectors of segments are d-point static averages of raw power vectors, the mean of $\Delta\mathbf{x}$ is unchanged, but its covariance matrix is changed due to the additional averaging. For an ON event of load m,

$$\Delta\mathbf{x} \sim N\left(\boldsymbol{\mu}_m\,,\ \tfrac{1}{dL_j}\mathbf{\Lambda}_{\mathbf{x}_j} + \tfrac{1}{dL_{j-1}}\mathbf{\Lambda}_{\mathbf{x}_{j-1}}\right). \tag{3.10}$$

For an OFF event,

$$\Delta\mathbf{x} \sim N\left(-\boldsymbol{\mu}_m\,,\ \tfrac{1}{dL_j}\mathbf{\Lambda}_{\mathbf{x}_j} + \tfrac{1}{dL_{j-1}}\mathbf{\Lambda}_{\mathbf{x}_{j-1}}\right). \tag{3.11}$$

In each case, j is the event segment and $j-1$ is the pre-event segment. $\mathbf{\Lambda}_{\mathbf{x}_{j-1}}$ can be computed by summing up the covariance matrices of the loads which are on. $\mathbf{\Lambda}_{\mathbf{x}_j}$ is obtained by adding $\mathbf{\Lambda}_m$ (ON event) to or subtracting it (OFF event) from $\mathbf{\Lambda}_{\mathbf{x}_{j-1}}$. It is obvious by now that it is the normal distribution of $\Delta\mathbf{x}$, not the distribution of a load, to be used for the hypothesis test. Thus when (3.5) is computed for each load class, the load covariance matrix needs to be replaced by the event covariance matrix (3.10 or 3.11). The advantage of the ML classifier against the minimum distance classifier is apparent here, because the latter has to perform the whitening transform for every event hypothesis due to the changing event covariance matrix determined by the nature of the signal.

The statistical properties of a load class are estimated from its samples. Because its true mean and covariance are unknown, they are usually approximated by the sample mean

and covariance. The quality of estimation increases with the increasing number of samples, if the random process is *ergodic* [Papoulis and Pillai, 2002].

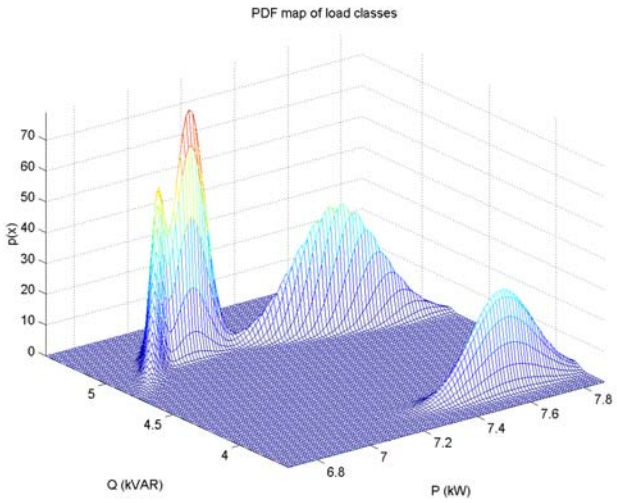


Figure 3-5 PDF map of four pump classes (in Figure 3-1).

In its original form (3.3), the hypothesis test simply looks for a load class that gives the largest conditional PDF assuming H_i is right. On a 2-dimensional power plane, with the third axis as the height of PDFs, the test can be conceived as a search for a hypothesis which gives the highest elevation for a given event point. Figure 3-5 shows the PDF map of four pump classes whose mean and covariances are obtained from sample means and covariances. However, it should be remembered that an event covariance matrix replaces the load covariance matrix at the time of test.

Two events generated in the previous section are the ON and OFF events of the CHWP2. The event messages are fed to the event classifier and the classification results are shown in Table 3-2.

Table 3-2 Classification results for the steady-state events of Table 3-1.

index	direction	ΔP (kW)	ΔQ (kVAR)	load name	ξ_i
54	on	6.8624	5.0133	CHWP2	66.08
230	off	-6.9047	-5.0128	CHWP2	40.25

The classifier successfully associated the events with the correct load class, with reasonable test statistic values. The classifier also updates the load status upon a successful classification. The load status is simply a vector with Boolean entries (1 for ON, 0 for OFF). If there are k loads being monitored, the load status vector $\mathbf{s}$ is represented by $\mathbf{s} = [s_1\ s_2\ \ldots\ s_k]^T, s_i \in \{0,1\}$.

In this section, a simple ML rule is developed as the event classifier and the statistical distribution of an event vector is specified. The resulting event classifier is easy to use and can be applied to any series of power vectors with an arbitrary dimension and time-

averaging. Upon the successful classification of a steady-state event, the classifier updates the load status.

3.5 Test of steady-state load classifier

The overall steady-state load classifier was assembled with the components developed in the previous sections and tested with a few off-line data sets. Figure 3-6 shows two sets of the CTWP1 ON and OFF events used to test the steady-state load classifier. The figure shows the real and reactive power after the 8-point static averaging. Due to the activities of the background load, the waveforms are highly irregular, suggesting the difficulty of the step-height estimation. The classifier succeeded in the second ON-OFF event set, but misclassified the first set as the CTWP2 events. The misclassification of the first OFF event is inevitable, because only the CTWP2 is assumed ON at the time of the search. Table 3-3 shows the classification results.

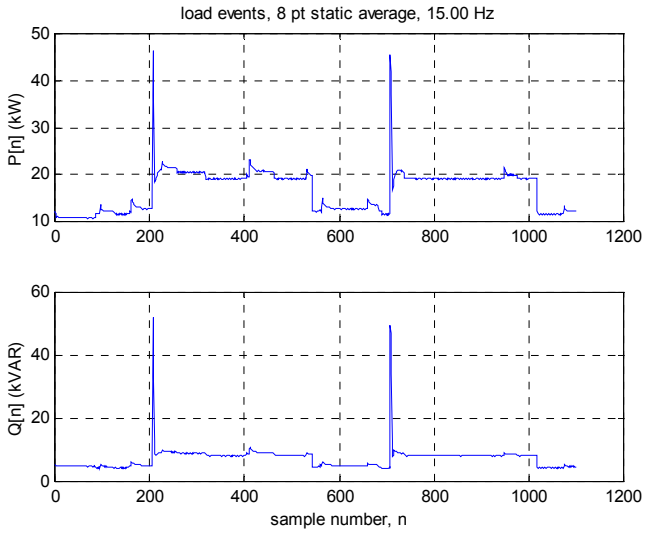


Figure 3-6 Two CTWP1 ON/OFF event sets.

Table 3-3 Classification results for the steady-state events in Figure 3-6.

index	direction	ΔP (kW)	ΔQ (kVAR)	load name	ξ_i
227	on	9.203	4.658	CTWP2	7.928×10^3
544	off	-7.785	-4.025	CTWP2	2.804×10^3
738	on	7.572	3.941	CTWP1	-1.528
1019	off	-7.575	-3.910	CTWP1	1.199

A closer examination of the figure reveals that a small background load is turned on at n = 226, before the CTWP1 reaches its steady state, whereas the second ON event is free of the background activities for the most of its steady state. Consequently, the first P step-height is overestimated by 1.5 kW (step-height of the background event), because the event associator is programmed to wait approximately two seconds after detecting a *pump-like* initial step change.

One may argue that by allowing an ample waiting time for the event associator, the problem of background activity may be solved. However, even with an extended idling time, there is no guarantee that the point at which the associator assumes a steady state is not subject to a background activity. A longer waiting time simply gives more chances to the background activities and thus the step-height measurements are prone to more errors. On the other hand, if the waiting time is too short, a steady state is declared when the turn-on response of a load is still in its transition portion, whose mean is different from the one of the steady state and usually is not repeatable. Either approach (shorter or longer waiting time) does not solve the problem of measuring the step-heights under active background load activities.

The difficulties in defining and detecting a steady state and in classifying it against proximal load classes are already discussed. These difficulties are partially addressed by designing the general change-of-mean/variance detector, event associator and event classifier. The additional difficulty is introduced by the background activity, which can interfere with the estimation of the step-heights. The steady-state load classifier alone can not solve the problem of background load activity. It needs help from its sibling– a *transient pattern recognizer*.

3.6 Load Classification based on transient pattern recognition

A load classifier based on the state space mapping was developed in the previous sections. By utilizing the fact that constant loads occupy distinctive locations in the harmonic power space, the classifier could classify ON/OFF events with a reasonable accuracy. However, the classifier performs poorly when the load classes are adjacent to each other in the power space and when there are many background load activities.

From this section, the transient response of a load is treated as a source of features and classified, rather than a nuisance in defining and detecting a steady state. Two classifiers work in parallel, utilizing two different aspects of a load response- the *transient* and the *steady state*- and thus making a better use of the available information that can be extracted from the harmonic power signal. A certain amount of coordination will be necessary to place two different classifiers simultaneously at work.

A transient response is observed when a load moves to and from its steady-state. Most loads, such as single-speed motors, incandescent bulbs and florescent lamps, show characteristic transients [Leeb, 1993; Leeb *et al.*, 1995]. However, transient patterns are not genuinely unique and some loads can have similar transient patterns, especially when they have similar ratings. The patterns can also have their own statistical properties.

Thus, the transient pattern recognizer is basically another classifier and is subject to the same limitations as those of the steady-state load classifier, i.e. the overlaps in time and space and the background load activities. However, it at least brings in a redundancy because it utilizes another aspect of a load event overlooked by the steady-state load

classifier. On the other hand, it is expected to perform better in detecting and classifying a load event, because the transient patterns have a more discernibility than the steady state powers (the steady state mapping is performed with only two or three points whereas the pattern recognition is accomplished with tens to hundreds data points). Also it will be less vulnerable to the background activities, because the pattern width is typically less than the waiting time of the event associator and their effects are lessened due to the higher power levels during a transient (higher SNR).

Figure 3-7 shows the transient responses of four 10-hp water pumps installed at 160 Sansome St. The ON transient shape of a single-speed motor typically consists of a tranquil pre-event segment, plateau (a relatively flat portion after the initial step increase), transition and steady state. Although the whole transient response shape can be used for pattern recognition, it will be wiser to utilize only the plateau and the beginning part of the transition both to minimize the influence of the background load activities and to reduce the computational burden.

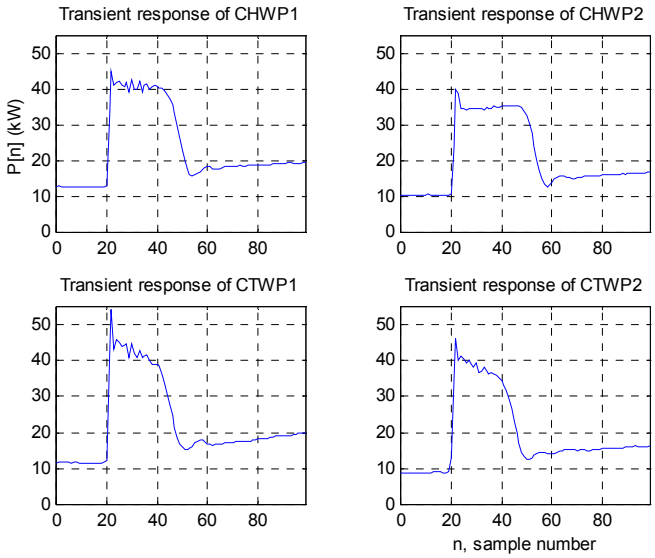


Figure 3-7. Transient response (ON) shapes of four 10-hp pumps.

A close examination of Figure 3-7 shows that the response shapes are generally similar but differ in detail. However, the CTWP1 and the CTWP2 have largely the same response shape, although the CTWP1 has a slightly higher plateau. The pattern recognizer may have a difficulty in differentiating them. However their steady-state clusters were, at least, separate (Figure 3-1). Thus it demonstrates the advantage of having both the pattern recognizer and the steady-state load classifier to complement each other.

The pattern recognition is essentially composed of two parts: *detection* of a pattern and its *classification*. In order to recognize a pattern, it first needs to be detected. Considering the fact that most transient responses of the loads are accompanied by the power level changes, a change-of-mean detection scheme, such as the cumulative sum (CUSUM) algorithm, can be used to detect a transient pattern [Basseville and Nikiforov, 1993; Leeb

et al., 1995]. Sometimes, a transient response (or a section of it) shows changes in its variance only, not accompanied by a power level change. A type of change-of-variance detector is necessary to capture this sort of transient. One approach is to compute a finite-impulse-response (FIR) filtered version of a signal and threshold the difference between the FIR filtered version and the original signal [Shaw, 2000]. This approach works well when there are sharp changes, but performs poorly when there are moderate changes and oscillations with relatively small magnitudes.

However, a very versatile change-of-mean/variance detector was already developed for the steady-state load classification and thus it can also be used to detect the transient patterns. This approach also makes sense in the viewpoint that because two classifiers work in parallel, tapping the same power signal, they need to be triggered by a single event generator. Thus the pattern classifier will receive its input (raw pattern) from a modified event associator that will prepare the outputs for both the steady-state classifier and the pattern classifier once an event is detected.

Once a transient pattern is detected and extracted, it is the job of the transient pattern classifier to compare the pattern against the exemplars of known transient responses. Each exemplar may have a different signal representation and the observed pattern may have to be decomposed accordingly for each exemplar to be compared. Each comparison produces a matching gain and in the end the classifier selects a load class that offers the most favorable gain. A pattern or exemplar can be composed of several sections. In this case, matching gains are computed for each section and the overall single gain is computed from the section gains.

However, when two classifiers work in parallel, there is no guarantee that two will produce the same result. Thus an *arbitrator* is necessary to determine a load event when they come up with contradicting results. Based on the outcomes of two separate classifiers, three situations are possible:

1) Both classifiers recommend the same load class and at least one of them has an acceptable classification gain: This is the most wanted case and the event is easily recognized with a high level of confidence.
2) Each classifier comes up with a different load class, but one classifier has a more favorable gain to the other: The load class with the more favorable gain is selected for the event. Obviously, the level of confidence is low.
3) Neither classifier comes up with an acceptable gain: This implies that both classifications are unreliable. The event is unclassified.

Figure 3-8 shows the block diagram of the overall load disaggregator with the steady-state load classifier and the transient pattern classifier placed in parallel.

There is also the question of how many channels should be utilized for the pattern recognition, i.e. whether to use only the real power or use the reactive and other harmonic powers. Initial study suggests that the real power (P) channel alone is enough for most pattern recognition purposes. If needs arise, other channels can be searched for useful

patterns. The patterns from other channels are considered as additional sections and the overall classification gain is computed as usual.

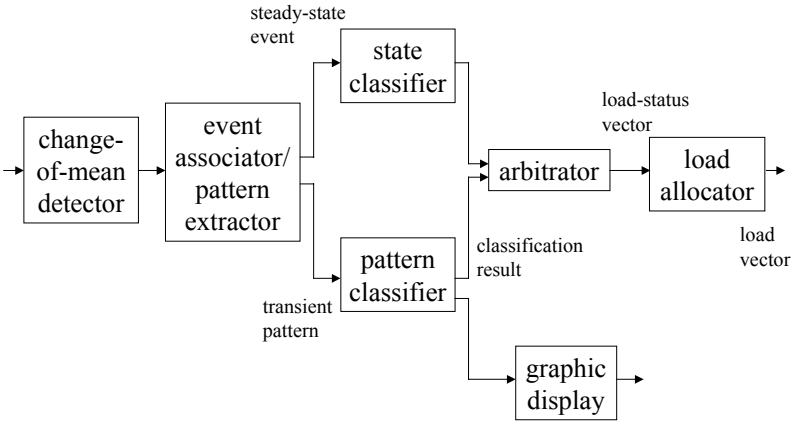


Figure 3-8 Block diagram of the load disaggregator with two classifiers.

There basically are two types of events for the purpose of the load classification- ON and OFF. An OFF event, in most cases, is a simple step-like downward transition with no or a few transition points between two different power levels. Thus the information extractable from an OFF pattern is merely its step-height.

On the other hand, the ON events are typically accompanied by the distinctive transient response shapes and thus their identification can be greatly benefited from the transient pattern recognition. In this chapter, the transient pattern recognizer is employed to identify the ON events only. The OFF events are solely identified by the steady-state load classifier. If needs arise, the pattern recognizer can be tapped later to help identifying the OFF events.

3.7 Transient pattern classifier

The transient pattern recognizer is essentially another type of classifier. It basically classifies the observed transient pattern by matching it against the known set of transient exemplars (classes). An exemplar with a best match is selected and the observed pattern is associated with its load.

To recognize a pattern, it first needs to be represented in a suitable format and then tested according to a classification rule. When a transient pattern is very long, it will be more economical to extract a few distinctive *sections* than to save the whole length of the signal.

A section needs to be represented by a basis, which can emphasize distinctive and repeatable features and discourage non-characteristic and non-repeatable features of the signal. Many types of basis are possible, such as time, wavelet, Fourier, etc. Nevertheless, whatever basis is used, once the signal is decomposed, the result is another finite length vector (or a set of vectors) in $L^2(\mathbb{Z})$ space, spanned by the selected basis. Thus, the way a pattern is represented does not affect the core design of a classifier, although provisions have to be made to accommodate different weighting strategies and the possibility of subsections depending on the basis choice.

Once the observed transient pattern is represented in the same manner as an exemplar to be matched against, its membership to the exemplar needs to be tested by a classification rule, or classifier. There can be various types of classifiers. Simple classifiers use Euclidean distances between the pattern and exemplars, or normalized inner products of the pattern and exemplars [Leeb, 1993]. A general purpose pattern classifier based on the method of least squares was developed and tested by Shaw [2000].

If the statistical properties of the transient responses are known completely, a Bayesian or ML type classifier can be designed. However, estimating the statistical property of a transient pattern poses a problem. Even if the probabilistic distribution of a pattern is assumed Gaussian, its mean and covariance need to be estimated to compute its PDF. The mean of a normal distribution can be estimated reliably with tens of samples. Nevertheless, the covariance estimation needs at least hundreds of samples. A few transient responses may be obtained during a training period. But obtaining one hundred responses is practically impossible.

Thus, the classifier is likely to have to make decisions in an environment where there exist a very crude estimation of the mean, obtained from one or a few samples, and virtually no information about the covariance of each class. In this situation, designing a classifier based on the method of least squares (LS) appears to be a viable strategy, because it does not require any probabilistic assumptions about the classes [Kay, 1993]. The advantage of the LS classifier is that it can be used in an environment where the additive noise is non-Gaussian or changes with time. The down side is that no optimality can be claimed because of the lack of the statistical structure.

Suppose that **y** represents a data vector (observed pattern) and **x** an exemplar vector. We want to quantify (or estimate) the amount of correlation between **y** and **x**, so that a decision can be made whether **y** belongs to **x**. We assume that **y** can be linearly approximated by **x**, with an offset term to account for a possible difference in the DC level. Any difference in the approximation is attributed to an unknown additive noise vector **w**.

$$\mathbf{y} = \alpha\mathbf{x} + \beta\mathbf{1} + \mathbf{w} = \mathbf{A}\mathbf{a} + \mathbf{w}, \tag{3.12}$$

where $\mathbf{A} = [\mathbf{x} \;\; \mathbf{1}]$ and $\mathbf{a} = [\alpha \;\; \beta]^T$. α is the correlation (matching) gain between the data and the exemplar. If **x** and **y** are identical, $\alpha = 1$ (and $\mathbf{w} = \mathbf{0}$). As **x** and **y** are uncorrelated,

α diverges from the unity. β is the offset term. We want to minimize the second norm of the error vector, by an optimal choice of the parameter vector **a**, i.e.

$$\text{minimize: } \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{A}\mathbf{a}\|^2. \tag{3.13}$$

This is a standard least squares problem. The solution is given by the following matrix formula:

$$\hat{\mathbf{a}} = [\hat{\alpha} \quad \hat{\beta}]^T = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}. \tag{3.14}$$

$\hat{\alpha}$ is the LS estimation of the correlation gain between **y** and **x**.

We define the residue of estimation *r* as the square of the minimum second norm of the error vector. It often is necessary to normalize the residue by the length of vectors. If both **x** and **y** are of length *M*, the normalized residue of estimation r_n is defined by

$$r_n = \tfrac{1}{M} r = \tfrac{1}{M} \|\mathbf{y} - \mathbf{A}\hat{\mathbf{a}}\|^2. \tag{3.15}$$

However, the above discussion does not include the fact that there can be a time offset (shift) between **y** and **x**, in the case of a time-based signal representation. Thus the best shift also has to be found as a part of the minimization process. Let **y** is of length *N* and **x** of length *M*. For simplicity, we assume $N \geq M$. The original minimization of a least squares error criterion now becomes a double minimization problem.

$$(\hat{\mathbf{a}}, \hat{i}) = \underset{a,\ 0 \leq i \leq N-M}{\arg\min} \|\mathbf{y}_i - \mathbf{A}\mathbf{a}\|^2, \tag{3.16}$$

where $\mathbf{y}_i = [y[i]\ y[i+1] \ldots y[i+M-1]]^T$. $\hat{\mathbf{a}}$ and $\hat{i}$ are the optimal estimations of the parameter vector and the shift, respectively. In implementation, for each *i* the parameter vector is estimated and the resulting residue is computed, then $\hat{i}$ and $\hat{\mathbf{a}}$ are determined by finding the minimum of the residues. Also, note that $(\mathbf{A}^T\mathbf{A})^{-1}$ needs to be computed only once because **x** is not shifted and thus at each *i* only $\mathbf{A}^T\mathbf{y}_i$ needs to be obtained.

When there exist multiple sections for a pair of pattern and exemplar, the LS estimation of the correlation gain and the resulting normalized residue is computed for each section. Suppose there are *k* sections. We also introduce a weighting factor w_j (0 – 1) for section *j*. Now we solve a least squares problem again to obtain the overall correlation gain *a*.

$$\begin{aligned} a\mathbf{H} &\approx \mathbf{b} \\ a &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{b}, \end{aligned} \tag{3.17}$$

where $\mathbf{H} = [w_1\ w_2 \ldots w_k]^T$ and $\mathbf{b} = [\alpha_1 w_1\ \alpha_2 w_2 \ldots \alpha_k w_k]^T$. α_j is the correlation gain of section *j*. The normalized residue of the overall estimation γ is also defined as:

$$\gamma = \frac{1}{\sum_{j=1}^{k} M_j} \sum_{j=1}^{k} r_j = \frac{1}{\sum_{j=1}^{k} M_j} \sum_{j=1}^{k} M_j r_{n_j} , \tag{3.18}$$

where M_j, r_j and r_{n_j} are the length, residue and normalized residue of section j, respectively.

In this manner, the correlation gain a and the normalized residue γ can be computed for each exemplar given an observed pattern vector **y**. Now it is time to determine whether **y** represents a load event and if so which load (exemplar). Again, this is a multiple hypothesis test problem.

$$\begin{aligned} &H_1 : \mathbf{y} \text{ belongs to load } 1 \, (\mathbf{x}_1) \\ &H_2 : \mathbf{y} \text{ belongs to load } 2 \, (\mathbf{x}_2) \\ &\vdots \\ &H_m : \mathbf{y} \text{ belongs to load } m \, (\mathbf{x}_m), \end{aligned} \tag{3.19}$$

where m is the number of loads being monitored. There can be a few test criteria, but the simplest is to choose the exemplar that gives a minimum normalized residue, i.e.

$$\hat{H}_k = \arg\min_{H_i} \gamma_i \qquad \gamma_i < \gamma_{thres} , \tag{3.20}$$

where γ_i is the normalized residue of exemplar $\mathbf{x}_i$ and γ_{thres} is the threshold of the test. This rule is appealing in the sense that it is a natural extension of the LS error criterion. The *minimum normalized-residue rule* appears to work best in most circumstances. However, it was observed to perform poorly when the (Gaussian) noise magnitude is large. It can be explained by the fact that as the noise magnitude increases (lowers SNR) the normalized residues also increase in general and are affected more by the random fluctuations of the noise and less by the differences between the true signal of an observation and exemplars.

Another measure which can be used for classification is the correlation gain a. Because the gain can be any real number, it first needs to be normalized before comparison. The gain a is normalized by the following normalization function, $f(a)$

$$f(a) = a_n = \begin{cases} 0, & \text{if } a \le 0 \\ a, & \text{if } 0 < a \le 1 \\ 1/a, & \text{if } a > 1, \end{cases} \tag{3.21}$$

where a_n is the normalized gain, which has a range of 0 to 1.

With the normalized gain for each exemplar, it is tempting to set up a classification rule that selects a hypothesis that yields a maximum normalized gain. However, the normalized gain classification rule performs poorly, because the gain is simply the slope of the first order approximation and lacks any measure of the scatter of the data from the approximation (e.g. $\mathbf{y} = [0\ 0\ 0\ 0\ 10\ 10]^T$ and $\mathbf{x} = [0\ 0\ 5\ 5\ 5\ 5]^T$ gives $a = 1$, though two vectors are radically different.).

Thus, it is necessary to take into account both the normalized gain and the normalized residue in selecting a load class. One solution is to rescale the normalized residue and multiply it with the normalized gain to obtain a composite gain ξ_i:

$$\xi_i = a_{n_i}\left(1 - \frac{1 - a_{thres}}{\gamma_{thres}}\gamma_i\right), \tag{3.22}$$

where $a_{thres} < a_{n_i} \le 1$ and $0 \le \gamma_i < \gamma_{thres}$. Now the *maximum composite-gain rule* simply looks for an exemplar (load) that gives a maximum composite gain:

$$\hat{H}_k = \arg\max_{H_i} \xi_i . \tag{3.23}$$

The maximum composite-gain rule works better than the minimum normalized-residue rule in an environment where Gaussian noises have large magnitudes.

When the transient pattern classifier and the steady state-load classifier produce different results, an arbitration is necessary. The only case when a true arbitration is necessary is when both classifiers have acceptable gains but differ in terms of the declared load classes. Though there can be many approaches, we design a simple arbitration rule based on the rescaling of the gain ξ of the steady-state load classifier.

$$\xi_n = \begin{cases} 1 & \xi \le 0 \\ 1 - \dfrac{1 - a_{thres}}{\xi_{thres}}\xi & 0 < \xi < \xi_{thres} \end{cases}, \tag{3.24}$$

where ξ_n is the normalized (rescaled) gain of ξ and ξ_{thres} is the threshold of ξ. a_{thres} is the threshold for the normalized gain a_n of the transient pattern classifier. Note ξ_n is normalized to the scale of a_n. Consider the following binary hypothesis test and its empirical test rule:

H_0: The output of the steady-state classifier is true.
H_1: The output of the transient pattern classifier is true.

$$a_n + \varepsilon \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \xi_n , \tag{3.25}$$

where ε is the weight of the test. If $\varepsilon > 0$, it favors the transient pattern classifier and, otherwise, the steady-state classifier. Because the pattern matching in general produces better results the weight is usually positive.

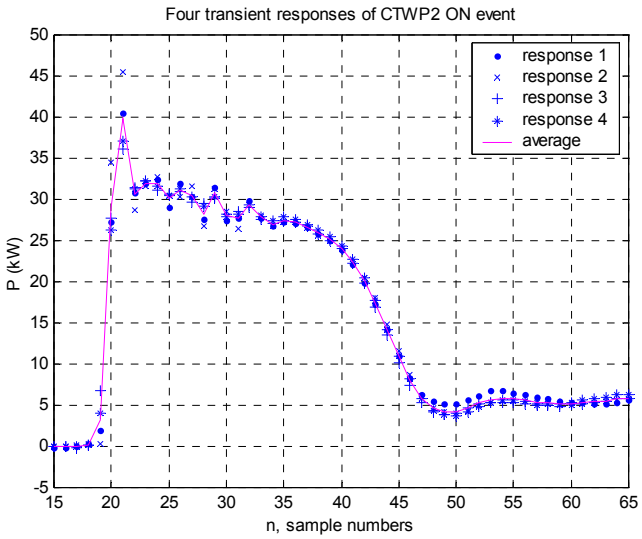


Figure 3-9 Four transient responses of CTWP2 ON event and their average.

As evident from the development of the classifier based on the least squares error criterion, it performs better when the exemplar of each class is close to the true mean of the transient response. Thus, when there are multiple samples available, it is better to average them, rather than to take a single sample as an exemplar. The *exemplar averaging* enhances the performance of the pattern classifier especially when there are similar loads. Figure 3-9 shows the average of four responses of the CTWP2 ON event.

In this section, the issue of representing a signal for recognition was discussed and a pattern classifier based on the least squared error criterion was designed. The classifier can handle a pattern with multiple sections, and does not require the statistical characterization of load classes but a single exemplar for each class. Selection rules were developed for the classifier and the arbitrator.

3.8 Test of transient pattern classifier

The transient pattern recognizer is assembled and placed in parallel to the steady-state load classifier. The overall load disaggregator with the two parallel classifiers was tested in a variety conditions.

The two sets of CTWP1 ON/OFF events (Figure 3-6) are used again to test the new classification scheme. We remember that the first ON event was misclassified as the CTWP2. The minimum normalized-residue rule is used to select the best matching

exemplar and the following test parameters are used: $\xi_{thres} = 1.5\times10^4$, $a_{thres} = 0.6$, $\gamma_{thres} = 40$ and $\varepsilon = 0.06$. Table 3-4 shows the result of ON event classifications.

Table 3-4 Transient pattern classification results for the ON events in Figure 3-6.

index	load name	no. of sections	a	γ
207	CTWP1	1	0.9786	0.8315
709	CTWP1	1	0.9924	0.2231

We observe that the first ON event is successfully matched to the CTWP1 ON by the pattern classifier and that the arbitrator will rule in favor of the pattern classifier. Figure 3-10 shows the pattern matching result. In the picture, solid line denotes the observed pattern and '×'s represent the matched exemplar. For the second ON event, both classifiers agreed that it belongs to the CTWP1. Thus the load disaggregator, with the help of the transient pattern classifier, correctly identified both ON events. During the test, for a variety of 34 ON events, the load disaggregator successfully classified all the events (0 % misclassification rate).

As discussed before, one advantage of the LS estimator is that it can be used in an uncharacterized noise environment. We add a WGN with the standard deviation $\sigma = 3$ (kW) to the CTWP1 waveforms and test how the pattern matcher performs. We simulate an unknown change in noise characteristic and thus the load database for the steady-state load classifier is not updated The maximum composite-gain rule is used to select the best matching exemplar and the following test parameters are used: $\xi_{thres} = 1.5\times10^4$, $a_{thres} = 0.6$, $\gamma_{thres} = 60$ and $\varepsilon = 0.06$.

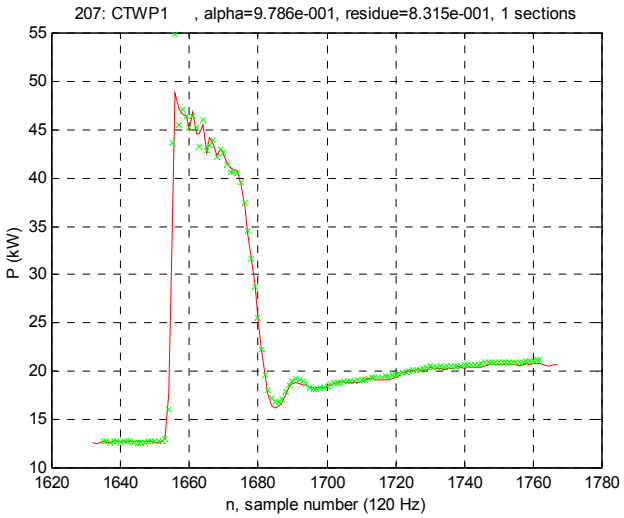


Figure 3-10 Matching result of CTWP1 ON event.

Table 3-5 shows the result of the steady-state load classification and Table 3-6 shows the result of the transient pattern classification. Due to the increased noise, the steady-state classifier misclassified both ON events. However, the transient pattern classifier correctly

classified the events with favorable gains so that the arbitrator can decide that those belong to the CTWP1.

Table 3-5 Steady-state load classification results for the waveforms in Figure 3-6 with a WGN (σ= 3 (kW)).

index	direction	ΔP (kW)	ΔQ (kVAR)	load name	ξ_i
221	on	9.542	4.512	CTWP2	2.244×10^4
544	off	-7.828	-4.600	CTWP1	1.875×10^3
716	on	8.965	4.592	CTWP2	1.038×10^4
1019	off	-6.654	-3.877	CTWP1	4.953×10^3

Table 3-6 Transient pattern classification results for the ON events in Figure 3-6 with a WGN (σ= 3 (kW)).

index	load name	no. of sections	a	γ
207	CTWP1	1	0.9631	7.940
708	CTWP1	1	1.033	11.49

Figure 3-11 shows the matching result of the first ON event. Thus, we verify that the LS transient pattern matcher performs very well even under a severe noise environment. Due to the pattern matcher, the load disaggregator becomes more robust to noise. For a variety of 34 ON events corrupted with the WGN, the load disaggregator misclassified only one event (2.94 % misclassification rate).

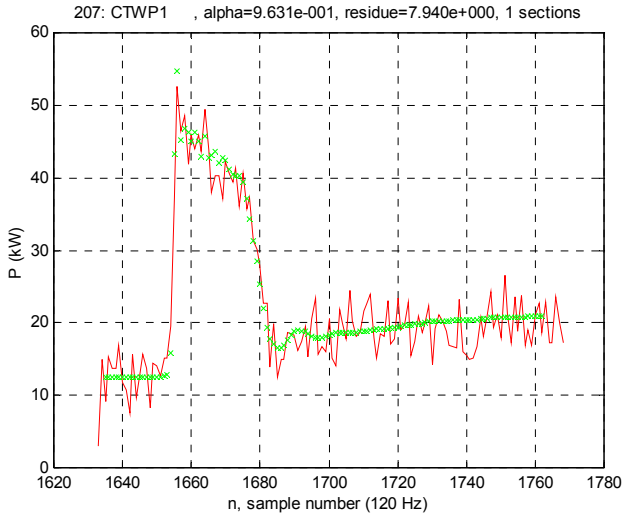


Figure 3-11 Matching result when the pattern is corrupted by a WGN (σ= 3 kW).

It often is the case that the noise added to a power signal is not Gaussian but has a certain bandpass characteristic. An unbalanced rotating machine can generate a harmonic noise. The harmonic (or periodic) noise can usually be estimated and thus can be filtered. However, in some cases, the estimation (or filtering) is not feasible, and the pattern matcher has to work with corrupted signal.

For the test, a set of CTWP1 events is corrupted by a harmonic noise, obtained from the 160 Sansome St. building, and an additional WGN with σ = 1 kW. For the test, the minimum normalized-residue rule is used to select the best matching exemplar and the following test parameters are used: $\xi_{thres} = 1.5\times10^4$, $a_{thres} = 0.6$, $\gamma_{thres} = 60$ and $\varepsilon = 0.06$

Table 3-7 Steady-state load classification results for a CTWP1 event set corrupted with harmonic and white Gaussian (σ= 1 kW) noise.

index	direction	ΔP (kW)	ΔQ (kVAR)	load name	ξ_i
86	on	8.159	4.414	CTWP1	1.855×10^3
2025	off	-7.906	-3.920	CTWP1	7.117×10^1

Table 3-8 Transient pattern classification results for a CTWP1 event set corrupted with harmonic and white Gaussian (σ= 1 kW) noise.

index	load name	no. of sections	a	γ
43	CTWP2	1	1.201	6.484

Table 3-7 and 3-8 show the classification results. The steady-state load classifier correctly identified the ON event, whereas the pattern matcher misclassified it to that of the CTWP2. Figure 3-12 shows the matching result, which shows that one of the noise peaks coincided with the beginning of the transient response. However, because the steady-state classification had a more favorable gain, the arbitrator could correctly determine that it was a CTWP1 event.

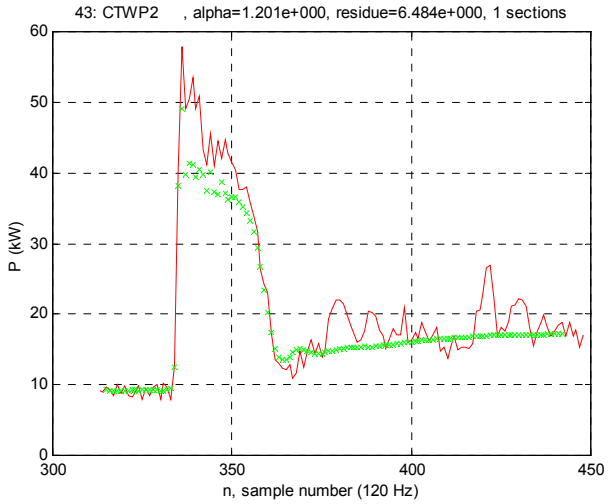


Figure 3-12 Matching result for a CTWP1 ON event corrupted by harmonic and white Gaussian (σ= 1 kW) noise.

This is an example where the transient pattern classifier, which usually performs better, makes a mistake and is corrected by the steady-state classifier. Unpredicted noise, spikes and fluctuations can jeopardize the performance of the pattern classifier. The steady-state load classifier is, in general, less affected by the temporal variation of the signal and adds a redundancy and confidence in such a situation. For a variety of 34 ON events corrupted

with the harmonic noise and WGN, the load disaggregator misclassified three events (8.82 % misclassification rate).

The pattern matcher works with an almost zero misclassification rate in a designed environment, and can also be used in a highly noisy environment (both Gaussian and non-Gaussian) although misclassification rate increases slightly. The development of the transient pattern classifier based on the least squares error criterion and its parallel placement with the steady-state load classifier have been very successful.

3.9 Chapter summary

A rich amount of information can be extracted from the non-intrusive power signal. The power signal carries the fingerprint of the operational status of a load in terms of its steady state power consumption and its transient response. By systemically exploiting those two properties of a load, its ON or OFF event can be identified and thus its operational status can be obtained. The knowledge of the statuses of the loads make it possible to disaggregate the total power observed into individual powers, in real-time.

The method of steady-state load classification utilizes the fact that a constant load consumes a fixed amount of power in its steady state and that its steady state power is distinctive in the harmonic power space. In this chapter, a simple steady-state load classifier based on the ON/OFF two-state load model was developed. The power consumed by a constant load was modeled by a Gaussian random process. A simple ML classification rule was introduced and the probability space of a load event was derived. The steady-state classifier generally works well, but can perform poorly when there exist proximal load classes and when there are many background load activities.

A transient pattern classifier based on least-squares error criterion was introduced to augment the task of load identification. The method of transient pattern recognition utilizes the source of information ignored by the method of steady-state load classification and is less prone to the errors caused by the background load activities. The transient pattern recognition can also have the problem of proximal load classes when the loads have similar ratings, but is less affected because the transient pattern of a load has a more discernibility than its steady state. An arbitration rule was developed for the case when two classifiers produce contradicting results.

The fact that electric loads are monitored from a single location, not individually, complicates the job of identifying their operational status the non-intrusive power signal. By utilizing the information offered by a load both at its steady state and transient state, the task of keeping track of its operational status becomes more amenable and realistic. The steady-state load classifier and the transient pattern classifier complement each other for the common purpose of load identification.

Chapter 4

Identification of Temporal Overlap of Electric Load Events via Event Space Search and Transient Pattern Composition

This chapter describes the general method of identifying constant electric loads when their ON and/or OFF events overlap in time. The identification of a single constant-load event was fully discussed in the previous chapter. The preceding chapter developed the methods of steady-state load classification and transient pattern recognition, to better utilize two different aspects of a load– steady-state response and transient response– for the purpose of its identification. This chapter further develops the idea to identify overlapping load events by the construction and search of the event space as a combination of individual events and the composition and match of transient patterns.

4.1 Introduction

Because only a single set of voltage and current sensors is used to monitor tens of different individual electric loads, the non-intrusive power signal is the (linear) superposition of individual power signals. It is undoubtedly a challenging task to identify the ON/OFF status of the loads from the observation of non-intrusive power signal alone. Fortunately, many constant electric loads have unique power consumption levels at their steady state and furthermore have distinctive transient responses. Exploiting these *fingerprints* of loads in a systematic manner, the ON or OFF event of a load could be identified reliably when load events are separated in time, as shown in the previous chapter.

However, the identification fails when multiple load events occur sufficiently close in time. For example, if two loads are turned on almost at the same time, resulting steady-state power step change will differ from any step change of a load and their (combined) transient response shape will not match any of single-load exemplar patterns. Thus, when a multi-event is detected, an appropriate event space needs to be created so that the measured step-heights can be classified, and suitable transient patterns need to be formed so that the observation can be matched.

Because the probability space of an individual event is known, the probability space of a combination of individual events can be created by a combination of individual probability spaces. Thus the multi-event identification, based on step-height measurements, is simply the matter of forming a set of hypotheses and testing to choose the best hypothesis. The set of feasible hypotheses is denoted as event space. However, because the number of possible combinations of individual load events is large, a systematic and efficient algorithm is necessary to create and search the event space.

By analogy, a transient pattern can be created for a hypothesis, as a combination of the individual exemplars specified by the hypothesis, and the most probable hypothesis can

be chosen from the set, which gives the best match against an observed response. Conceptually, the transient pattern composition to identify a multi-event is no different from the event space search. However, because relative time-indices of individual exemplars are unknown for a given hypothesis, the indices have to be found as a part of the hypothesis test process. This index-search problem adds an enormous computational burden. Thus, in reality, the transient pattern composition can not be implemented for every hypothesis, but for a handful of *best* candidates.

4.2 Overview of multi-event handling

Because the majority of load events are single events, the identification of an event should be performed with the assumption of a single event whenever possible to reduce computational cost. A multi-event handler will be called when the uni-event handler fails or when the event is detected as a multi-event in first place.

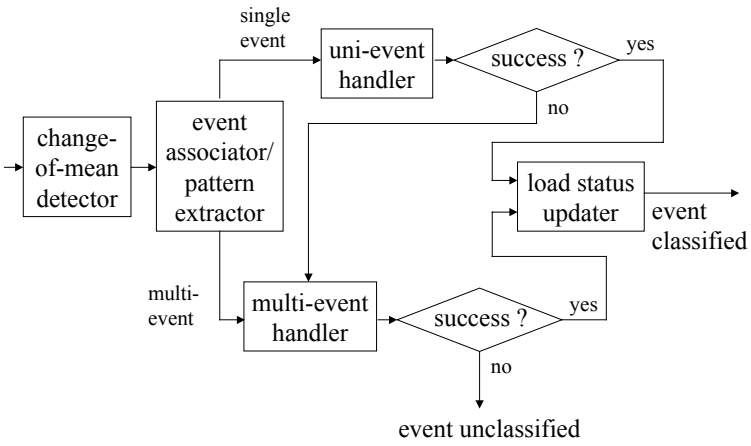


Figure 4-1 Schematic of the overall load event classifier.

Figure 4-1 shows the schematic of the overall event classification scheme. The harmonic power vectors (i.e. [P Q P_3 Q_3 …]) are generated at a power coding rate. The number of vectors (or samples) is first reduced by a proper downsampling. Downsampled power vectors are inspected by the change-of-mean detector, which produces a series of constant-mean segments from the stream of power vectors. The event associator inspects the constant-mean segments to identify possible load events and to extract relevant transient responses for the pattern matching. The event associator, in general, will be able to detect whether the event of interest is a uni-event or multi-event. For example, if a second ON transient is detected before a first ON transient is settled, the associator will be able to recognize two transients as a multi-event. However, the associator can not be perfect and it will sometimes misjudge a multi-event as a uni-event when its transients happen sufficiently close in time.

However, a multi-event, first recognized as a single event, will generate a very poor classification result and reenter the multi-event handler for a second chance. If the classification result is poor even by the multi-event handler, the event is declared *unclassified*. Each handler utilizes both the steady-state load classifier and the transient pattern recognizer and analyzes both outputs to classify an observed event. Upon a successful classification, the load status (ON or OFF for each load) is updated.

As discussed before, because the transient pattern composition search takes an enormous time, it is desirable to conduct the search for a handful of hypothesis candidates. Thus, the multi-event handler first performs the event space search. A few hypotheses are selected from the search, which best account for the step height change observations. Then, the transient pattern composition and matching is conducted for each selected hypothesis. A separate arbitrator determines a best hypothesis from the results of both searches, based on an appropriate arbitration rule.

4.3 Event space search

Suppose a load event $\Delta\mathbf{x} = [\Delta P\ \Delta Q\ \ldots\ ]^T$ is detected. ΔP, ΔQ, … are the increments of the steady-state power levels due to the event. Now, we want to classify it by forming and searching the event space, i.e. the set of event hypotheses based on the steady-state behavior of individual loads.

We assume that there are L electric loads of interest, of which M are on and N are off before the event ($L = M + N$). We define the event direction vector $\mathbf{u}$ as

$$\mathbf{u} = [u_1\ u_2\ \ldots u_L]^T = \left[\underbrace{[-1\ -1\ \ldots\ -1]}_{M} \underbrace{[1\ 1\ \ldots\ 1]}_{N} \right]^T . \tag{4.1}$$

Note that $u_i = -1$ indicates that load i has a potential to change its status from ON to OFF during the event. Similarly $u_i = 1$ indicates an OFF-to-ON change. However, we still do not know what loads are participating the event and change their status as a result. We define the event status vector $\mathbf{v} = [v_1\ v_2\ \ldots\ v_L]^T$, whose entry can be either 1 for change or 0 for non-change. Thus, the product $u_i v_i$ completely characterizes the behavior of load i at the event (1 for OFF-to-ON, 0 for non-change and –1 for ON-to-OFF). Because, there are L loads, there exist $2^L - 1$ possible selections of $\mathbf{v}$. In terms of hypothesis test, for a given $\mathbf{u}$,

$$\begin{aligned}
&H_1 : \mathbf{v} = [1\ 0\ \ldots\ 0]^T \\
&H_2 : \mathbf{v} = [0\ 1\ \ldots\ 0]^T \\
&H_3 : \mathbf{v} = [1\ 1\ \ldots\ 0]^T \\
&\vdots \\
&H_{2^L-1} : \mathbf{v} = [1\ 1\ \ldots\ 1]^T .
\end{aligned} \tag{4.2}$$

A simple conditional maximum likelihood (ML) rule can be conceived for the test:

$$\hat{H}_k = \arg\max_{H_i} p(\Delta\mathbf{x} \mid H_i). \tag{4.3}$$

Each load i is assumed to exhibit a Gaussian (normal) distribution at its steady state with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Lambda}_i$, i.e. $\mathbf{s}_i[n] \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$ where $\mathbf{s}_i[n]$ is the load vector of load i at discrete time n. Thus, the event vector $\Delta\mathbf{x}$ is also Gaussian, because it is a linear combination of independent load vectors. Its probability density function (PDF) under H_i is given by the following formula:

$$p(\Delta\mathbf{x} \mid H_i) = \frac{1}{(2\pi)^{\frac{l}{2}} \det^{\frac{1}{2}}(\boldsymbol{\Lambda}_{\Delta\mathbf{x}|H_i})} \exp\left(-\frac{1}{2} (\Delta\mathbf{x} - \mathbf{m}_{\Delta\mathbf{x}|H_i})^T \boldsymbol{\Lambda}_{\Delta\mathbf{x}|H_i}^{-1} (\Delta\mathbf{x} - \mathbf{m}_{\Delta\mathbf{x}|H_i}) \right), \tag{4.4}$$

where l is the length of $\Delta\mathbf{x}$. The mean and covariance are given by the following formulae:

$$\begin{aligned} \mathbf{m}_{\Delta\mathbf{x}|H_i} &= \sum_{k=1}^{L} u_k(v_k \mid H_i)\boldsymbol{\mu}_k, \\ \boldsymbol{\Lambda}_{\Delta\mathbf{x}|H_i} &= \tfrac{1}{dL_j}\boldsymbol{\Lambda}_{\mathbf{x}_j} + \tfrac{1}{dL_{j-1}}\boldsymbol{\Lambda}_{\mathbf{x}_{j-1}}, \quad \boldsymbol{\Lambda}_{\mathbf{x}_j} = \boldsymbol{\Lambda}_{\mathbf{x}_{j-1}} + \sum_{k=1}^{L} u_k(v_k \mid H_i)\boldsymbol{\Lambda}_k. \end{aligned} \tag{4.5}$$

dL_j and $dL_{j\text{-}1}$ are the numbers of samples for the event segment and the pre-event segment, respectively. $\boldsymbol{\Lambda}_{\mathbf{x}_{j-1}}$ is the sum of load covariance matrices that are on at the pre-event segment. The original ML rule is converted to a simpler minimization problem by taking a negative logarithm of the PDF:

$$\begin{aligned} \hat{H}_k &= \arg\min_{H_i} \xi_i, \\ \xi_i &= -2\ln p(\Delta\mathbf{x} \mid H_i) - l\ln 2\pi \\ &= (\Delta\mathbf{x} - \mathbf{m}_{\Delta\mathbf{x}|H_i})^T \boldsymbol{\Lambda}_{\Delta\mathbf{x}|H_i}^{-1} (\Delta\mathbf{x} - \mathbf{m}_{\Delta\mathbf{x}|H_i}) + \ln\left(\det(\boldsymbol{\Lambda}_{\Delta\mathbf{x}|H_i})\right). \end{aligned} \tag{4.6}$$

Now the hypothesis test is simply the matter of evaluating ξ_i for each hypothesis and selecting a minimum. Although simple conceptually, this hypothesis test poses two problems. The first is that we are interested in finding a few (e.g. three or five) best candidates, not a single minimum, for further evaluations using transient patterns. The second is that due to its combinatorial nature, no polynomial-time algorithm exists even to find a single minimum[4].

[4] This is a nondeterministic polynomial-time (NP) optimization problem with $O(2^L)$ complexity.

A systematic and efficient method is required to list and search every possible **v**. Although many techniques can be conceived, the fact the entries of **v** are either 1 or 0 (binary) suggests us a structure similar to binary tree [Cormen *et al*., 2001] as shown in Figure 4-2.

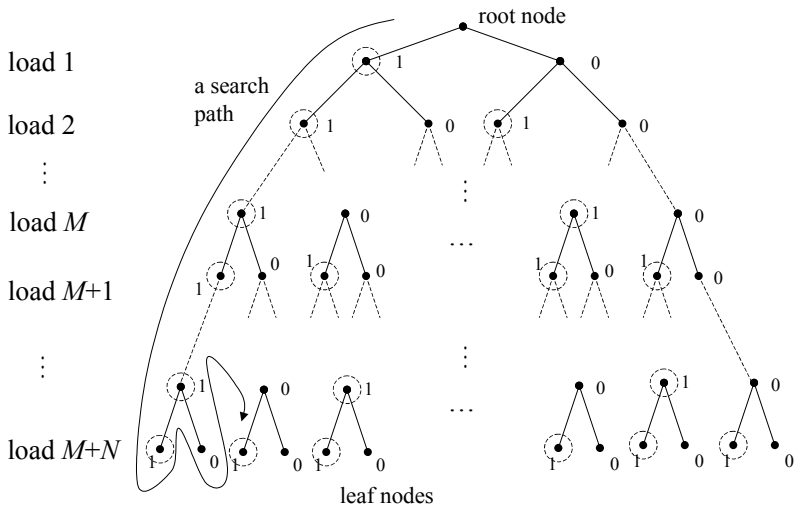


Figure 4-2 Binary search tree of the possible event status vector **v**.

Here, the tree has the height of L and each height level corresponds to a load. Each node has a left child designated as 1 and a right child designated as 0. The tree has 2^L leaf nodes and they represent all possible combinations of **v** (except the rightmost node). The specific value of **v** of a leaf node can be found by backtracking its ancestors. For example, the second leftmost leaf node gives $\mathbf{v} = [1\ 1\ \ldots\ 1\ 0]^T$. However, if we implicitly assume $\mathbf{v} = [0\ 0\ \ldots\ 0]^T$ at the root node, every possible **v** can be found by evaluating **v** at each 1 node, which is circled in the figure. Thus, the left child of the root node gives $\mathbf{v} = [1\ 0\ \ldots\ 0]^T$ and its left child gives $\mathbf{v} = [1\ 1\ 0\ \ldots\ 0]^T$. We also note that the right child of a node gives the same **v** as its parent and thus does not need to be evaluated.

Thus, starting from the root node, by finding **v** and evaluating ξ_i at each 1 node, the set of hypotheses can be searched exhaustively and the best candidates can be selected by sorting the output. However, this scheme requires an exhaustive search of the tree and sorting, which are very time consuming. A heuristic method is often added to the tree search to expedite the process. Heuristics is problem-specific information to guide the search path, or to help decide which node is the best one to expand next [Nilsson, 1998]. Although many different types of heuristics can be conceived, for the given problem, we define the Euclidean distance between the mean (of node n) and the event vector in selected dimensions as a heuristic (evaluation) function $f(n)$ of node n or, equally, hypothesis H_n:

$$f(n) = \left\| \widetilde{\mathbf{m}}_{\Delta\mathbf{x}|H_n} - \Delta\widetilde{\mathbf{x}} \right\|. \tag{4.7}$$

~ (tilde) indicates that the distance is evaluated for a few selected dimensions (e.g. P and Q only), for the reason to be explained later. The heuristic function indicates the distance between a current node and the nominal target $\Delta\mathbf{x}$, and can be used to steer the search direction. In selecting a search direction, we nominally choose to evaluate the children of a node first rather than its peers, a strategy that can be termed depth-first search (compared to *breadth-first* search). Thus, a typical search path goes all the way down to the leftmost leaf node by evaluating the left children of nodes first, comes back up to evaluate the right child of an immediate parent node, etc. A typical search path is shown in Figure 4-2.

We note that in this depth-first search scheme, $\mathbf{v}^T\mathbf{v}$ increases monotonically along a search path from an intermediate node to its leftmost leaf node[5]. We select the dimension of the mean and event vector for the heuristic function as the dimension of a subspace of the harmonic power space, in which the elements of every mean load vector $\boldsymbol{\mu}_i$ are positive. Thus, if every load vector has a positive mean in its P (real power), the dimension of P is selected. If every load vector also has a positive mean in its Q (reactive power), the dimension of Q is selected in addition to P [6]. The elements of $\tilde{\mathbf{m}}_{\Delta\mathbf{x}|H_n}$ decrease monotonically from height 1 to height M, and increase monotonically from height M+1 to height M+N.

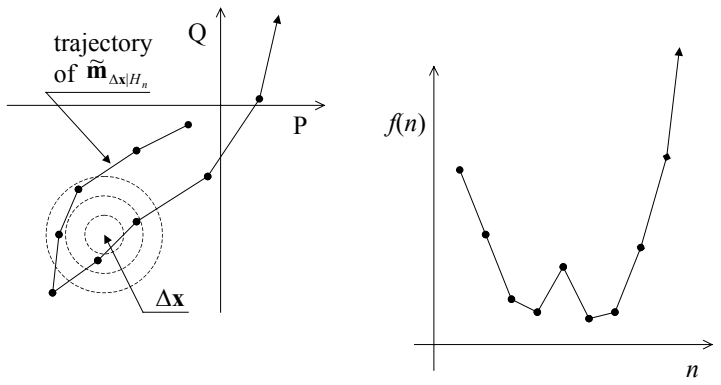


Figure 4-3 Typical search trajectory and resulting behavior of the heuristic function.

Figure 4-3 shows an exemplary search trajectory on the P-Q plane and the resulting behavior of the heuristic function, when the event vector $\Delta\mathbf{x}$ is located at the third quadrant of the plane (a net-OFF event). The exact shape of $f(n)$ depends on the relative location of $\Delta\mathbf{x}$ with respect to the trajectory. However, we note that after the trajectory makes a 'U-turn' at height M, $f(n)$ becomes a convex function, i.e. it reaches a minimum

[5] The leftmost leaf node of a node is a leaf node located farthest left, which can be reached from it without going back to its ancestors.
[6] This is the case where all loads have inductive components (a lagging power factor or a positive reactive factor).

in the vicinity of $\Delta\mathbf{x}$ and increases monotonically afterwards. Thus, with a threshold radius r_{thres}, it is unnecessary to continue search, once the trajectory leaves a circle with the radius of r_{thres} after the U-turn.

We designate the following pruning rule to eliminate a whole branch from a search:

Prune the branch from node n, if $(f(n) \geq r_{thres})$ & $(\nabla f(n) > 0)$ & (height $\geq M$+1). (4.8)

The second (gradient) condition is necessary to ensure that $f(n)$ increases monotonically and the third condition is necessary to apply the pruning rule after the U-turn. In terms of search efficiency, pruning should happen as early as possible. The examination of the pruning rule reveals that pruning occurs early when M is small ($M << N$) and when $\Delta\mathbf{x}$ is located far away from the trajectory after the U-turn. However, the above pruning rule is inefficient when M is large ($M >> N$). In this case, we may swap the event direction vector $\mathbf{u}$, such that

$$\mathbf{u} = [u_1\ u_2\ \ldots u_L]^T = \left[[\underbrace{1\,1\ldots 1}_{N}][\underbrace{-1\,-1\ldots -1}_{M}] \right]^T . \quad (4.9)$$

Also, the binary search tree is constructed with N loads to be turned on closer to root node and M loads to be turned off closer to leaf nodes (N-to-M direction). In this case, the search trajectory first travels to the first quadrant, makes a U-turn and heads toward the third quadrant (when the P-Q plane is selected as the evaluation dimension of the heuristic function). In this case, the pruning rule is the same, except that the third condition is replaced by (height $\geq N$+1).

An ambiguity can occur when M and N are relatively the same ($M \sim N$). Either search direction can be used, but we may choose the M-to-N direction when $\Delta P < 0$, assuming a net-OFF event and the N-to-M direction when $\Delta P > 0$, assuming a net-ON event.

In this manner, every node of a tree is visited and its heuristic function is evaluated. If $f(n) < r_{thres}$, the node (or hypothesis) is moved to a separate bin. If the node satisfies the pruning rule, the entire branch stemming from the node is eliminated from the search. After the search is complete, ξ_i is evaluated for each hypothesis in the bin and the results are sorted to select a number of best minima.

A practical consideration is whether to generate a static tree at program initialization or to generate trees dynamically when needs arise. A static tree is convenient when the number of nodes is small and when its presence expedites a search process. However, in our case, load states change constantly, requiring a different tree at each load event. Also, there really is no advantage in search speed by keeping a static tree. Thus, it makes more sense to create a tree dynamically when a load event occurs.

Also, considering the fact that a node is completely identified by its event status vector $\mathbf{v}$ and its height, it is desirable to create nodes during a search, rather than to create a whole

tree before the search. We create a stack that holds nodes to be searched. Initially the stack holds the root node (or equivalently $\mathbf{v} = [0\ 0\ \ldots\ 0]^T$ and height = 0) only. First, the root node is removed from the stack and its two children are placed in the stack (with the left child on top of the right child). From then on, the top element of the stack is removed, its heuristic function is evaluated and its two children are placed back into the stack unless it is a leaf node or a pruning has occurred. In this manner, every node is created and tested until the stack is empty. We note that the maximum depth of the stack is at most L+1, and hence requires a very small amount of memory.

Because the number of nodes is proportional to 2^L, it is advantageous to reduce the number of loads whose status changes are searched for. For example, if certain loads are known to have turned on, turned off or remain unchanged, these loads are separated from the set of total loads and the search is conducted among the rest[7].

4.4 Transient pattern composition

Let's assume that m event status vectors ($\mathbf{v}$'s) are selected from the event space search, along with a single event direction vector $\mathbf{u}$. We are interested in selecting the best hypothesis (or $\mathbf{v}$) by creating a composite exemplar from the combination of individual exemplars for each hypothesis, and comparing each composite exemplar against an observed transient pattern $\mathbf{y}$ to find the best match. We assume the transient vector $\mathbf{y}$ is of length N, i.e. $\mathbf{y} = \left[y[0]\ y[1] \ldots\ y[N-1]\right]^T$.

For a given hypothesis, we assume that there are $k_{\max}$ loads that either turn on or off. Let n_{j-1} represent the time index of exemplar j, i.e. load j is turned on at time n_{j-1}. Let $\mathbf{x}_j$ represent the length-N exemplar vector of load j. $\mathbf{x}_j$ is filled with zeros between index 0 and $n_{j-1}-1$, and the rest is the exemplar of load j whose transient is referenced from zero DC level. Then the composite exemplar vector $\mathbf{x}$ is simply the sum of individual exemplar vectors:

$$\mathbf{x} = \sum_{j=1}^{k_{\max}} \mathbf{x}_j \,. \tag{4.10}$$

We also make $\mathbf{y}$ referenced to zero DC level, by subtracting the average of prevent segment β, i.e. $\tilde{\mathbf{y}} = \mathbf{y} - \beta$. The computation of β can be done easily because the location of the first event is available from the change-of-mean detector.

We want to minimize the distortion J between $\tilde{\mathbf{y}}$ and $\mathbf{x}$, which is defined as the second norm of the error vector,

$$\text{minimize } J = \left\|\tilde{\mathbf{y}} - \mathbf{x}\right\|^2. \tag{4.11}$$

[7] The evaluations of the heuristic function and ξ_i should still take into account of those separated loads.

Obviously, the smaller is the distortion, the higher is the similarity between the pattern and the exemplar. The distortion is computed for each hypothesis and the hypothesis with a minimum distortion is selected as a match.

However, the time indices of individual exemplars need to be fixed to evaluate J. With no prior knowledge of the time indices, they have to be found as a part of the minimization process by enumeration. Thus, the minimization amounts to $O(N^{k_{\max}})$ complexity to evaluate J for a given hypothesis. In most cases, we can make an educated guess of index bounds from a careful analysis of the change-of-mean detector output. However, because two different events may happen exactly at the same time so that they can appear as a single event, the index bound assessment has to be performed under worst-case scenarios. If average index bound length is reduced to n ($n < N$) after the assessment, the complexity can be reduced to $O(n^{k_{\max}})$. However, the complexity is still exponential in terms of $k_{\max}$.

At this point, one may wonder if we can use *Dynamic Programming* (DP) to find the minimum of J. If applicable, the DP can reduce the exponential complexity of a problem to a polynomial complexity. For a minimization problem to be DP-applicable, its cost function should be represented in terms of *states*, the overall cost needs to be the sum of additive costs which are defined by the states, and the selection of a current state must be independent from the selection of previous states, except its immediate previous state (history independence or Markovianity). For our current problem of minimizing the distortion, the time indices can be chosen as states, but because their relative locations are unknown the additive costs can not be defined.

Thus to make the problem DP-applicable, we have to impose a certain constraint to the time indices. Suppose the time indices are ordered, i.e. $n_{j-1} \le n_j \quad \forall j$. Then the kth additive cost can be defined:

$$\Delta_k[n_{k-1}, n_k - 1] = \sum_{n=n_{k-1}}^{n_k-1} (\tilde{y}[n] - x[n])^2 \quad k = 0, 1, \ldots, k_{\max}, \tag{4.12}$$

with $n_{-1} = 0$ and $n_{k_{\max}} = N$. Note that the kth additive cost is the distortion between indices n_{k-1} and $n_k - 1$ and that the non-zero portion of exemplar k starts at n_{k-1}. The overall cost is simply the sum of additive costs:

$$J = \sum_{k=0}^{k_{\max}} \Delta_k[n_{k-1}, n_k - 1]. \tag{4.13}$$

To find the minimum of J, we can conceive kth intermediate optimal cost at index L:

$$I_k[L] = \min_{\substack{n_0, n_1, \ldots, n_{k-1} \\ n_{-1}=0,\, n_k = L+1}} \sum_{j=0}^{k} \Delta_j[n_{j-1}, n_j - 1]. \tag{4.14}$$

To solve the problem using the DP, we need to convert the intermediate optimal cost to a recursive form, which requires the Markovianity of the states. In other words, the minimum additive cost of current state ($n_{k\text{-}1}$) should be independent from the selection of previous states ($n_{k\text{-}2}$, …, n_0). However, a close examination of the cost model reveals that this condition is not satisfied. For example, a current additive cost changes depending on the start index of its immediately preceding exemplar, unless the exemplar has a constant value after the start index. Thus, the DP is not applicable in a strict sense. However, if we assume that the optimal locations of preceding exemplars give a minimum current additive cost, the DP can be used in an approximate sense. The rationale is that previous best selections conforming to an observed pattern will give rise to the best selection of a time index to add a new exemplar. Now, we can write the intermediate optimal cost in a recursive form:

$$I_k[L] = \min_{n_{k-1}} \left(I_{k-1}[n_{k-1} - 1] + \Delta_k[n_{k-1}, L] \right). \tag{4.15}$$

The overall optimal cost is found when $k = k_{\max}$ and $L = N - 1$.

As explained before, it is advantageous to limit the index bounds, if possible. Let n_j^- be the lower bound and n_j^+ be the upper bound of n_j, i.e. $n_j^- \le n_j \le n_j^+$. The upper and lower bounds should increase monotonically, i.e. $n_{j\text{-}1}^- \le n_j^-$ and $n_{j\text{-}1}^+ \le n_j^+$. Also, $1 \le n_0^-$ and $n_{k_{\max}-1}^+ \le N - 1$. Now the minimum distortion, $I_{k_{\max}}[N-1]$, for a given index order can be found by the following DP algorithm:

1. Let $k = 0$ and $I_{-1}[\cdot] = 0$. For $L = n_0^- - 1, n_0^-, \ldots, n_0^+ - 1$, compute

$$I_0[L] = \Delta_0[n_{-1} = 0, L] = \sum_{n=0}^{L} (\tilde{y}[n])^2.$$

2. Let $k = 1$. For $L = n_k^- - 1, n_k^-, \ldots, n_k^+ - 1$, compute

$$I_k[L] = \min_{n_{k-1}^- \le n_{k-1} \le \min(n_{k-1}^+, L+1)} \left(I_{k-1}[n_{k-1} - 1] + \Delta_k[n_{k-1}, L] \right).$$

Also, for each L, determine $\hat{n}_{k-1}$ at which $I_k[L]$ becomes minimum and call it $n_{k\text{-}1}[L]$.

3. Let $k \leftarrow k + 1$ and iterate until $k = k_{\max}$.

4. Let $k = k_{\max}$. For $L = N - 1$, compute

$$I_k[L] = \min_{n_{k-1}^- \le n_{k-1} \le \min(n_{k-1}^+, L+1)} \left(I_{k-1}[n_{k-1} - 1] + \Delta_k[n_{k-1}, L] \right).$$

Also, find $\hat{n}_{k_{\max}-1} = n_{k_{\max}-1}[N-1]$, i.e. the start index of last segment.

Optimal time indices are found by the following backward recursion:

$$\begin{aligned}
\hat{n}_{k_{\max}-1} &= n_{k_{\max}-1}[N-1] \\
\hat{n}_{k_{\max}-2} &= n_{k_{\max}-2}[\hat{n}_{k_{\max}-1}-1] \\
&\vdots \\
\hat{n}_0 &= n_0[\hat{n}_1-1].
\end{aligned} \tag{4.16}$$

We note that the time indices also need to be found during the execution of the DP algorithm to compute additive costs, i.e. to construct **x**.

In this manner, the minimum distortion for a given index order can be found efficiently. We define the normalized residue (of estimation) γ as the minimum distortion per sample:

$$\gamma = \frac{1}{N} I_{k_{\max}}[N-1]. \tag{4.17}$$

We also define the correlation gain or shape factor (of estimation) $\hat{\alpha}$, as the one which gives a minimum distortion between $\tilde{\mathbf{y}}$ and $\alpha\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is the optimum **x** found from the DP algorithm.

$$\hat{\alpha} = \arg\min_{\alpha} \left\| \tilde{\mathbf{y}} - \alpha\hat{\mathbf{x}} \right\|^2. \tag{4.18}$$

$\hat{\alpha}$ is the solution of the standard least-squares (LS) problem, i.e.

$$\hat{\alpha} = (\hat{\mathbf{x}}^T\hat{\mathbf{x}})^{-1}\hat{\mathbf{x}}^T\tilde{\mathbf{y}}. \tag{4.19}$$

The above optimal cost was found by assuming an order in $k_{\max}$ time indices. Because the exact order is usually unknown, the optimal cost has to be found for every possible order and an order that gives a minimum is to be selected. Thus, we need to run the DP algorithm $k_{\max}!$ times for a given hypothesis to obtain the minimum distortion. Thus, finding a minimum distortion for a given hypothesis using the DP algorithm has a complexity on the order of $O(k_{\max}!\, k_{\max} N^2)$. Compared to the original enumeration scheme, the DP algorithm is advantageous whenever $k_{\max}!\, k_{\max} < N^{k_{\max}-2}$. Because $k_{\max}$ is usually less than 10 and N is on the order of 1000, the DP algorithm is almost always more efficient. This remains true even if N is replaced by n, the average length of index bounds, which is typically on the order of 100. Thus we substantially reduced the computational cost of the original exponential complexity problem.

However, readers will remember that permutation grows very quickly. Table 4-1 shows the permutations of a few selected numbers. Thus, it may be impossible to run the DP algorithm if $k_{\max}$ is very large, although the exact upper bound will be determined by the

computational speed and load of an ELIS computer. Depending on test sites, this may or may not be a problem. For example, if a site has a very low activity rate and event overlaps rarely occur, the worst case scenario may not happen or even if it happens the computer may have enough time to try every possible sequence. On the other hand, if a site has a very high activity rate and event overlaps are common, based on computational budget, the computer may have to declare a failure if k_{max} turns out to be very large.

Table 4-1 Permutations of a few selected numbers.

3!	6
4!	24
5!	120
6!	720
7!	5040
8!	40320
9!	362880

If a failure is declared, the classification of a multi-event is either aborted or can be done based on the event space search only. In latter, classification result is not very reliable and it may need to be double checked by a status estimation technique to be developed later. Another option is to launch a separate process to do the exemplar composition search. The parent process puts the classification on hold until the result from the child process is available, while continuing its nominal tasks.

It also is possible to reduce the number of permutations by an educated guess of possible sequences. For example, if three similar pumps and a drastically different fan compose a hypothesis, we may first want to fix the location of the fan (four possibilities) and then fix the order of the pumps (additional six possibilities), instead of trying all twenty-four possible sequences. This is another example of using problem-specific information to reduce the size of search space. This and other computation saving techniques need to be investigated in future.

After all the possible permutations of time indices are evaluated, the best sequence has to be selected for a given hypothesis. The simplest is to choose a sequence that gives a minimum normalized residue. This is in line with the philosophy of minimizing the distortion between an observation and composite exemplars. However, when there are large amounts of noise, (lower SNR), this minimum normalized residue rule often loses its discernability. In this case, a maximum composite gain rule may be proposed by normalizing the shape factor and residue, compute their product (composite gain) and finding the maximum composite gain.

A practical consideration is how to generate all $k_{max}!$ permutations when k_{max} is given. There exist several methods to produce them in a systematic manner. Lexicographic ordering is most natural to humans, but is computationally expensive. An alternative is to produce a next permutation by changing only two elements of current permutation (transposition) [Nijenhuis and Wilf, 1975; Reingold *et al*., 1977]. Because the number of

total permutations is known, all the permutations can be generated sequentially by the transposition starting from an arbitrary sequence, until a counter reaches $k_{\max}!$.

So far, we have treated each exemplar as having a single section. But in reality, an exemplar can have multiple sections. In terms of finding the best exemplar sequence to account for an observation, these additional sections can be treated as individual exemplars, except that there are sequence constraints. For example, the second section of an exemplar can not be placed ahead of its first section. Suppose that r_j ($r_j \geq 1$) is the number of sections of load j. Then the total number of permutations increases to:

$$\frac{r!}{r_1! r_2! \cdots r_{k_{\max}}!}, \tag{4.20}$$

where $r = \sum_{j=1}^{k_{\max}} r_j$. In the implementation viewpoint, $r!$ permutations are generated as usual by the transposition and only those sequences that satisfy the sequence constraint have chances to initiate the DP algorithm.

In this manner, the best index sequence is selected for a given hypothesis, along with γ, $\hat{\alpha}$ and $\hat{\mathbf{x}}$, to best account for an observation $\mathbf{y}$. However, because there are m hypotheses, the procedure is repeated for each hypothesis, and the best hypothesis is selected based on a criterion. We note that ξ_i's are also available from the event space search. There can be many possibilities in designing the classification criterion based on the three different measures– the normalized residue, shape factor and event space search test statistic.

To streamline the discussion, we normalize the measures. Let $a = \hat{\alpha}$, then the normalized shape factor a_n is defined by the following normalization function:

$$a_n = f(a) = \begin{cases} 0, & \text{if } a \leq 0 \\ a, & \text{if } 0 < a \leq 1 \\ 1/a, & \text{if } a > 1 \end{cases}. \tag{4.21}$$

Two other measures are normalized to the scale of a_n.

$$\gamma_n = 1 - \frac{1 - a_{thres}}{\gamma_{thres}} \gamma, \tag{4.22}$$

where a_{thres} is the threshold of a_n ($0 < a_{thres} < 1$) and γ_{thres} is the threshold of γ ($\gamma_{thres} > 0$).

$$\xi_n = \begin{cases} 1 & \xi \leq 0 \\ 1 - \dfrac{1 - a_{thres}}{\xi_{thres}} \xi & \text{otherwise,} \end{cases} \tag{4.23}$$

where ξ_{thres} is the threshold of ξ ($\xi_{thres} > 0$). Now we define four classification rules. The first two rules utilize the transient pattern composition result only and the remaining two utilizes both results. We assume that there are m hypotheses to choose ($H_1, H_2, \ldots, H_m$).

Rule 1: $\hat{H}_k = \arg\max_{H_i} \gamma_{n|H_i}$.

Rule 2: $\hat{H}_k = \arg\max_{H_i} \gamma_{n|H_i} a_{n|H_i}$.

Rule 3: $\hat{H}_k = \arg\max_{H_i} \gamma_{n|H_i} a_{n|H_i} \xi_{n|H_i}$.

Rule 4: $\hat{H}_k = \arg\max_{H_i} \gamma_{n|H_i} \xi_{n|H_i}$.

Rule 1 uses the event space search simply to find candidates for the transient pattern composition and the best hypothesis is the one that gives a minimum residue. This criterion works well where noise is small but the event space search is relatively unreliable (e.g. due to high background load activity). If noise becomes an issue, Rule 2 can perform better than Rule 1 because the shape factor is relatively immune to noise. Rule 3 and 4 take into account the results of the event space search, and thus can be used where the search yields very reliable results. Depending on noise content, either Rule 3 (high noise) or Rule 4 (low noise) can be used.

4.5 Test of multi-event handler

In this section, we test the multi-event handler with a few multi-events created from single event data. The data were obtained from a test building in 160 Sansome St., San Francisco, CA. The building has six constant loads. Table 4-2 shows their list with their identification numbers.

Table 4-2 Load names and identification numbers.

load id. no.	load name
1	Chilled water pump 1 (CHWP1)
2	Cooling tower pump 1 (CTWP1)
3	Cooling tower fan 1 (CTFan1)
4	Chilled water pump 2 (CHWP2)
5	Cooling tower pump 2 (CTWP2)
6	Cooling tower fan 2 (CTFan2)

We also define a load identification vector $\mathbf{w} = [w_1 \; w_2 \ldots w_L]^T$, whose individual entry holds a load identification number. The load identification vector is to be used in conjunction with the event direction vector **u** and the event status vector **v**, to characterize the load state changes of an event.

Figure 4-4 shows an example of multi-event where three loads turn off almost simultaneously. The original 120 Hz data are downsampled by the 8-point static averaging and then fed to the change-of-mean detector. The figure shows the waveforms of P and Q after downsampling. Because three ON events are distinct, they are recognized as single events and correctly identified by the uni-event handler. However, the last OFF event is the combination of three individual OFF events and it triggers the multi-event handler.

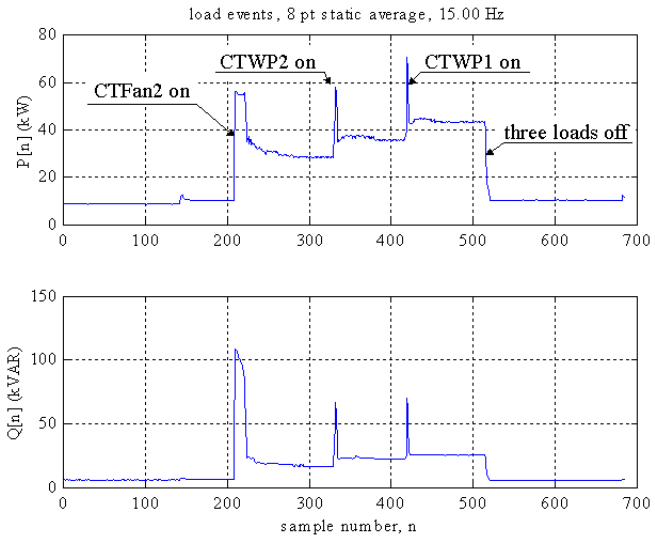


Figure 4-4 A multi-event where three loads turn off simultaneously.

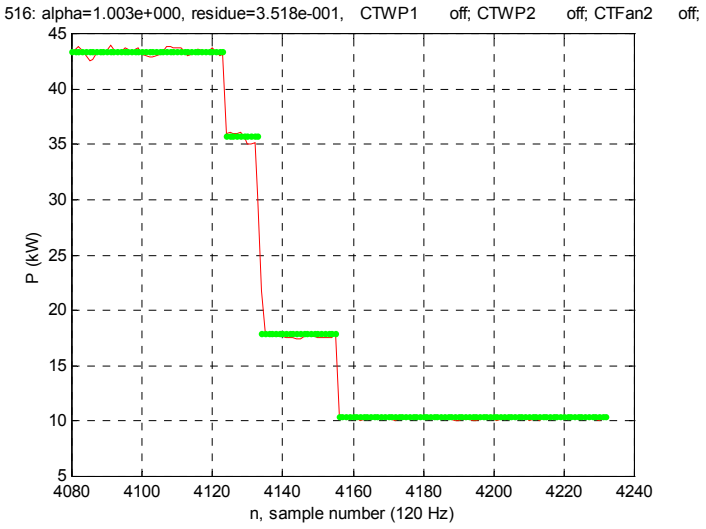


Figure 4-5 Matching result of a composite exemplar for a three-load-off event.

With $\mathbf{w} = [2\ 5\ 6\ 1\ 3\ 4]^T$ and $\mathbf{u} = [-1\ -1\ -1\ 1\ 1\ 1]^T$, the event space search yields a single $\mathbf{v} = [1\ 1\ 1\ 0\ 0\ 0]^T$ with $\xi = -8.1687$. The search result suggests that there was no other state combination within the threshold radius from the event vector ($\Delta\mathbf{x}$). Because there is only one state combination candidate, the transient pattern composition search simply finds the

best time indices of the individual exemplars to account for the observation. Figure 4-5 shows the result of the matching along with the shape factor and residue. The observation is shown in a solid line and the composed exemplars are shown in dots. The graph suggests that the multi-event classification has been very successful.

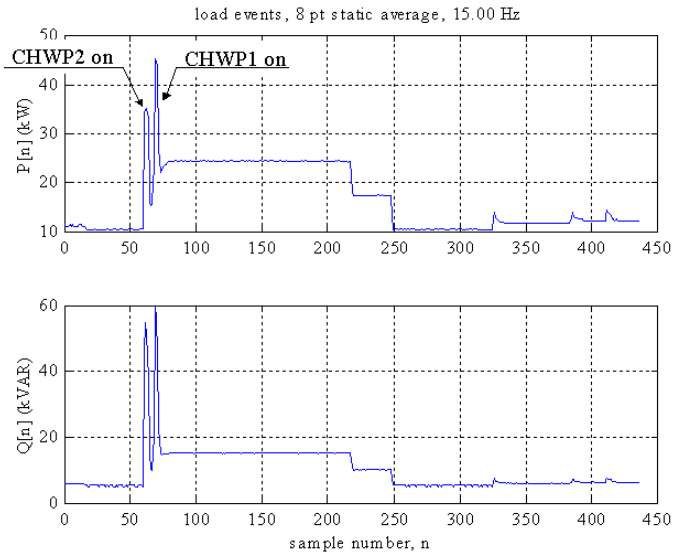


Figure 4-6 A multi-event where two pumps turn on almost at the same time.

Figure 4-6 shows another example of multi-event where two pumps turn on. Because two events happen sufficiently close in time, the event associator recognizes it as a single event and sends it to the uni-event handler. However, because the uni-event handler can not find any match, it declares a failure and sends the event to the multi-event handler.

With $m = 3$, $\mathbf{w} = [1\ 2\ 3\ 4\ 5\ 6]^T$ and $\mathbf{u} = [1\ 1\ 1\ 1\ 1\ 1]^T$, the event space search yields three event status vectors as shown in Table 4-3. The actual event, CHWP1 and CHWP2 ON, is the first candidate with the smallest ξ. The remaining two candidates are other two-pump-on events, which are perfectly legitimate.

Table 4-3 Three event status vector candidates for the multi-event in Figure 4-6.

$\mathbf{v}^T$	ξ
[1 0 0 1 0 0]	2.6301e+002
[0 0 0 1 1 0]	1.8722e+003
[1 0 0 0 1 0]	2.5557e+003

Table 4-4 Measures of the exemplar composition search for the candidates in Table 4-3.

$\mathbf{v}^T$	a	γ
[1 0 0 1 0 0]	9.9111e-001	2.5758e-001
[0 0 0 1 1 0]	9.8452e-001	3.5331e+000
[1 0 0 0 1 0]	9.4626e-001	7.1403e+000

The multi-event handler performs the exemplar composition search for these three candidates, which returns the measures shown in Table 4-4. With Rule 1 as the classification rule, the multi-event handler correctly selects the first hypothesis, which has the smallest residue. Figure 4-7 shows the matching result.

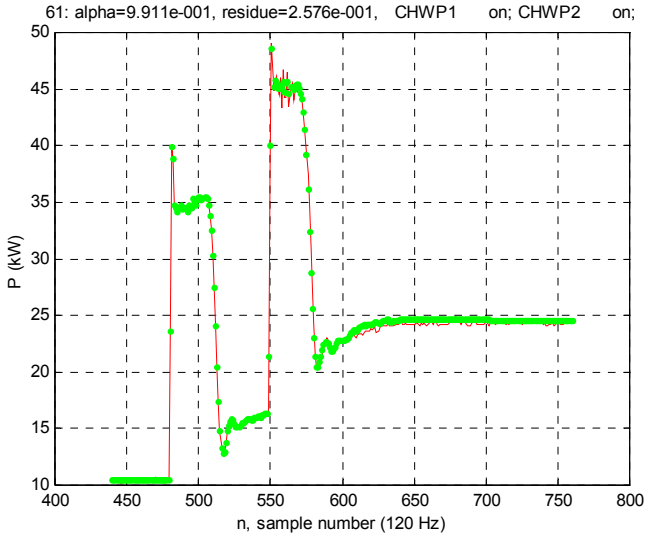


Figure 4-7 Matching result of a composite exemplar for two-pump-on event.

Although rare, it is possible to observe a mixed multi-event, i.e. certain loads turn on while other loads are turned off almost at the same time. Figure 4-8 shows an example of mixed event. The CTWP1 first turns on, and later the CTWP2 and CTFan2 turn on almost at the same time. The CTWP1 turns off slightly after the CTFan2 is turned on. Thus, in this mixed multi-event, two loads are turned on while a third load is turned off.

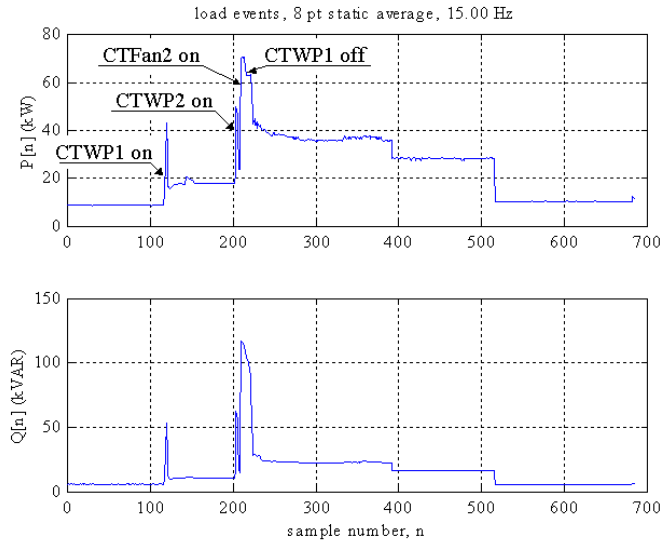


Figure 4-8 A multi-event where one load turns off while two loads turn on.

With $m = 3$, $\mathbf{w} = [1\ 3\ 4\ 5\ 6\ 2]^T$ and $\mathbf{u} = [1\ 1\ 1\ 1\ 1\ -1]^T$, the event space search yields three event status vectors as shown in Table 4-5. Readers will recognize that the actual event is selected as the second candidate. The first candidate is simply the turn-on event of the CTFan2. This example suggests that the event space search can be very unreliable by itself. The reason is that, depending on the nature of individual loads, state combinations can be densely populated in certain regions and, as a consequence, a certain combination may look more probable than others. Also, the event space search is very vulnerable to a background load activity. The exemplar composition search is indispensable, although its cost is extremely high.

Table 4-5 Three event status vector candidates for the multi-event in Figure 4-8.

$\mathbf{v}^T$	ξ
[0 0 0 0 1 0]	7.6631e+002
[0 0 0 1 1 1]	1.3155e+004
[0 0 1 0 1 1]	1.7497e+004

Table 4-6 shows the measures of the exemplar composition search. With Rule 1, the multi-event handler correctly selects the second hypothesis. Figure 4-9 shows the matching result. It is interesting to observe that the CTWP1 turns off in the middle of the plateau of the CTFan2 ON event. This example also shows the versatility and flexibility of the exemplar composition search based on the DP algorithm and the permutation, even though the DP algorithm was applied in an approximate sense.

Table 4-6 Measures of the exemplar composition search for the candidates in Table 4-5.

$\mathbf{v}^T$	a	γ
[0 0 0 0 1 0]	1.0184e+000	3.3293e+001
[0 0 0 1 1 1]	9.8272e-001	3.3517e+000
[0 0 1 0 1 1]	1.0119e+000	6.7727e+000

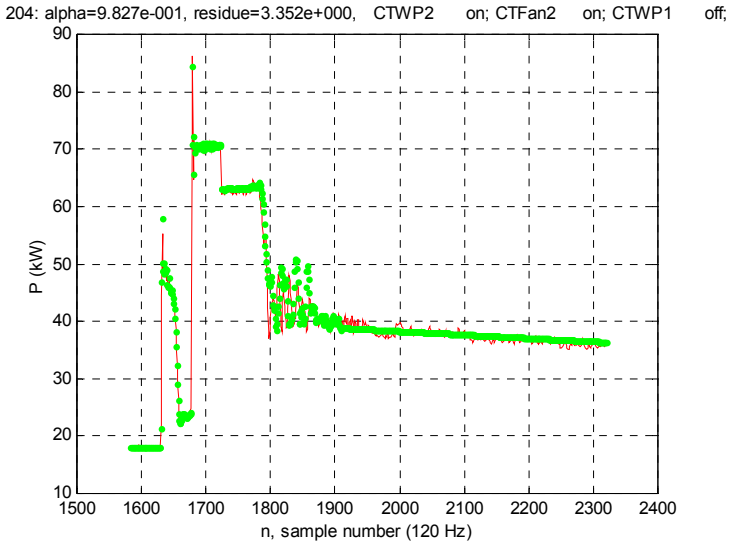


Figure 4-9 Matching result for two-load-on & one-load-off event.

4.6 Chapter summary

A systematic method has been developed in this chapter to identify the ON/OFF status of constant electric loads when their events overlap in time. A set of hypotheses, i.e. which loads turn on or off, is first proposed for a given event observation. The probability space of a hypothesis is created as a linear combination of the probability spaces of individual loads, and a few hypotheses that give largest probabilities are selected as candidates. A binary-tree-like structure with heuristics was developed to generate and search the event space efficiently.

The exemplar composition search is conducted for the hypothesis candidates selected from the event space search. A composite exemplar is formed as the sum of individual exemplars based on a hypothesis to best account for the transient response observation. This is a non-deterministic polynomial-time optimization problem, but its cost is substantially reduced by the permutation of time-index sequences and the application of dynamic programming in an approximate sense. The cost is further reduced by an educated guess of index bounds. However, the overall cost is still high due to the permutation. The best hypothesis is selected according to a classification rule, based on the measures from both searches.

Test results prove that the multi-event handler correctly classifies overlapping events, and also demonstrate the need to conduct the two searches to identify a multi-event. The two searches are connected in series for two reasons. First, because the exemplar composition search has an enormous complexity, it is imperative to conduct the event space search first to confine the search space of an exemplar composition. Second, certain regions of the state space can be so densely populated with possible state combinations that the event space search may lose its discernability when an event vector is located in one of those regions. Also, the event space search is vulnerable to background load activity. Thus, it is essential to double-check an event space search result with the exemplar composition search.

One practical concern is the time cost involved in the multi-event handling, especially that of the exemplar composition search. Although the exact upper bound has to be determined after the multi-event handler is implemented in real time, it is obvious that it will not be able to function when tens of events overlap. After an upper bound is determined, there are two possibilities depending on the nature of test sites. The first case is when the maximum number of overlapping events rarely exceeds the upper bound. In this case, the multi-event handler developed in this chapter can be employed without any modification. The second, more general, case is when the maximum number of overlapping events indeed exceeds the upper bound. In this case, further refinements of the multi-event handling algorithm may be necessary, such as reducing the size of event space by eliminating certain loads whose statuses are clearly known and reducing the number of permutations in the exemplar composition using heuristic information. However, we have to admit that the multi-event handler sometimes can declare a failure when the complexity of a given multi-event exceeds its capability.

The possibility of multi-event handler failure suggests that we can not rely solely on load events to identify the status of loads. We may have to utilize the steady-state outputs of loads. State estimation (based on steady-state outputs only) is necessary to identify the initial status of loads when the program first starts and to make a better decision when the edge estimation (based on event outputs) either fails or produces an unreliable result. We will next study the state estimation problem.

Chapter 5

Electric Load State Estimation

This chapter describes the general method of identifying the ON/OFF status of constant electric loads from the observation of steady-state power waveforms. A set of hypotheses, concerning the status of electric loads, is first proposed and, when conditions are met, the observed segments of steady-state powers are used to determine the best hypothesis. The state estimation, combined with the identification techniques based on load events, provides a complete solution to identify the status of multiple constant electric loads from the harmonic power waveforms generated at a single service entry.

5.1 Introduction

It usually is a difficult task to identify the ON/OFF status of electric loads from the observation of the non-intrusive power signal alone. Fortunately, many constant electric loads have unique power consumption levels at their steady states and furthermore have distinctive transient responses. Exploiting these *fingerprints* in a systematic manner, the ON or OFF event of a load can be identified reliably. When multiple loads turn on or off at the same time, the resulting event can still be identified by combining individual load events to best emulate the observation, as shown in the previous chapter.

However, these approaches rely on the changes of load status, i.e. load events, to estimate the status. In other words, if there is no load event, the load status can not be determined. An immediate problem is the initial status estimation, i.e. when an identification program is first initialized it does not know what loads are on at that time. Even after a few load events are identified, the initial *unknowns* remain unknown. This may not be a problem, when the load status is identified off-line by analyzing a sequence of load events. But in a real time situation, the initial load status needs to be known to make timely decisions. On the other hand, when a load event is misclassified, there is no mechanism to correct the error, and the program is prone to further misclassifications due to the erroneous status information generated by the prior misclassification. Inter-event power signal has been under-utilized, which may hold potential information to determine the load status.

We introduce the concept of *state* as the combination of the ON/OFF status of individual loads at a given instant, and *edge* as the boundary at which a state changes. The previous approaches rely on edge outputs to determine states. In this chapter, we develop a state estimation technique based on state outputs. The state estimator provides a means to identify the initial load status. It also can be used to augment the edge classifier by correcting any misclassification error. Also, depending on the circumstances, a separate decision making authority may be necessary to make a decision concerning the load status based on both the state and the edge output.

5.2 Overview of load state estimation

The state estimator utilizes the harmonic power signal, generated when certain loads are at their steady ON state while the others are completely OFF, to determine their collective ON/OFF status. Thus, the state estimator[8] compliments the edge classifier that relies on edge outputs, and hence makes a better use of two different characteristics of loads– state and edge– for the objective of reliable real-time load status identification.

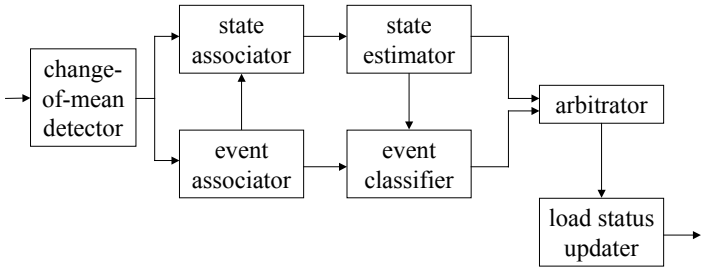


Figure 5-1 Schematic of the overall load status identifier.

Figure 5-1 shows the overall load identification scheme, where two estimators are placed in parallel. A stream of harmonic power vectors is first examined by the change-of-mean detector and is converted to multiple constant-mean segments. The event associator examines the constant-mean segments to find a load event. A load event typically involves a transient that can span over several constant-mean segments. The event associator assembles a load event from the analysis of those transient segments, and sends it to the event classifier. Also, the event associator marks those transient segments and forward the information to the state associator.

The state associator assumes a distinct load state for a series of constant-mean segments unmarked by the event associator. The state associator generates several state hypotheses from the first segment of the series and the test metric for each hypothesis is updated with the accumulation of the constant-mean segments. The series is terminated either when a segment is marked by the event associator or when a decision has to be made by the arbitrator. A new series is created immediately after a new unmarked segment appears. Thus, a homogeneous state is assumed throughout a series and, at the end of the series, state hypothesis candidates are available with their test metrics.

[8] Strictly speaking, it should be called a state *classifier*, because it involves coming up with a finite number of state hypotheses and choosing the best. Estimator is usually reserved for a decision making process, where the number of possibilities is infinite or the variable of interest can take a continuous value. However, in this thesis, we slightly abuse the terminology and a classifier is also called an estimator.

The arbitrator makes an intelligent decision about the load status, based on the outputs of both the state and the edge estimator. The decision strategy may depend on the nature of the monitored loads. For example, if the state estimation is fairly reliable after a certain observation interval, it may make a decision based on the state estimation output only. On the other hand, if the state estimation is not very reliable, it may not make a decision until a load event is observed. In this case, the event estimator performs event classification repeatedly based on each state hypothesis and the hypothesis that best accounts for the observed event is finally selected. Thus, the state estimation is coupled with the event classification under this scenario.

The state estimator can also be used to check the event classification result. For example, if the load status determined from an event classification is not found among the state hypotheses generated by the state estimator after the event, the arbitrator can determine that there was an error in the previous event classification. The arbitrator may choose another possibility proposed by the prior event classification or simply designate the current status as *unknown* and enter the initialization mode.

5.3 State estimation

We assume that there are m constant electric loads to be monitored by the ELIS. In addition to those m loads, we assume the presence of a background load. The background load can be either a collection of small physical loads, such as lighting or plug loads, which tap currents from the main trunk where the current sensor is located, or simply the DC offsets present in harmonic power vectors even when those m loads are all off. We define a load status vector, $\mathbf{v} = [v_0\ v_1\ v_2\ \dots\ v_m]^T$, whose ith entry is either 1 for the ON status or 0 for the OFF status of load i. The zeroth entry v_0 represents the status of the background load, which is always 1.

For a constant mean segment of length N, there exist N harmonic power vectors, each of length l:

$$\mathbf{x}[n] = \left[P[n]\ Q[n] \dots\ \right]^T \qquad 0 \le n \le N-1. \tag{5.1}$$

The dimension of the power vector is determined by the nature of the constant loads. Typically, only the real and reactive powers are used for the status identification and hence $l = 2$. However, other harmonic powers, such as the third, can be used if all the loads generate constant amounts of those harmonic powers when they are on.

Each load i is assumed to exhibit a Gaussian (normal) distribution at its steady state with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Lambda}_i$, i.e. $\mathbf{s}_i[n] \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$, where $\mathbf{s}_i[n]$ is the load vector of load i at discrete time n. We assume that the load status remains unchanged for the duration of the segment. Thus, the power vector can be represented as a linear combination of the load vectors that are on.

$$\mathbf{x}[n] = \sum_{i=0}^{m} v_i \mathbf{s}_i[n]. \tag{5.2}$$

Because each load is statistically independent, the power vector also follows a Gaussian distribution:

$$\mathbf{x}[n] \sim N\left(\boldsymbol{\mu}_{\mathbf{x}} = \sum_{i=0}^{m} v_i \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_{\mathbf{x}} = \sum_{i=0}^{m} v_i \boldsymbol{\Lambda}_i \right). \tag{5.3}$$

We organize the N power vectors into a single large $lN \times 1$ data vector $\mathbf{y}$:

$$\mathbf{y} = \left[\mathbf{x}[0]^T \ \mathbf{x}[1]^T \ \ldots \mathbf{x}[N-1]^T \right]^T. \tag{5.4}$$

This type of arrangement is called *temporal ordering* or *column rollout* in array signal processing [Kay, 1998]. Because each power vector is Gaussian and statistically independent, the data vector $\mathbf{y}$ is also Gaussian with,

$$\text{mean: } \mathbf{m}_{\mathbf{y}} = \left[\boldsymbol{\mu}_{\mathbf{x}}{}^T \ \boldsymbol{\mu}_{\mathbf{x}}{}^T \ \ldots \boldsymbol{\mu}_{\mathbf{x}}{}^T \right]^T \tag{5.5}$$

$$\text{covariance: } \boldsymbol{\Lambda}_{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Lambda}_{\mathbf{x}} & & & \\ & \boldsymbol{\Lambda}_{\mathbf{x}} & & \\ & & \ddots & \\ & & & \boldsymbol{\Lambda}_{\mathbf{x}} \end{bmatrix}. \tag{5.6}$$

Note that the covariance matrix is a block diagonal matrix with zero off-diagonal block entries.

Now the load status estimation, given the length-N constant-mean segment, is simply the matter of determining $\mathbf{v}$ among 2^m possible alternatives. In the form of a hypothesis test,

$$\begin{aligned} H_0 &: \mathbf{v} = [1\,0\,0 \ldots 0]^T \\ H_1 &: \mathbf{v} = [1\,1\,0 \ldots 0]^T \\ H_2 &: \mathbf{v} = [1\,0\,1\,0 \ldots 0]^T \\ H_3 &: \mathbf{v} = [1\,1\,1\,0 \ldots 0]^T \\ &\vdots \\ H_{2^m-1} &: \mathbf{v} = [1\,1\,1 \ldots 1]^T. \end{aligned} \tag{5.7}$$

A simple conditional maximum likelihood (ML) rule can be conceived for the test:

$$\hat{H}_k = \arg\max_{H_i} p(\mathbf{y} \mid H_i). \tag{5.8}$$

The probability density function (PDF) under H_i is given by the following formula:

$$p(\mathbf{y} \mid H_i) = \frac{1}{(2\pi)^{\frac{lN}{2}} \det^{\frac{1}{2}}(\mathbf{\Lambda}_{\mathbf{y}|H_i})} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{m}_{\mathbf{y}|H_i})^T \mathbf{\Lambda}_{\mathbf{y}|H_i}^{-1} (\mathbf{y} - \mathbf{m}_{\mathbf{y}|H_i}) \right). \tag{5.9}$$

Because $\mathbf{\Lambda}_\mathbf{y}$ is block diagonal, the conditional PDF can be simplified by using the following relationships:

$$\det \mathbf{\Lambda}_\mathbf{y} = \prod_0^{N-1} \det \mathbf{\Lambda}_\mathbf{x} = (\det \mathbf{\Lambda}_\mathbf{x})^N \text{ and } \mathbf{\Lambda}_\mathbf{y}^{-1} = \begin{bmatrix} \mathbf{\Lambda}_\mathbf{x}^{-1} & & & \\ & \mathbf{\Lambda}_\mathbf{x}^{-1} & & \\ & & \ddots & \\ & & & \mathbf{\Lambda}_\mathbf{x}^{-1} \end{bmatrix}. \tag{5.10}$$

Now, the PDF becomes:

$$p(\mathbf{y} \mid H_i) = \frac{1}{(2\pi)^{\frac{lN}{2}} (\det \mathbf{\Lambda}_{\mathbf{x}|H_i})^{\frac{N}{2}}} \exp\left(-\frac{1}{2} \sum_{n=0}^{N-1} (\mathbf{x}[n] - \mathbf{\mu}_{\mathbf{x}|H_i})^T \mathbf{\Lambda}_{\mathbf{x}|H_i}^{-1} (\mathbf{x}[n] - \mathbf{\mu}_{\mathbf{x}|H_i}) \right). \tag{5.11}$$

The original ML rule is converted to a simpler minimization problem by taking a negative logarithm of the PDF:

$$\begin{aligned} \hat{H}_k &= \arg\min_{H_i} \xi_i, \\ \xi_i &= -2 \ln p(\mathbf{y} \mid H_i) \\ &= lN \ln 2\pi + N \ln\left(\det \mathbf{\Lambda}_{\mathbf{x}|H_i}\right) + \sum_{n=0}^{N-1} (\mathbf{x}[n] - \mathbf{\mu}_{\mathbf{x}|H_i})^T \mathbf{\Lambda}_{\mathbf{x}|H_i}^{-1} (\mathbf{x}[n] - \mathbf{\mu}_{\mathbf{x}|H_i}). \end{aligned} \tag{5.12}$$

There can be two scenarios of the hypothesis test, depending on the nature of the background load. If the background load has a fixed and known mean, the above ML rule can be used without any modification. However, in a more general case, the background load can have a variable mean. Yet, it is assumed to be constant for a given segment. In this case, the background mean has to be estimated as a part of the hypothesis test process and thus the above ML rule needs to be modified.

We first examine the fixed background mean case. To conduct an effective search among 2^m possible hypotheses, we use the depth-first search binary tree introduced in the previous chapter. Figure 5-2 shows the schematic of the binary tree. The load status vector $\mathbf{v}$ is assumed to be $[1\ 0\ 0\ \ldots\ 0]^T$ at the root node. At height i, the ith entry of $\mathbf{v}$

changes its status at 1 nodes and remains unchanged at 0 nodes. In this manner, every possible **v** is generated from the binary tree.

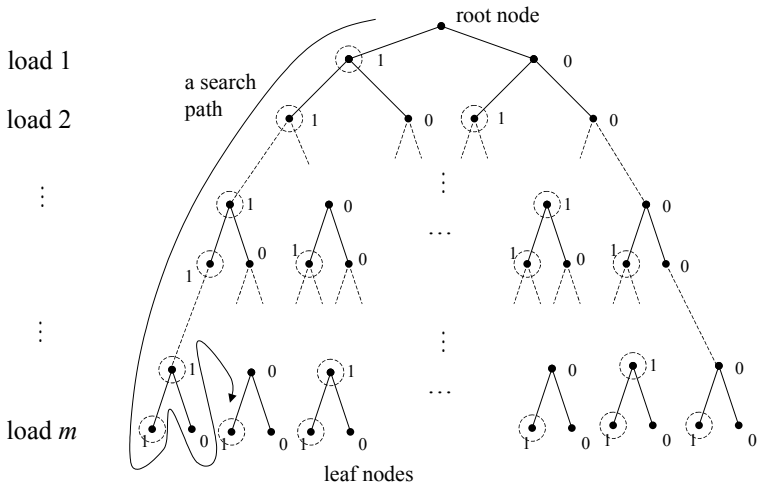


Figure 5-2 Binary search tree for the possible load status vector **v**.

The Euclidean distance between the mean and the segment average is selected as a heuristic function to prune unnecessary branches. For a node *n*,

$$f(n) = \left\| \tilde{\boldsymbol{\mu}}_{\mathbf{x}|H_n} - \tilde{\mathbf{x}}_{avg} \right\|, \tag{5.13}$$

where $\mathbf{x}_{avg}$ is the segment average of the *N*-point power vectors, i.e. $\mathbf{x}_{avg} = \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{x}[i]$. Tilde indicates that the dimension used for the evaluation of the heuristic function is the dimension of the subspace of the harmonic power space, in which the elements of every mean load vector $\boldsymbol{\mu}_i$ are positive. Now, it is easy to show that $f(n)$ is a convex function and we can conceive the following pruning rule:

$$\text{Prune the entire branch from node } n, \text{ if } (f(n) \geq r_{thres}) \;\&\; (\nabla f(n) > 0), \tag{5.14}$$

where r_{thres} is the threshold radius. Pruning occurs early when the minimum of the heuristic function is located close to the root node. Thus the above search direction is efficient when $\left\|\tilde{\mathbf{x}}_{avg}\right\| \leq a \max_{H_n}\left\|\tilde{\boldsymbol{\mu}}_{\mathbf{x}|H_n}\right\|$, where a is a fraction. If $\left\|\tilde{\mathbf{x}}_{avg}\right\| > a \max_{H_n}\left\|\tilde{\boldsymbol{\mu}}_{\mathbf{x}|H_n}\right\|$, we may swap the search direction by assigning $\mathbf{v} = [1\ 1\ 1\ \ldots\ 1]^T$ at the root node. The heuristic function and the pruning rule remain unchanged in this case.

In this manner, each node of the tree is visited and its heuristic function is evaluated. If $f(n) < r_{thres}$, a copy of the node (or hypothesis) is moved to a separate candidate bin. If a node satisfies the pruning rule, the entire branch stemming from the node is eliminated from the search. After the search is complete, the test statistic ξ_i is evaluated for each hypothesis in the bin.

The tree search is conducted once for the first segment of a common load status series. For its subsequent constant-mean segments, just the test statistic is evaluated and recorded for each hypothesis in the candidate bin. If there are M segments for a given series, the overall test statistic ξ_i for H_i is simply the weighted sum of individual segment test statistics:

$$\xi_i = \frac{\sum_{k=1}^{M} N_k \xi_{i,k}}{\sum_{k=1}^{M} N_k}, \tag{5.15}$$

where N_k is the length and $\xi_{i,k}$ is the test statistic of the kth segment.

In the fixed background load mean case, the state estimation is fairly reliable if an observed series is long enough. Thus, the arbitrator may make an appropriate decision based on the overall test statistics in the bin and the accumulated segment length, i.e. it selects a hypothesis with a minimum test statistic when the accumulated length exceeds a certain threshold.

We next examine the case when the background load has a varying mean, which commonly occurs at real buildings where small auxiliary loads are connected to the panel designed for the major (HVAC) loads. Figure 5-3 shows an example of the background load activity when all major loads are off. The data were taken at a test building in San Francisco, CA. There are many small ON/OFF activities in the background. However their step heights are too small (~1 kW), compared to other major loads (~10 kW), and their collective ON/OFF activities are too frequent to be kept track of reliably.

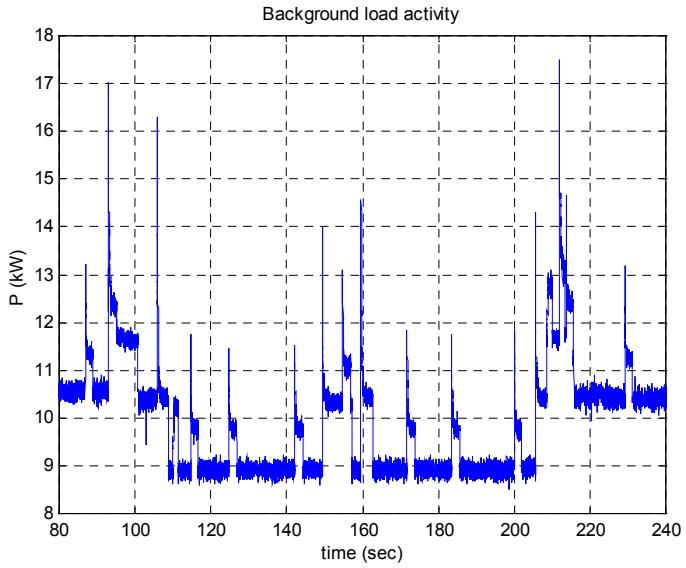


Figure 5-3 An example of background load activity.

However, it is reasonable to assume that the background load has a constant mean for a short duration of time. The changes of the background mean can normally be captured by the change-of-mean detector, except when the changes are masked by the larger transients of the major loads. We also assume that there are upper and lower bounds in the value the background mean can take and that those bounds are known. We further assume that the background load has a constant covariance matrix, even though its mean changes with time.

Because the background mean $\boldsymbol{\mu}_0$ is unknown, it has to be estimated as a part of the hypothesis test procedure for a given constant-mean segment. We assume that $\boldsymbol{\mu}_0$ is a deterministic but unknown quantity with a certain set of known bounds. The conditional PDF is rewritten with $\boldsymbol{\mu}_0$ as an unknown parameter:

$$p(\mathbf{y};\boldsymbol{\mu}_0 \mid H_i) = \frac{1}{(2\pi)^{\frac{lN}{2}} (\det \boldsymbol{\Lambda}_{\mathbf{x}|H_i})^{\frac{N}{2}}} \exp\left(-\frac{1}{2} \sum_{n=0}^{N-1} (\mathbf{x}[n] - \boldsymbol{\mu}_0 - \textstyle\sum_{k=1}^{m} v_{k|H_i} \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_{\mathbf{x}|H_i}^{-1} (\mathbf{x}[n] - \boldsymbol{\mu}_0 - \textstyle\sum_{k=1}^{m} v_{k|H_i} \boldsymbol{\mu}_k) \right). \quad (5.16)$$

We use the generalized ML rule to account for the additional uncertainty involved in estimating $\boldsymbol{\mu}_0$ [Kay, 1998]:

$$\begin{aligned} \hat{H}_k &= \arg\min_{H_i} \xi_i , \\ \xi_i &= -2 \ln p(\mathbf{y}; \hat{\boldsymbol{\mu}}_{0,i} \mid H_i) + \ln \det\left(\mathbf{I}_{\mathbf{y}}(\hat{\boldsymbol{\mu}}_{0,i})\right), \end{aligned} \quad (5.17)$$

where $\hat{\boldsymbol{\mu}}_{0,i}$ is the ML estimation of $\boldsymbol{\mu}_0$ assuming H_i is true. $\mathbf{I}_{\mathbf{y}}(\boldsymbol{\mu}_0)$ is the Fisher information matrix [Papoulis and Pillai, 2002] assuming H_i is true. It is straightforward to show that

$$\hat{\boldsymbol{\mu}}_{0,i} = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}[n] - \sum_{k=1}^{m} v_{k|H_i} \boldsymbol{\mu}_k . \quad (5.18)$$

The evaluation of the Fisher information matrix is rather involved. We simply state the result:

$$\mathbf{I}_{\mathbf{y}}(\boldsymbol{\mu}_{0,i}) = E\left(\left(\frac{\partial \ln p(\mathbf{y};\boldsymbol{\mu}_{0,i})}{\partial \boldsymbol{\mu}_{0,i}} \right)^T \left(\frac{\partial \ln p(\mathbf{y};\boldsymbol{\mu}_{0,i})}{\partial \boldsymbol{\mu}_{0,i}} \right) \right) = N \boldsymbol{\Lambda}_{\mathbf{x}|H_i}^{-1}. \quad (5.19)$$

If we substitute these two relationships into the generalized ML rule, after a few algebraic steps, the test statistic ξ_i becomes:

$$\xi_i = lN \ln 2\pi + N \ln \det \boldsymbol{\Lambda}_{\mathbf{x}|H_i} + \sum_{n=0}^{N-1} (\mathbf{x}[n] - \mathbf{x}_{avg})^T \boldsymbol{\Lambda}_{\mathbf{x}|H_i}^{-1} (\mathbf{x}[n] - \mathbf{x}_{avg}) + l \ln N - \ln \det \boldsymbol{\Lambda}_{\mathbf{x}|H_i} . \quad (5.20)$$

Compared to the test statistic of the constant background mean case, the sum of the load means under a hypothesis is replaced by the segment average. The last two terms are

penalty terms introduced by the Fisher information matrix. Because the fourth term is common among the hypotheses given a segment, only the fifth term is a real penalty factor in evaluating different hypotheses. We also note that the influence of the penalty term diminishes with increasing N.

The above test statistic has an interesting feature. Because $\boldsymbol{\mu}_{\mathbf{x}|H_i}$ is replaced by $\mathbf{x}_{avg}$, the sum of the load means under H_i plays no role in determining the test statistic. Any difference between the segment average and the sum of the individual load means under a hypothesis is attributed to the background load. Of course, because the bounds for the background mean are known, certain hypotheses that give out-of-bound background mean estimations can be eliminated from the consideration. Then, it is the covariance matrix that allows distinguishing different hypotheses. The above generalized ML rule has more resolving power when loads have more distinctive covariance matrices. However, when loads show similar noise patterns, the test may not be reliable.

The hypothesis candidates can be found by using the binary tree shown in Figure 5-2. Let's assume the bounds for the background mean are known for the real power (P):

$$\mu_{0P}{}^{-} \le \mu_{0P} \le \mu_{0P}{}^{+}, \tag{5.21}$$

where μ_{0P} is the real power component of the background mean vector $\boldsymbol{\mu}_0$. For the heuristic function, we select:

$$f(n) = \hat{\mu}_{0P}(n), \tag{5.22}$$

i.e. the real power component of the ML estimation of the background mean at node n. If $\mathbf{x}_{avg}[1] \le a\sum_{i=1}^{m} \mu_{iP}$, where $\mathbf{x}_{avg}[1]$ is the real power component of the segment average, μ_{iP} is the real power component of the mean vector of load i and a is a fraction, the load status vector $\mathbf{v}$ is set to $[1\ 0\ 0\ \ldots\ 0]^T$ at the root node. With this search direction, it is easy to show that the heuristic function decreases monotonically along a branch. We designate the following pruning rule:

Prune the entire branch from node n, if $f(n) \le \mu_{0P}{}^{-}$. (5.23)

If $\mathbf{x}_{avg}[1] > a\sum_{i=1}^{m} \mu_{iP}$, we set $\mathbf{v} = [1\ 1\ 1\ \ldots\ 1]^T$ at the root node. In this case, the heuristic function increases monotonically. The pruning rule changes to:

Prune the entire branch from node n, if $f(n) \ge \mu_{0P}{}^{+}$. (5.24)

Other harmonic powers, such as the reactive power, can be jointly used for pruning, if the background mean bounds are known and all loads have positive means for the harmonic powers of interest.

In this manner, each node of the tree is visited and its heuristic function is evaluated. If the heuristic function, the ML estimation of the background mean, satisfies the bounds, a copy of the node is moved to a separate candidate bin. As before, the tree search is conducted once for the first segment of a common load status series, and just the test statistic is evaluated for each hypothesis in the bin in subsequent constant-mean segments. When the series is terminated, the overall test statistic for each hypothesis is computed as a weighted sum using (5.15).

As discussed before, because the sum of the load means under a hypothesis does not play any role in determining the best hypothesis, the state estimation in the varying background mean case is not reliable in most cases, except when the loads have distinct covariance matrices. Thus, it is risky for the arbitrator to make a decision based on the state estimator output only. One of the viable strategies in this case is to couple the state and the edge estimator to make a more intelligent decision. For example, the edge estimator may perform its estimation based on the state hypotheses provided by the state estimator. On the other hand, the result of the edge estimation may be checked against the state hypotheses generated by the state estimator. We propose two arbitration strategies– when the load status is unknown and when it is known.

When the load identification program is first started, the load status is unknown and the program is in an initialization mode. The state estimator first conducts the binary tree search to find state hypothesis candidates. Their test statistics are updated with the accumulation of the constant-mean segments, and the process is terminated when a load event is observed. We designate a threshold width, $\Delta\xi \geq 0$, for the test statistics of the state hypotheses. The hypotheses, which have their test statistics within $\Delta\xi$ from the minimum test statistic, are selected as valid ones to be forwarded to the edge estimator.

The edge estimator conducts its test to identify the observed load event, assuming a pre-event load status for each forwarded state hypothesis. After the test, the hypothesis which gives rise to the best event classification result is selected. At this point, the load event is identified. The arbitrator further examines if other state hypotheses yielded the same load event classification result. If so, the state hypothesis with a minimum test statistic among those hypotheses is selected as the state of the pre-event. Otherwise, the original state hypothesis is retained. Thus, the load status prior to the load event is also resolved after the load event identification. The initialization mode is released after the successful edge event classification.

However, the edge estimation sometimes fails. For example, no valid state hypothesis may exist at the time of edge estimation, no state hypothesis may give rise to a reliable edge classification result, or if the computational complexity of particular event identification is too high the edge estimator may voluntarily declare a failure. In this case, the arbitrator decides not to make a decision and simply enters the initialization mode, starting to generate new state hypotheses and waiting for a next decisive edge event.

Once the initialization mode is released, the load status is known and the edge estimator can perform its duty without the assistance of the state estimator. Ideally, subsequent load status identifications can be performed solely by the edge estimator. However, in a real life situation, the edge estimator may fail or misclassify a load event. If the edge estimator fails, the arbitrator simply enters the initialization mode. The consequence of an edge event misclassification can be subtle. Depending on the circumstances, a misclassification happens once and it may not greatly affect the outcomes of future edge classifications. On the other hand, it can lead to a series of future misclassifications.

Thus, it is necessary for the arbitrator to have a list of *legal* state hypotheses, even when the current load status is known after a seemingly successful edge classification. The legal hypotheses are those who have the test statistics within the threshold width from the minimum test statistic, among the state hypothesis candidates generated by the state estimator. Once the accumulated constant mean segment length exceeds a certain threshold, the arbitrator performs a test to verify if the current *known* status is among the legal hypotheses. If so, the previous edge estimation result is confirmed and no further action is necessary. If not, the arbitrator searches the list of edge event classification candidates, archived by the edge estimator after the previous edge estimation. The arbitrator searches from the most favorable to the least favorable candidate, and examines if the resulting load status, assuming the candidate were true, is among the legal state hypothesis. If so, the candidate is selected, the previous edge estimation is nullified, the current load status is updated and the search is terminated. If the search terminates without producing a result, the arbitrator sets the initialization mode.

5.4 Test of load state estimator

In this section, we test the state estimator. The test data are obtained from a test building at 160 Sansome St., San Francisco, CA. The building has six constant loads. Table 5-1 shows the list of loads with the load identification numbers. The building has a varying background load mean, and the test program is written accordingly following the strategy outlined in the previous section. In that building, the background mean has the following bounds: μ_{0P}^{-} = 8 kW and μ_{0P}^{+} = 14.5 kW.

Table 5-1 Load names and identification numbers.

load id. no.	load name
1	Chilled water pump 1 (CHWP1)
2	Cooling tower pump 1 (CTWP1)
3	Cooling tower fan 1 (CTFan1)
4	Chilled water pump 2 (CHWP2)
5	Cooling tower pump 2 (CTWP2)
6	Cooling tower fan 2 (CTFan2)

Figure 5-4 shows an example of load events, where the CTWP2 is turned on and then turned off in about twenty seconds later. The original 120 Hz data are downsampled by eight by taking static averages before the analysis. Background load activities in the

beginning and around the CTWP2 OFF event are notable. All other major loads are off before the events. However, because the program has no knowledge about the load status in the beginning, it starts to generate state hypotheses and the hypotheses are evaluated at the first event.

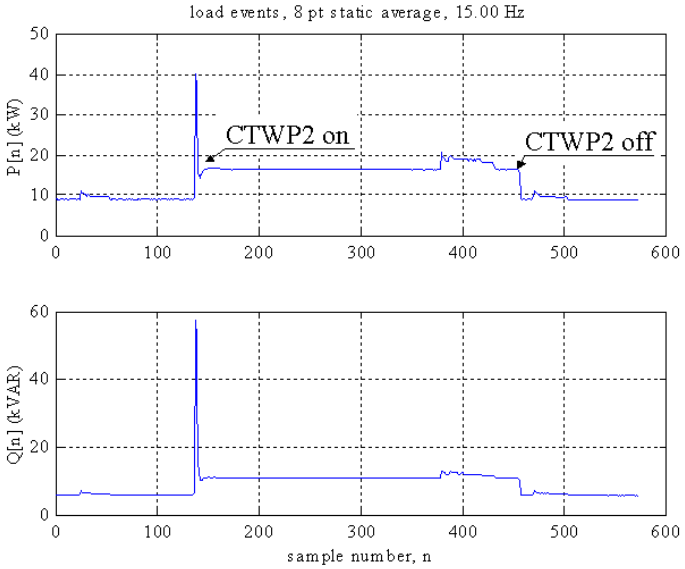


Figure 5-4 CTWP2 on and off events with unknown initial status.

Table 5-2 shows the state hypotheses forwarded to the edge estimator when the first event is evaluated. The table shows that only a single state hypothesis (all loads off, except the background load) was generated by the state estimator before the event.

Table 5-2 State hypotheses before the first event in Figure 5-4.

$\mathbf{v}^T$	ξ
[1 0 0 0 0 0 0]	-1.0427e+003

Table 5-3 List of post-event load status generated by the edge estimator after the first event in Figure 5-4.

$\mathbf{v}^T$
[1 0 0 0 0 1 0]
[1 0 1 0 0 0 0]
[1 1 0 0 0 0 0]
[1 0 0 0 1 0 0]

Table 5-3 shows the list of possible post-event load statuses generated by the edge estimator after the successful identification of the first event. The most probable status is listed on the top and the arbitrator accepts the recommendation. Now the pre-event load status is identified (no load ON) and the current load status is identified (only CTWP2 ON).

However, the arbitrator still has to check if the event classification result gives one of legal state hypotheses. After 928 points (at 120 Hz) are observed, the arbitrator runs a state check. The legal state hypotheses available at the time provided by the state estimator are shown in Table 5-4. The status resulted from the previous edge estimation is one of those legal hypotheses, although it is not the one with the minimum test statistic. Thus, the arbitrator confirms the previous edge estimation result and the current load status is completely determined.

Table 5-4 Legal state hypotheses after the evaluation of the first event in Figure 5-4.

$\mathbf{v}^T$	ξ
[1 1 0 0 0 0 0]	-9.3992e+002
[1 0 0 0 1 0 0]	-9.3957e+002
[1 0 0 0 0 1 0]	-9.2614e+002

Table 5-5 shows the load status history generated in the end by the program. It specifies the start and end index (at 15 Hz) and the ON/OFF status of loads at each constant load status series. By default, the load status is not assigned during transients.

Table 5-5 Load status history for the events in Figure 5-4.

start index	end index	CHWP1	CTWP1	CTFan1	CHWP2	CTWP2	CTFan2
35	135	OFF	OFF	OFF	OFF	OFF	OFF
147	455	OFF	OFF	OFF	OFF	ON	OFF
457		OFF	OFF	OFF	OFF	OFF	OFF

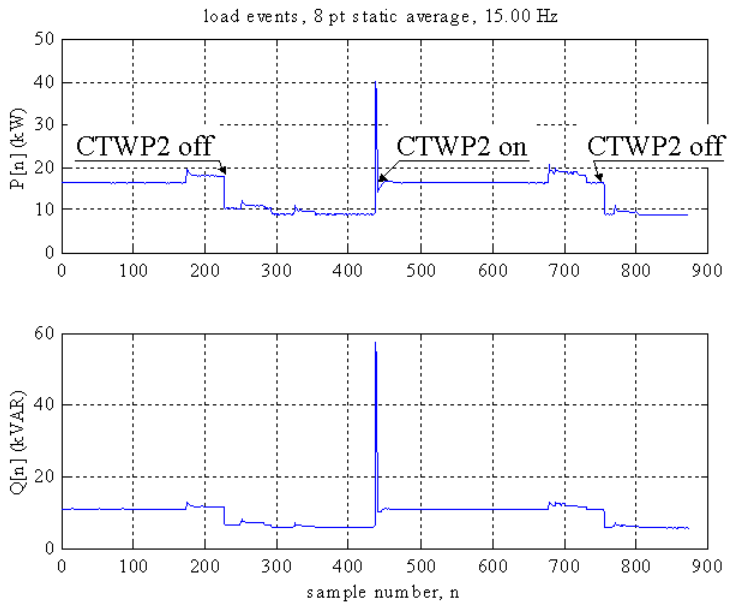


Figure 5-5 CTWP2 events with CTWP2 ON in the beginning.

Figure 5-5 shows another CTWP2 ON/OFF event set. However, in this case, the program is started when the CTWP2 is already on. The first event is the CTWP2 OFF event. Table

5-6 shows the state hypotheses available at the time of the first event classification. Because four pumps have similar mean levels and covariance matrices, each one of them shows up as the only load in each valid state hypothesis. Although the true hypothesis (fourth one) has the minimum test statistic, other hypotheses in the table also have test statistics close to the minimum. This example illustrates the difficulty in performing a state estimation based on state outputs only and the necessity to make a decision in conjunction with an edge event and its classification.

Table 5-6 State hypotheses before the first event in Figure 5-5.

$\mathbf{v}^T$	ξ
[1 1 0 0 0 0 0]	-1.0880e+003
[1 0 1 0 0 0 0]	-1.0687e+003
[1 0 0 0 1 0 0]	-1.0882e+003
[1 0 0 0 0 1 0]	-1.0885e+003

Table 5-7 Load status history for the events in Figure 5-5.

start index	End index	CHWP1	CTWP1	CTFan1	CHWP2	CTWP2	CTFan2
41	226	OFF	OFF	OFF	OFF	ON	OFF
228	435	OFF	OFF	OFF	OFF	OFF	OFF
468	755	OFF	OFF	OFF	OFF	ON	OFF
757		OFF	OFF	OFF	OFF	OFF	OFF

The edge estimator correctly identifies the CTWP2 OFF event based on the fourth hypothesis in the table and the arbitrator accepts the result. Subsequent edge event identifications can be performed with known load statuses. Table 5-7 shows the resulting load status history.

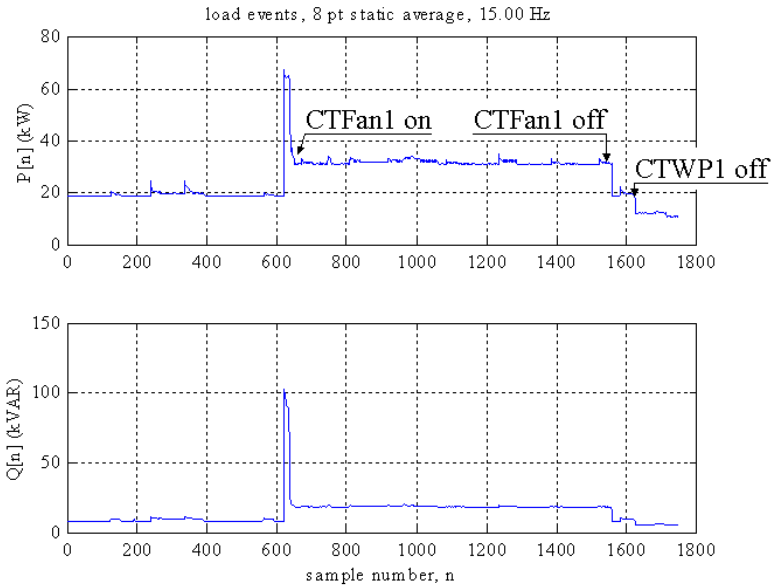


Figure 5-6 CTFan1 events with CTWP1 ON in the beginning.

Figure 5-6 shows another example of load events, where the CTWP1 is already on in the beginning and the CTFan1 is turned on and off on top of it. Table 5-8 shows the state hypothesis candidates at the time of the first event evaluation. The situation is similar to Table 5-6, where the true hypothesis (second one) has the smallest test statistic but with a very narrow margin.

Table 5-8 State hypotheses before the first event in Figure 5-6.

$\mathbf{v}^T$	ξ
[1 1 0 0 0 0 0]	-9.7627e+002
[1 0 1 0 0 0 0]	-9.9193e+002
[1 0 0 0 1 0 0]	-9.6583e+002
[1 0 0 0 0 1 0]	-9.8422e+002

Table 5-9 Load status history for the events in Figure 5-6.

start index	End index	CHWP1	CTWP1	CTFan1	CHWP2	CTWP2	CTFan2
41	618	OFF	ON	OFF	OFF	OFF	OFF
674	1558	OFF	ON	ON	OFF	OFF	OFF
1560	1623	OFF	ON	OFF	OFF	OFF	OFF
1625		OFF	OFF	OFF	OFF	OFF	OFF

In this case, every state hypothesis gives rise to the same load event of CTFan1 ON and, as a result, the state hypothesis with the minimum test statistic is selected as a valid pre-event load status. Table 5-9 shows the resulting load status history.

5.5 Chapter summary

A systematic method has been developed in this chapter to estimate the ON/OFF status of constant electric loads based on the observation of the harmonic power waveforms under a uniform load status. A list of state hypotheses is generated via the binary tree search in the beginning of a uniform load status series. Their ML test statistics are evaluated and updated with the accumulation of subsequent constant-mean segments. If the background load has a fixed mean, the hypotheses can be evaluated to identify the current load status, once the accumulated observation length exceeds a certain threshold.

However, in a more general situation, the background load can have a varying mean, assumed to be constant over a constant-mean segment. In this case, the arbitrator does not make a decision until a load event occurs. The edge estimator classifies the observed load event repeatedly based on each state hypothesis forwarded by the state estimator. The arbitrator selects the state hypothesis which gives rise to the best event classification result. The state is estimated in conjunction with a following load event.

Once a load event is successfully identified, the edge estimator performs the status estimation. However, the arbitrator occasionally runs a state check to verify that the result of a previous edge estimation gives one of current legal state hypotheses generated by the state estimator. If so, the previous edge estimation is confirmed. If not, the result of the

previous edge estimation is searched to find a viable candidate. If a candidate is found, the current load status is updated. If not, the arbitrator sets the initialization mode.

The overall load identification program is written and tested with the data collected in a real building, which has a variable background mean. The test result is quite satisfactory and shows that the overall algorithm is accurate and very robust.

Previous load identification approaches have relied on the limited aspects of edge estimation only. They were prone to frequent misclassifications, were vulnerable to noise spikes and disturbances, and lacked the capability to perform the identification in real time. The approach presented in this chapter utilizes both aspects of the power waveforms– state and edge– to better identify load status. The resulting algorithm is more accurate, reliable and robust, because it can tolerate the mistakes and failures by the edge estimator. It also does not require any prior load status information in beginning. It provides a complete solution to identify the status of constant electric loads in real time.

Chapter 6

Estimation of Variable Speed Drive Power Consumption from Higher Harmonic Powers

It is a difficult problem to estimate the real and reactive powers consumed by variable speed drives[9] (VSDs) from the non-intrusive power signal, especially in the presence of other constant electric loads. Variable speed drives continuously change their power consumption and thus do not have clearly defined steady states. Furthermore, their time-varying nature can interfere with the load status identifier, designed to keep track of the ON/OFF status of constant loads based on their steady-state power consumption levels. Hence, it is necessary to estimate the power consumed by VSDs, not only to trace their power usage but also to guarantee the functionality of the constant load status identifier.

Like any good detection and estimation routine in signal processing, the VSD power estimation has to rely on a certain signal model that best represents the characteristic of VSDs. In this chapter, a VSD power estimation method is proposed, which relies on the measurement of higher harmonic powers generated by a VSD and its distinctive relationship between the harmonic and fundamental powers. For simplicity, it is assumed that a single set of VSDs is present along with other constant loads.

6.1 VSD load model

Variable speed drives are widely used AC motor drives, especially in industrial and commercial applications, where single-speed operations can not meet the demands of various physical loads. Machine tools, fans, pumps and chillers are typical applications of VSDs. A VSD is an efficient way to control a motor when the connected load requires a wide range of speed or torque adjustment. A VSD typically consists of a rectifier, a direct current bus, and an inverter [Bose, 1993]. Figure 6-1 shows the topology of a VSD, along with a typical circuit model. Three-phase AC currents are converted to a DC via a rectifier and the DC current is inverted to AC currents (though may not be sinusoidal) by an inverter. The output frequency is adjusted by controlling the inverter timing.

The three-phase rectifier forms a bridge with six diodes as shown in Figure 6-1. The load is fed via a three-phase half-wave connection, and the return current path is another half wave connection to one of the three supply lines [Lander, 1987]. The output voltage is further rectified typically by an *LC* filter, before being fed to an inverter. Chokes (inductors) can be placed between the supply lines and the diodes to suppress the propagation of harmonics back to transmission lines.

At any moment, two diodes are on– one in the supply path and the other in the return path. For example, when the A phase supply diode is on, either the B or the C phase

[9] Variable speed drives are also called variable frequency drives (VFDs) and adjustable speed drives (ASDs). We will use them interchangeably.

return diode is on. Similarly, the A phase return diode is on when either the B or the C phase supply diode is on. As a result, the A (B or C) phase current typically shows four peaks– two in a positive swing and two in a negative swing– in a single cycle [Xu *et al*., 1999].

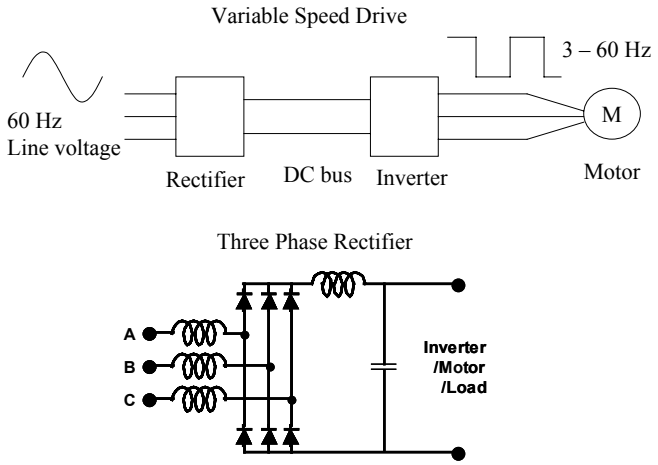


Figure 6-1 Topology and typical circuit model of variable speed drive.

We have already observed a representative VSD current waveform in Chapter 2 (Figure 2-5), which clearly shows four peaks per cycle. The harmonic powers computed from the waveform showed significant amounts of P_1 (or P), P_7, Q_1 (or Q), Q_5 and Q_7 (Figure 2-6). Thus, we observe that VSDs generate significant amounts of higher harmonics, especially the fifth and seventh.

Because constant linear loads do not generate harmonics[10], except real and reactive powers at the fundamental frequency (60 Hz), higher harmonic powers are free from the influences of the activities of constant linear loads. Thus, if a site mostly has constant linear loads in addition to a VSD, e.g. an HVAC room, higher harmonics may be used to keep track of the fundamental powers consumed by a VSD. This approach requires the knowledge of a certain relationship between the higher harmonic powers and the fundamental powers of a VSD.

Figure 6-2 shows VSD harmonic powers as functions of time. The data were collected at a test building at 160 Sansome St. San Francisco, CA. The building has two VSDs to drive a supply fan (100 hp) and a return fan (75 hp) for its main air circulation system. These fans are turned on and off at the same time and connected via the air duct system. Thus, these two VSDs can be conceived as a single unit for VSD load tracking purpose, because they are operated in tandem and serve the same physical system.

[10] However a constant linear load can change the relative phase relationship between its current and voltage waveforms. If there is no phase difference between the two waveforms, the load is pure resistive (Q = 0). If current waveform lags voltage waveform, the load is inductive (positive Q). If current leads voltage, the load is capacitive (negative Q).

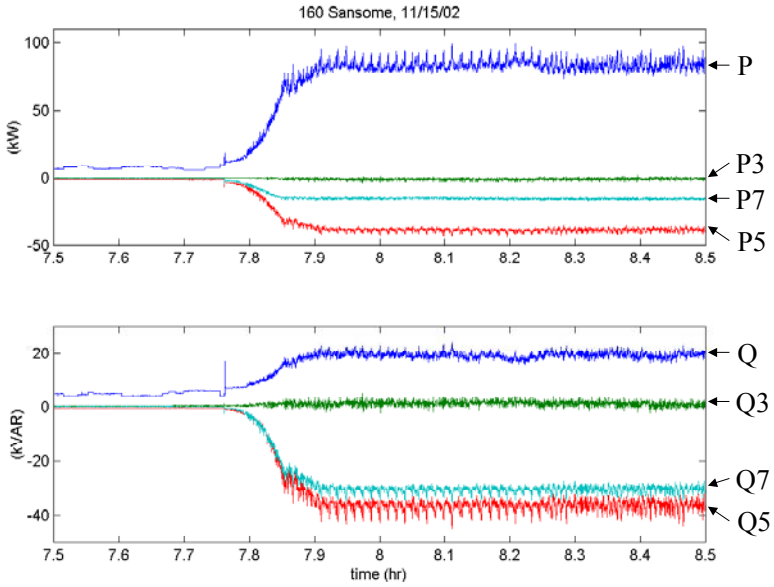


Figure 6-2 Harmonic powers consumed by a pair of VSDs.

The graph shows an early morning startup of the VSDs. To avoid a sudden pressure buildup in the air duct, VSD speeds are slowly increased over six minutes. A pressure loop controls fan capacities (i.e. speeds) afterwards. The graph clearly shows that there exist positive relationships between the fundamental powers and the higher harmonic powers of a VSD (or a set of VSDs). The third harmonic powers are essentially zero, although their noise variances increase with the magnitudes of the fundamental powers. There are almost constant biases in the fifth and seventh harmonic powers before the VSD set is turned on. These biases need to be subtracted from observations to measure the true harmonic powers generated by the VSDs. The biases of the fundamental powers are not constant because of the occasional activities of the background load that consists of light, plug and other small loads. However, their variations are relatively small, and thus can be subtracted by taking averages.

Like constant loads, the VSD load vector $\mathbf{s}_i[n]$ can be modeled as Gaussian, i.e. $\mathbf{s}_i[n] \sim N(\boldsymbol{\mu}_i[n], \boldsymbol{\Lambda}_i[n])$. A load vector $\mathbf{s}_i[n] = [P[n]\ Q[n]\ P_3[n]\ Q_3[n]\ P_5[n]\ \ldots\ ]^T$ is the harmonic power vector at discrete time n consumed by load i. A constant load has constant mean vector and covariance matrix. On the other hand, because VSDs do not have fixed steady states, their means and covariances change with time. However, to make the Gaussian VSD model practical for the purpose of tracking VSD fundamental powers, we have to impose certain restrictions in the time-varying nature of its mean and covariance.

It is reasonable to assume that the covariance matrix is proportional to the square of mean real power:

$$\boldsymbol{\Lambda}_i[n] \approx \left(\frac{\boldsymbol{\mu}_{i,1}[n]}{\boldsymbol{\mu}_{i,1}[n_r]} \right)^2 \boldsymbol{\Lambda}_i[n_r], \tag{6.1}$$

where $\boldsymbol{\mu}_{i,1}[n]$ is the first component (real fundamental power) of the mean vector at time n, and $\boldsymbol{\mu}_{i,1}[n_r]$ and $\boldsymbol{\Lambda}_i[n_r]$ are reference mean real power and covariance matrix taken at time n_r, respectively. The references are usually taken when VSD i is running under its most common operating conditions. Thus, the covariance matrix at time n is determined once the mean real power is known.

In estimating the mean power vector, we assume that there exist certain relations among its components, i.e.

$$\boldsymbol{\mu}_{i,j}[n] = f\left(\boldsymbol{\mu}_{i,1}[n], \ldots, \boldsymbol{\mu}_{i,j-1}[n], \boldsymbol{\mu}_{i,j+1}[n], \ldots\right). \tag{6.2}$$

However from the perspective of tracking fundamental powers using higher harmonics, we are interested in the relationships between the higher harmonic means (especially fifth and seventh) and the fundamental means. We further assume that a higher harmonic mean can be faithfully described by either the real or the reactive mean:

$$\boldsymbol{\mu}_{i,j}[n] \approx f\left(\boldsymbol{\mu}_{i,1}[n]\right) \approx g\left(\boldsymbol{\mu}_{i,2}[n]\right) \quad j \geq 5\,. \tag{6.3}$$

We basically are assuming that there exist strong correlations among mean harmonic powers so that one can be estimated reliably from another mean harmonic power. If the above functional relationships are one-to-one, the real and the reactive mean can be obtained from the measurement of a selected $\boldsymbol{\mu}_{i,j}[n]$, using inverse functions.

The mean of a harmonic power is typically measured by taking averages of the observed harmonic powers for a given duration. However if the data are biased, the time average is not a reliable estimate of true mean. Currently, the harmonic powers generated by the power coder are biased because of the phase bias. This phase bias is related to the uncertainty in finding the exact zero crossing of the fundamental voltage waveform during sampling and the amount of bias increases with the harmonic number, as explained in Chapter 2.

Currently, we obtain biased harmonic powers from the power coder. However, we still need an unbiased measure of a higher harmonic power to keep track of the fundamental powers. One such a measure is the kth apparent power A_k, defined in Chapter 2:

$$A_k = \sqrt{P_k^{\,2} + Q_k^{\,2}}\,. \tag{6.4}$$

A_k is simply the magnitude of the kth complex power, whose real part is the kth real power and imaginary part is the kth reactive power. A_1 has units of (VA). Because A_k is the magnitude of the DFT of the current waveform (with a scale factor), it is free from the time-shift of the reference voltage waveform.

With A_k selected as an unbiased measure, we need to establish the relationships between the higher harmonic apparent powers and the fundamental powers. As an example, Figure 6-3 and 6-4 show the correlation graphs between the fundamental powers and the fifth and seventh apparent powers. The data were collected at the 160 Sansome St. building on 12/05/02 in early morning. The graphs clearly show the positive relationships between the fundamental powers and the higher harmonic apparent powers. However the relationship is not linear and the slope of a graph decreases with increasing P or Q. Thus when the VSD set is running at its nominal operating condition, it tends to draw higher harmonics less aggressively. This can be conceived as a kind of saturation effect.

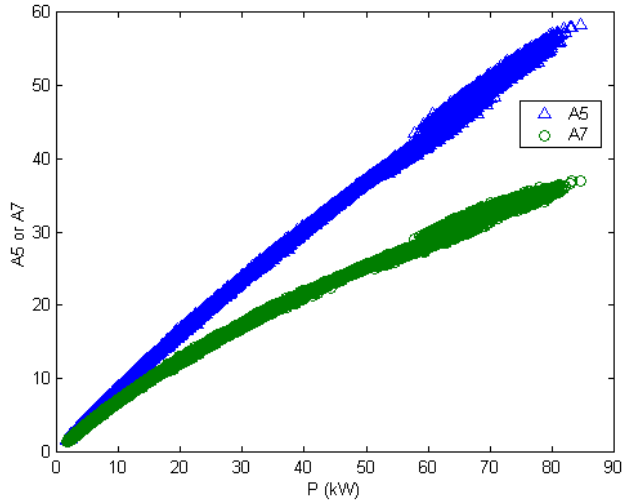


Figure 6-3 Correlation graphs between real power and selected apparent powers.

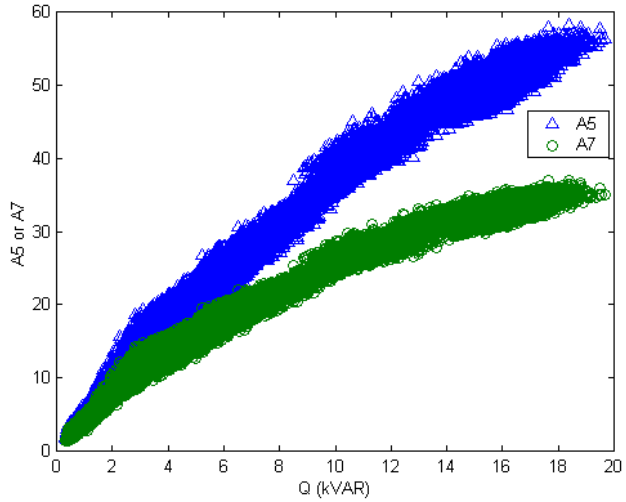


Figure 6-4 Correlation graphs between reactive power and selected apparent powers.

Figure 6-5 shows another set of correlations taken from the supply fan VSD of the Children's Court building in LA county. The data were taken on 11/24/02 after midnight, while other loads are off, by manually running the VSD at different speeds (as evidenced

by distinct clouds). The power signal from that building is quite noisy, but the overall correlation trend is similar. We note that the correlation of A_7 is relatively flat and will be less useful in tracking P, compared to that of A_5.

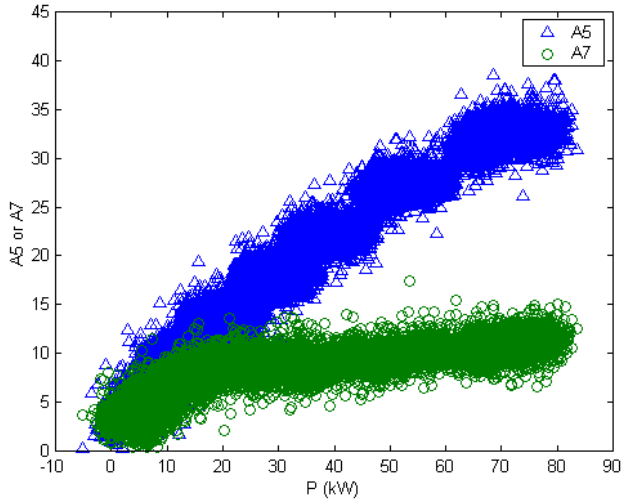


Figure 6-5 Correlation graphs between real power and apparent powers (LA building).

These correlations are complex results of how a VSD is constructed, what kind of physical load is attached to it (e.g. fan, duct and building space) and how the VSD is controlled. Theoretically, if we can accurately model these subsystems, we should be able to predict these correlations. However, overall physical system is often difficult to model and its exact modeling is unnecessary for our purpose. Here, we will obtain these relationships empirically from the measurements.

The nature of the physical load, i.e. the building air handling system, explains why we observe the saturation effect on the correlation graphs. For example, when the VSD is turned on in the morning, the building space is not pressurized and thus the VSD can easily induce airflow. However, as the space is being pressurized, the VSD has to work harder to maintain airflow, changing the condition at which the VSD is operated and hence the correlations between the fundamental powers and the higher harmonic apparent powers. This shows up as gradual changes in the correlation graph slopes. Especially, the correlation graphs of Q seem to have three or four distinct operation ranges.

Also, the time-varying nature of the physical load implies that the correlations are not fixed. If the details of the physical load change, they will create a different operating condition for a VSD and thus will change the correlations. The duct and building space will have different resistance, inertance and capacitance values depending on how air terminal boxes are regulated to meet varying heating, cooling and ventilation loads at different locations in a building. Thus the correlations, in general, will differ day by day and even vary slowly within a single day.

However, it also is true that there exist representative nominal operating conditions for VSDs, especially the ones used for HVAC systems, and the variations from these

representative correlations are typically small, i.e. 5 – 10 % range. Thus, we use the representative correlations obtained from measurements to keep track of VSD powers in an approximate sense, although true correlations changes continually.

Now, the task is to obtain parameterized estimations of the correlations. The estimations will be used to establish the inverse relationships of (6.3), in order to obtain fundamental mean powers from higher harmonic mean powers. Because the slope of a correlation graph gradually changes, we can not use a linear fit (straight line). A higher order polynomial may give an acceptable fit for the whole range, but the resulting function is difficult to invert. One viable approach is to use a piecewise continuous function to represent a correlation graph. This approach is appealing in the sense that a correlation graph consists of different regions of operation conditions. A linear fit within a region may be acceptable, but will produce a poor result unless each region is sufficiently narrow. A better alternative is to use a power function with a coefficient and an exponent. Piecewise power functions provide good fits to observed correlation graphs and three or four different regions are typically good enough to represent the whole range.

For simplicity, we denote the kth apparent, real and reactive mean power as A_k, P_k and Q_k, respectively. Consider a correlation between P and A_k for a give region; then A_k can be represented as a power function of P:

$$A_k = f(P) = aP^b, \tag{6.5}$$

where a is the coefficient and b is the exponent of the power function, respectively. If we take logarithm on both sides,

$$\ln A_k = \ln a + b \ln P. \tag{6.6}$$

Thus it is a linear function in terms of the logarithms of powers.

Suppose that we are interested in estimating a and b from a length-N observation. Let $\mathbf{y} = \left[\ln A_k[0] \;\; \ln A_k[1] \;\; \ldots \;\; \ln A_k[N-1]\right]^T$ and $\mathbf{x} = \left[\ln P[0] \;\; \ln P[1] \;\; \ldots \;\; \ln P[N-1]\right]^T$. Then,

$$\mathbf{y} \approx \ln a + b\mathbf{x} = \mathbf{H}\mathbf{a}, \tag{6.7}$$

where $\mathbf{H} = [\mathbf{1}\ \mathbf{x}]$ and $\mathbf{a} = [\ln a \; b]^T$. The estimation of $\mathbf{a}$ can be obtained easily using the method of least squares (LS):

$$\hat{\mathbf{a}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}. \tag{6.8}$$

The procedure can be repeated for different regions to obtain the set of parameters to describe the correlation between A_k and P for the whole range. This procedure can also be repeated for different apparent powers and for the reactive power Q.

Figure 6-6 shows the parameter estimation result of a region for the correlations shown in Figure 6-3. The resulting parameterized correlations are shown in solid lines along with the estimated parameters. The lower graph is in a logarithmic scale and the data points are almost perfectly linear, demonstrating that the power function model is a reasonable choice.

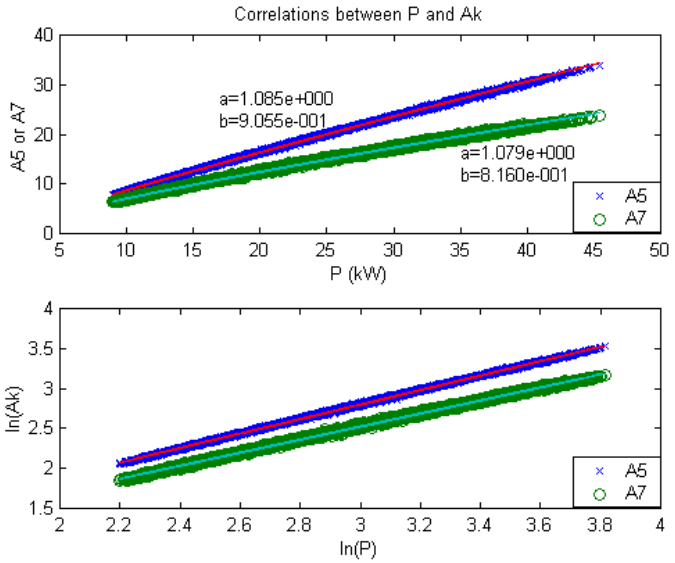


Figure 6-6 Parameter estimation for one of the regions shown in Figure 6-3.

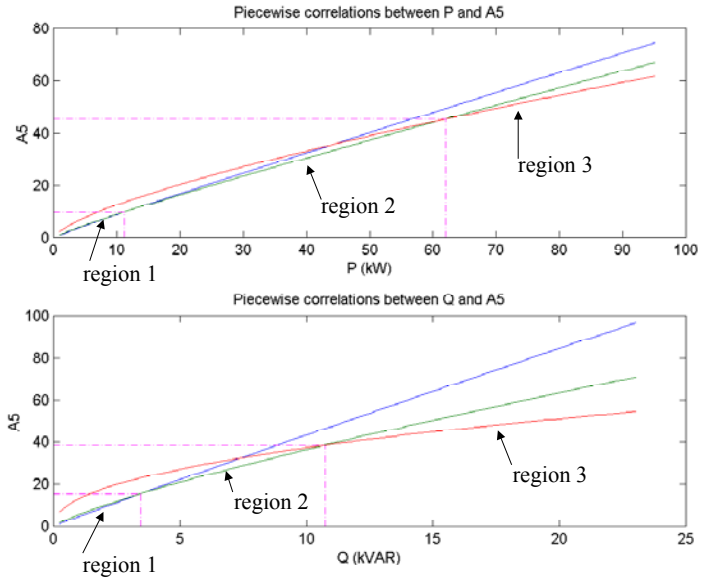


Figure 6-7 Piecewise correlations between fundamental powers and A_5.

In this manner, correlations can be estimated for each region. These correlations can be plotted over the whole range of P or Q, to obtain a piecewise continuous correlation curve. Figure 6-7 shows an example of the correlations between P and A_5 and between Q and A_5, plotted for the whole ranges. The boundary between neighboring regions is

determined as where two correlation curves meet. The regional boundaries are shown on the graphs as the vertical and horizontal dash-dot lines. At each region, the innermost curve is selected as the regional correlation. Connecting these regional correlations yields a piecewise continuous overall correlation curve. The overall correlation is one-to-one and thus invertible.

When a mean estimation of A_5 is given, an appropriate region is first selected based on its numerical value. Then, the corresponding mean value of P or Q is computed using the coefficient and exponent obtained from a lookup table based on the selected region.

6.2 VSD power estimation method

A simple VSD load model was developed in the previous section for the purpose of estimating VSD fundamental powers from the measurement of higher harmonic powers. In this section, we develop a systematic methodology to estimate VSD powers from a continuous stream of data and to condition the fundamental powers for further load status analysis. The overall schematic is shown in Figure 6-8.

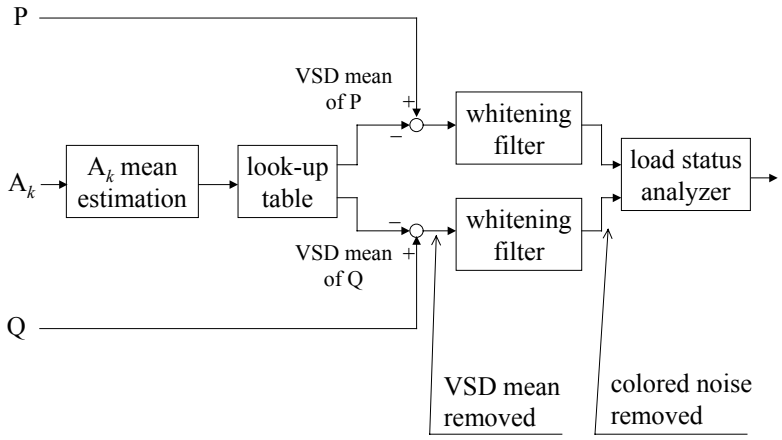


Figure 6-8 Schematic of VSD power estimation and signal conditioning.

The mean of a selected higher harmonic apparent power is first estimated. It is then fed to the look-up table that yields the estimations of the mean fundamental powers consumed by a set of VSDs, based on the piecewise continuous correlations between the apparent power and the fundamental powers. The VSD mean estimations are subtracted from the observations of P and Q, for further load status analysis, i.e. to identify the ON/OFF status of other constant loads. With the VSD mean subtracted, P and Q goes through whitening filters, because VSDs typically generate colored noise due to their mechanical load imbalances and the slow rotation speeds of their end loads. The filtered fundamental powers are free from the undulations caused by the time-varying VSD mean, and free from the colored noise. However, any white noise added by the VSD is retained and

accounted for by the load status analyzer. The load status analyzer identifies the load status of other constant loads by detecting and classifying load (ON/OFF) events, and by estimating the load states from the stream of fundamental powers, if applicable.

It also is possible to use more than one apparent power to estimate the fundamental powers consumed by a VSD. For example, we may use both A_5 and A_7 to estimate fundamental powers separately and take averages to obtain the final estimation results. However, in this chapter, we use A_5 only for simplicity.

Because the mean of A_k has to be estimated, the scheme has to be performed for a window (or block) of data. A time-window of data is first collected, and the whole chain of VSD mean power estimation and signal conditioning is performed. Then the process is repeated for the next window of data. Depending on the circumstances, there may or may not be an overlap between two neighboring windows.

The window size and the mean estimation method for A_k are determined, in most part, by the nature of the signal. Figure 6-9 shows the time graphs of P and A_5, collected at 120 Hz, from the 160 Sansome building on 12/04/02. The graphs show the undulating nature of P, and that the undulation of P is readily reflected by A_5. The undulation has a weak sinusoidal nature (with a period of approximately 30 seconds), but its frequency and phase are not fixed but change with time. The oscillation of time-wise power consumption is related to the nature of the air handling system and how it is controlled. Appendix B explains the observation as a *limit cycle* [Gibson, 1963].

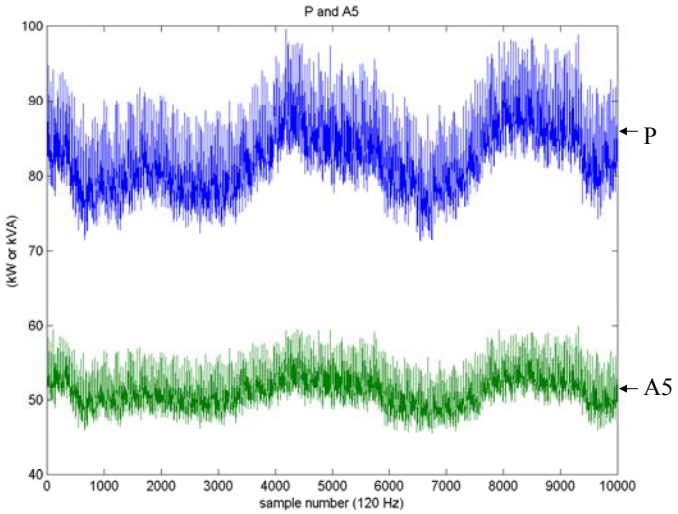


Figure 6-9 Time graphs of P and A_5.

Estimating the mean of A_5 as a sinusoidal function is, in general, not applicable because its sinusoidal nature changes even within a cycle, making the faithful estimation of sinusoidal parameters almost impossible. Another option is to select a reasonable size window, i.e. $N = 240 \sim 360$, and assume that the mean is described by a polynomial function of time indices. A higher-order polynomial will give a better fit for a given

window of data, but associated computational cost will be higher. Also, too high a polynomial order often results in an over-fit. In this chapter, we select the polynomial of order one (line) to describe the mean of A_5 due to its simplicity and efficiency. The first order polynomial mean estimator is described in Appendix C.

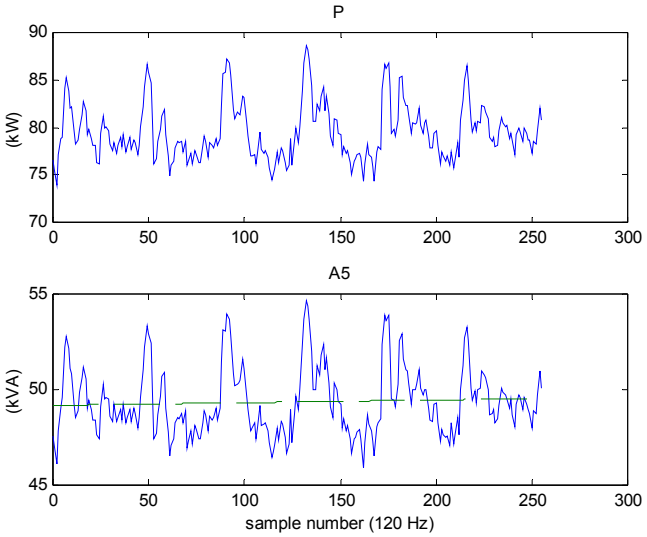


Figure 6-10 Detailed view of P and A_5.

Figure 6-10 shows a detailed view of P and A_5 with the window size $N = 256$, taken at $n = 7000$ from Figure 6-9. The bottom figure also shows the result of the first order mean estimation for A_5 as a dashed line. The graphs show large peaks and they appear to have a periodic nature. An interesting observation is that these noise peaks appear in phase with the same periodicity in both P and A_5. We have already seen that P and A_5 pick up the large time scale undulation (~30 sec) in unison. We notice that they also exhibit these small time scale noise peaks (~ 3 Hz) in phase. However, they differ in amplitudes.

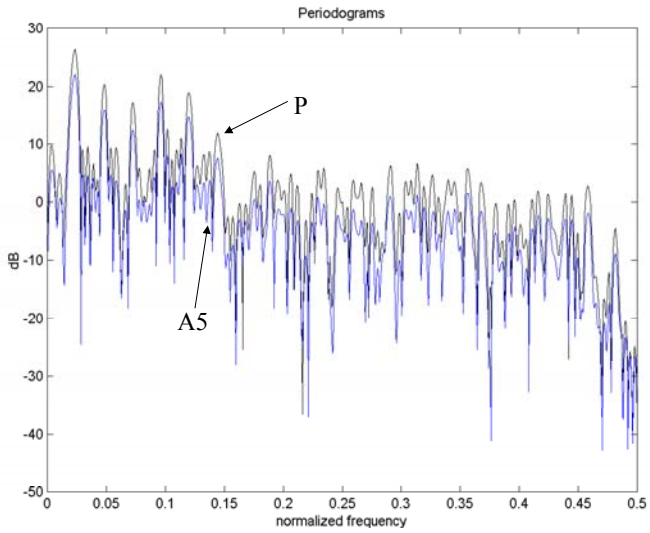


Figure 6-11 Periodograms of signals in Figure 6-10.

Figure 6-11 shows the periodograms[11] of the Figure 6-10 signals in a logarithmic magnitude[12] scale. The frequency is normalized so that the unit normalized frequency is equal to the sampling frequency (120 Hz). Both periodograms show large spectral peaks, especially at low frequencies. A closer examination of the graph reveals that the locations of the spectral peaks are the same for both P and A_5. Furthermore, the spectral peaks of P are taller than those of A_5 almost by a constant amount in the logarithmic scale, which corresponds to a multiplicative scaling factor in magnitudes. Thus, we conclude that P and A_5 have almost identical noise spectral distributions, which are different only in magnitudes by a constant scaling factor.

These spectral components are often called interharmonics, or subharmonics, because their frequencies are not the integer multiples of the fundamental frequency, i.e. 60 Hz [Yacamini, 1996]. The interharmonics are typically caused either by a periodically varying load or a frequency difference between two AC systems when they are connected by a DC link [Li *et al*., 2003]. The VSD interharmonics can be caused either by a mechanical load imbalance (periodic load) or by a DC ripple due to the frequency mismatch between the rectifier and the inverter.

If an interharmonic is generated by a frequency mismatch between the rectifier and inverter set, its frequency f_i is given by the following formula [Gunther, 2002]:

$$f_i = (p_1 m \pm 1) f \pm p_2 n f_0 , \tag{6.9}$$

where f is the fundamental frequency of the rectifier supply voltage, f_0 is the inverter output frequency, p_1 and p_2 are the pulse numbers of the rectifier and the inverter, respectively, and m and n are integers. The first term is the harmonic frequency and the second term is the frequency of the DC ripple, proportional to the output frequency. The interharmonic frequency can be considered as a modulation of the harmonic frequency by the ripple frequency.

When the interharmonic frequency is in the vicinity of a harmonic frequency, it will cause a periodic change of the amplitude of voltage or current waveform, or *beating*. This phenomenon is commonly known as *flicker*, because it is believed to make the fluorescent lights flicker. The beating frequency f_b is given by the following equation:

$$f_b = \left| f_i - f_k \right| , \tag{6.10}$$

where $f_k = kf$ is a harmonic frequency close to f_i. If we insert (6.9) into (6.10) and rearrange terms:

$$f_b = \left| (p_1 m \pm 1 - k) f \pm p_2 n f_0 \right| . \tag{6.11}$$

[11] The periodogram is defined in Chapter 7.
[12] The logarithmic magnitude of x is defined as $10 \log_{10} \|x\|$ (decibel, dB).

We note that the first term is a harmonic frequency and the second term varies linearly with the output frequency. The beating frequency will vary linearly with the output frequency but its range will be very short, because the beating occurs when two frequencies are sufficiently close to each other. Thus, if the output frequency is increased from zero, the beating should intermittently show up as the second term of (6.11) nears one of the harmonic frequencies. Also, the beating frequency should decrease and increase as the second term passes by a harmonic frequency.

Figure 6-12 shows the DFT magnitude plot of the 160 Sansome VSD set start-up in a morning. A rectangular window of size 1024 is applied on the time series data of P to compute the DFT. Then, the window is shifted by 2000 points (16.7 sec) to compute the next DFT. The graph clearly shows that the (first) spectral peak, or beating frequency, increases almost linearly with the ramp-up of the VSD set. Thus, it is more likely that the interharmonics are caused by the mechanical load imbalance of the fan output shaft(s). The load imbalance causes a periodic pulsation in the VSD power consumption and its frequency is directly proportional to the shaft rotation frequency.

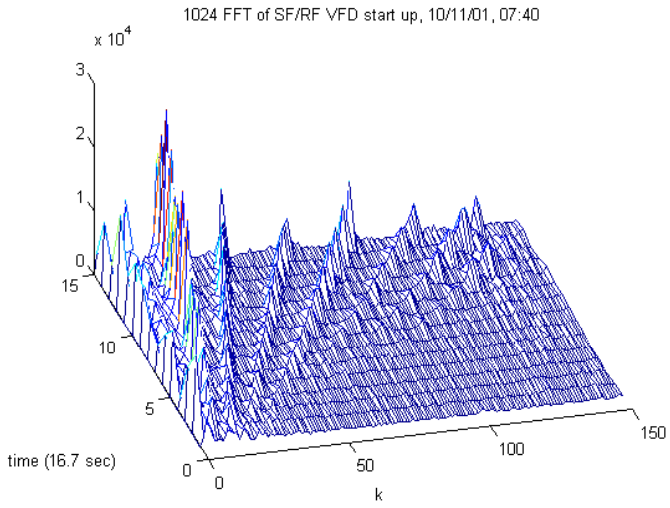


Figure 6-12 DFT magnitude plot of VSD ramp up.

The locations of the spectral peaks suggest that the interharmonics, or colored noise, have a harmonic nature, i.e. the locations are integer multiples of a fundamental frequency. The harmonic nature is especially evident from the first six peaks at low frequency. The lowest peak occurs at approximately 3 Hz and next peaks occur with a 3 Hz interval. Because the maximum speed of a VSD is often the nominal speed of an induction motor (30 Hz for a double-pole pair motor) and because there typically exists a 10:1 speed reduction between a motor and an air handling fan via a pulley-belt system, the speed of 3 Hz is approximately the rotational speed of a (centrifugal) fan. Thus, it is more plausible to speculate that the colored harmonic noise is originated by the mechanical load imbalance of an output shaft and its harmonic components are generated via a type of resonance mechanism through the pulley-belt system (the pulley is mostly inertive (L) while the belt has a compliance component (C)).

If the colored noise is generated by a load imbalance, the estimation of its magnitude can be used to quantify its severity. This information can be useful for a diagnostic purpose, i.e. to monitor the state of a fan drive system and take a maintenance action when necessary. Another implication is that the location of the first noise peak indicates the current speed of a fan. Thus, if we can reliably estimate the fundamental frequency of the harmonic noise, we can monitor the rotational speed of the fan. In other words, we can have a *virtual tachometer*. We expect the location of the first peak (or spacing between peaks) on the spectrum to change with time, reflecting the change in fan speed.

Although the harmonic colored noise can be useful for diagnosis and speed monitoring, its large magnitude can interfere with the load status identification system. For example, its large peak can confuse the load event detector, which relies on a clear sharp change of power levels to detect a load event. Also, the (constant) load status analyzer expects white noise, and thus can not guarantee optimality in its decision when colored noise is present. White filtering is thus necessary for both P and Q even after their VSD means are removed.

There can be two approaches in eliminating the zero-mean additive signal generated by a VSD (set). One approach is to treat the colored noise as a signal of interest and try to estimate the signal. Then, the estimated signal is subtracted from the original observation, and the difference is the filtered signal. This approach is useful when we are interested in using the estimated harmonic signal for diagnosis and speed monitoring. The other approach is to treat the harmonic signal simply as a colored noise to be removed by passing through a filter. The construction of an optimal whitening filter is closely related to the estimation of the power spectral density (PSD) of the signal.

In this chapter, we assume the presence of an ideal whitening filter to filter the colored noise of VSDs. The estimation and elimination of the zero-mean additive harmonic signal generated by a VSD will be fully discussed in the next chapter.

6.3 Test of VSD power estimation method

As a test, the overall VSD mean power estimation and signal conditioning scheme is performed for a span of data, collected at 120 Hz, from the 160 Sansome building on 12/04/02. The span consists of 30 non-overlapping rectangular windows each with the size of $N = 256$.

Figure 6-13 shows the time waveform of A_5 (upper graph). It also shows the first order mean estimation of A_5 as a solid line within the graph. Because the mean estimation is piecewise continuous across windows but linear within a single window, it is not as smooth as the natural undulation of A_5. It nevertheless represents the time varying nature of the A_5 mean reasonably well. The bottom graph shows the time waveform of A_5 after it is white-filtered. The output of a whitening filter is often called *innovations* (*process*).

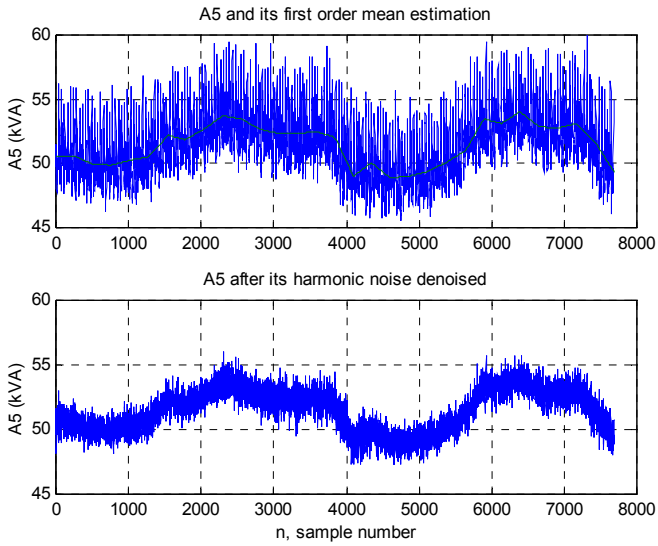


Figure 6-13 Time graph of A_5 with its first order mean estimation and innovations.

Figure 6-14 shows the time waveform of P (upper graph). Its VSD mean estimation from the A_5 mean estimation via the look-up table is also shown as a solid line within the graph. If we are interested in estimating the mean real power consumed by a VSD (set), the proposed scheme can provide the estimation as a function of time. The bottom graph shows the innovations of P after its VSD mean is subtracted and the difference is passed through a whitening filter.

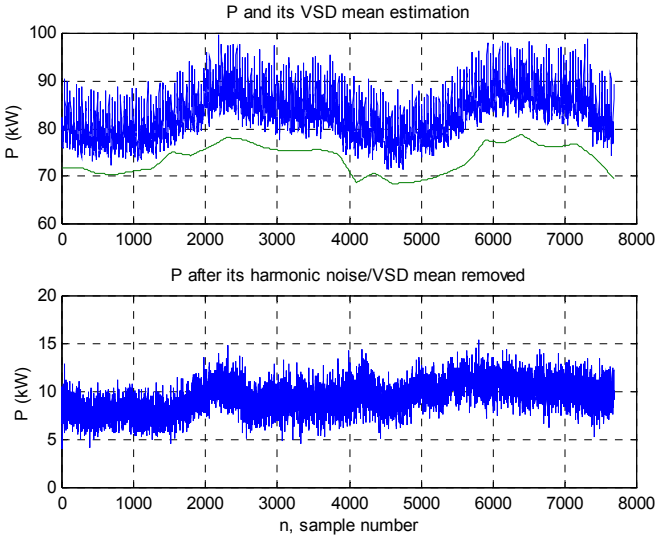


Figure 6-14 Time graph of P with its VSD mean estimation and innovations.

Ideally, we expect flat innovations of P when other constant loads remain unchanged. However, the innovations still show a slight undulation carried over from the original waveform. The VSD mean power estimation from the mean estimation of a higher harmonic apparent power is not fully accurate, because both means do not increase by a constant scaling factor. However, on average, the overall scheme performs reasonably

well in removing the effect of a VSD (set) from P, both the mean and the colored noise of the VSD. Also, on that day, no other constant load was on except background loads, which consist of small loads with random ON/OFF events. Thus, the innovations are the real power collectively consumed by these background loads, and it is possible that they have a few activities within the span of the data we observe.

Figure 6-15 shows the time waveform of Q (upper graph). Its VSD mean estimation from the A_5 mean estimation is also shown as a solid line. The bottom graph shows the innovations of Q as before. Here again, the VSD mean estimation from A_5 does not remove completely the undulation in the original waveform, although the result can still be used for the load status analysis purpose. The innovations of Q show more undulations then the innovations of P do, mainly because Q generally has a smaller magnitude than P. Its undulation is not as aggressive as that of P and when compensated by the mean estimation from A_5, whose undulation is more aggressive than that of Q, the result is a slight over-compensation.

Another possibility is the phase bias of harmonic powers. As explained in chapter 2, the fundamental power has a phase bias of 2.81 (degree). This is a relatively small angle, which may be ignored. However, when one component (in this case the reactive power or imaginary part) has a substantially smaller magnitude than the other, the component with a smaller magnitude suffers more from the phase instability. For example, if P = 80 kW and Q = 25 kVAR, with a 2.81 degree phase uncertainty Q can change by 16 % of its nominal value while P changes only by 1.5 %.

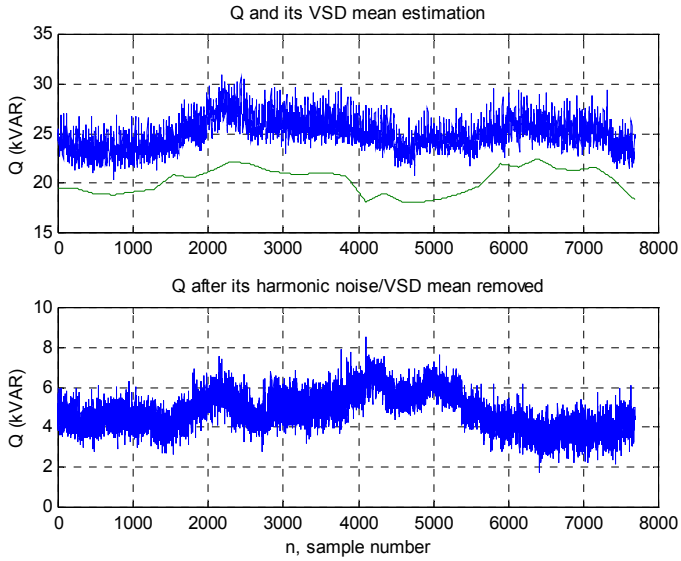


Figure 6-15 Time graph of Q with its VSD mean estimation and innovations.

Figure 6-16 shows the standard deviations of the white filtered signals (innovations). Although certain amounts of fluctuations exist, the standard deviation for a signal is relatively stable, in part due to the effectiveness of the white filtering. The white noise characteristic of the VSD can be estimated from the innovations.

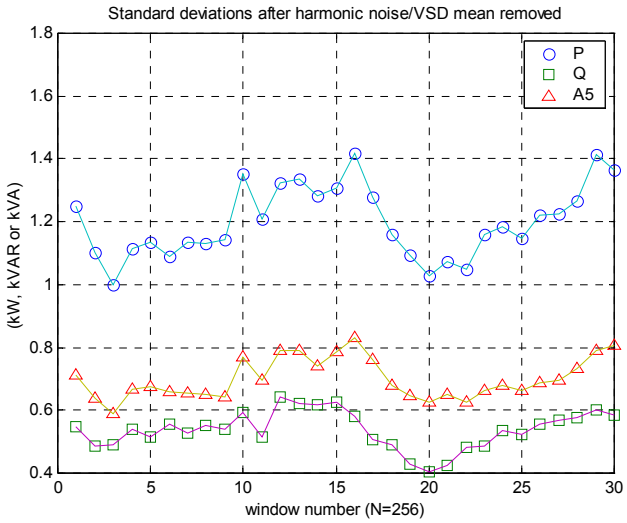


Figure 6-16 Standard deviations of innovations for each window.

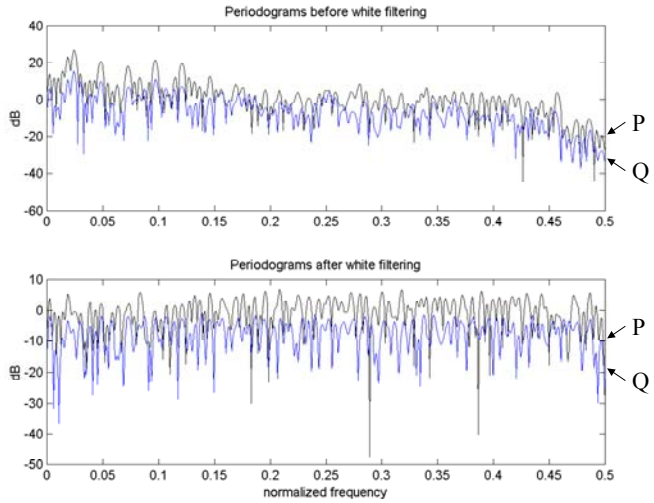


Figure 6-17 Periodograms of *P* and *Q* before and after white filtering.

Figure 6-17 shows the periodograms of *P* and *Q* before and after white filtering for a window of data (second window of the whole span). The top graph shows the periodograms of *P* and *Q* before filtering. Although the locations of spectral peaks are the same (i.e. same fundamental and harmonic frequencies), the heights of the spectral peaks are largely different between *P* and *Q*. *Q* in general has smaller spectral peaks and relatively higher peaks at mid-frequency range. Apparently *Q* is less noisy than *P*. The bottom graph shows the periodograms after white filtering. Large spectral peaks disappeared, and the periodograms are relatively flat for the overall frequency range. Thus, we verify the effectiveness of the white filtering. The standard deviations of the innovations (PSD magnitude after white filtering) are the inherent properties of the VSD white noise.

6.4 Chapter summary

A practical VSD power estimation method has been developed in this chapter, which relies on the mean estimation of a higher harmonic apparent power and the correlation between the apparent power and the fundamental powers. Certain higher harmonic powers are suitable for VSD load tracking, because there exist strong correlations between them and the fundamental powers consumed by a VSD and because they are not masked by the activities of constant linear loads. Thus, the real and reactive power consumed by a VSD can be estimated reliably as a function of time, even in the presence of the activities of other constant linear loads.

A Gaussian random process model is proposed for a VSD with a time-varying mean vector and a time-varying covariance matrix. The model is simplified by assuming that the covariance matrix is proportional to the square of real power and that there exist strong correlations between mean harmonic powers. The correlations were further reduced so that there exists a one-to-one relationship between a higher harmonic mean and a fundamental mean. The apparent power is introduced because of the phase bias of the power coder.

The actual correlations were obtained from the observations and the piecewise continuous power function model was adopted to represent a correlation curve over the whole range of P or Q. A power function becomes linear on a logarithmic scale and the method of least squares is applied to estimate its parameters. The overall correlation curve can be generated from these parameters, and the region boundaries are recorded in a look-up table so that appropriate parameters can be selected based on the mean estimation of a higher harmonic apparent power. The mean of a selected apparent power is estimated as a first order polynomial.

The nature of the colored noise generated by a VSD was discussed. The harmonic noise generated by the VSDs at the 160 Sansome building was attributed to the load imbalance at the fan shaft(s). Due to its physical origin, the magnitude of the colored noise can be used to monitor the severity of the imbalance and its fundamental frequency can be used to monitor the rotational speed. The strategy to estimate the harmonic signal was discussed and the whitening filter was introduced to eliminate the colored noise. The VSD mean and colored noise are removed from the real and reactive powers before they are examined by a load status analyzer. The white noise generated by a VSD still remains in them and the analyzer has to take it into account in analysis.

The test results show the effectiveness of the VSD power estimation method based on the observation of higher harmonic powers, thought the method is not fully accurate and leaves many rooms for further refinements. Nonetheless, it is a great achievement in the sense that no VSD power estimation method has existed unless a dedicated power meter is installed. Once this method is established and adopted, subsequent improvements and better alternatives will naturally follow.

Chapter 7

Estimation and Elimination of Harmonic Signals from Power Waveforms

While any time-varying load can generate non-white noise, variable speed drives (VSDs) are the most common source of colored noise on power waveforms. The colored noise of a VSD has a periodic nature and its spectral peaks are often harmonically related to their fundamental frequency proportional to the rotational speed of the VSD. Furthermore, the colored noise generally has a time-varying nature, i.e. the magnitudes of its spectral peaks and its fundamental frequency change with time. It is desirable to estimate the harmonic signal generated by a VSD to monitor the drive. It also is necessary to suppress the colored noise generated by a VSD for the functionality of a constant load status identifier, which relies on the white noise model of the power signal generated by constant loads. In this chapter, various techniques are introduced to estimate and eliminate the harmonic signals generated by VSDs.

7.1 Introduction

Identifying load status is the most basic task of an electric load information system based on the NIPM and provides a necessary basis to perform further diagnosis and optimization study. Analyzing the status of loads has to rely on a certain signal model, so that the current observation of harmonic powers can be represented as a sum of individual signal sources, each of which describes the behavior of a specific load. The Gaussian random process model was introduced for constant electric loads in earlier chapters, in which a load generates harmonic powers while it is on, following a Gaussian distribution with a fixed mean vector and a constant covariance matrix.

A VSD can also be modeled as a (Gaussian) random process, but with a time-varying mean and a time-varying covariance matrix. The elimination of the time-varying mean was discussed in the previous chapter. Due to the colored noise generated by a VSD, the Gaussian model can not be applied in a strict sense, because it implicitly assumes that the noise is white. Nonetheless, the Gaussian random process model is still valid if we model the VSD as a sum of two signal sources– a Gaussian source with a white noise and a colored noise source– and if we can eliminate the latter. *White filtering* refers to the procedure of eliminating the additive colored noise of a signal and restoring its white noise content. White filtering is necessary not only to apply a Gaussian model for a VSD but also to guarantee the functionality of the constant load status identifier, which relies on the Gaussian model and the whiteness of the observations. In this chapter, we are interested in developing a whitening filter for the colored noise generated by a VSD.

The colored noise of a VSD is often be related to the load imbalance of its end shaft. Here, the imbalance can be due to the rotor itself, e.g. shaft misalignment, uneven distribution of rotor inertia, uneven friction, etc., and/or to the medium impelled by the rotor, e.g. air in a fan, refrigerant in a compressor, etc. This implies that the colored noise

reflects the nature of this load imbalance and that by estimating the colored noise we can analyze its characteristic. For example, the severity of the imbalance is reflected by the magnitude of the colored noise, and by constantly monitoring the magnitude we can diagnose whether the mechanical drive of a VSD is in a good shape.

Due to its origin, the colored noise of a VSD has a periodic nature, with its frequency proportional to the rotational speed of its end shaft. Furthermore, the colored noise often has a harmonic nature amplified through a resonance mechanism. The PSD of the colored noise often has several harmonically related spectral peaks with their fundamental frequency equal to the rotational speed of the end device (fan, compressor, etc.). This implies that if we can estimate the fundamental frequency of the harmonic signal generated by a VSD, we can monitor the rotational speed of the end shaft. In other words, we can build a *virtual tachometer* solely from the non-intrusive power signals. This can be helpful when one tries to build a fault detection and diagnosis system for a VSD, in which the rotor speed is constantly monitored and compared to a command input, if available, or to a nominal operational speed range. Any discrepancy is reported as a fault.

It may be unfair to treat the harmonic signal generated by a VSD simply as a colored noise due to its interesting nature, i.e. the mechanical load imbalance. Thus, we may treat the harmonic signal as a signal of interest and estimate the signal. The estimation can be further analyzed for further monitoring and diagnosis. Nevertheless, the goal of filtering the colored noise can still be achieved by subtracting the estimation from the observation. A good estimation routine can perform the dual role of signal extraction and elimination. In this chapter, we will develop several methods to estimate the harmonic signal generated by a VSD.

Colored noise is not necessarily an exclusive byproduct of a VSD, but can be produced by any motor load that has a dynamic (rotational) component. However, we rarely observe a colored noise from a constant motor load. A single-speed induction motor typically runs at 60 Hz or 30 Hz depending on the number of its pole pairs. Even if there exists a certain load imbalance, this is too high a speed to be captured by the 120 Hz power coder unless there is a significant gear reduction ratio. The resulting noise will be more white than colored. Also, small single-speed induction motors are typically used to pump liquid (mainly water) through a closed pipe loop, whereas large induction motors powered by VSDs impart movement to gas medium (air in fan, refrigerant in compressor, etc.). The liquid has more inertia and viscosity than the gas, which in turn helps to stabilize the rotor. This in part explains why we do not observe colored noise from constant motor loads, whose majority are small water pumps.

This chapter first examines the method of detecting and estimating harmonic signals from periodograms. Then, its two varieties are proposed. We also develop a white filtering method by treating the harmonic signal simply as a colored noise. The pros and cons of these methods will be discussed at the end of the chapter.

7.2 Detection and estimation of harmonic signal from periodogram

To estimate a harmonic signal, we first have to *detect* its presence. Running a harmonic signal estimation routine, when there is no such a significant signal, is a waste of resource and further the routine will either fail or return a questionable result. Thus, a harmonic signal detector is a prerequisite to estimate harmonic signals reliably and accurately. Also, the development of a detector often gives rise to its accompanying estimator.

The estimation of a harmonic signal hinges on the estimation of its fundamental frequency (or period). The estimation of other components, such as amplitudes and phases, are usually straightforward once the fundamental frequency is available. Estimating the fundamental frequency of a harmonic signal is closely related to estimating its power spectral density (PSD).

When $x[n]$ is a real discrete-time wide-sense stationary (WSS) random process, its PSD $P_{xx}(\omega)$ is defined by the Wiener-Khinchine theorem [Papoulis and Pillai, 2002]:

$$P_{xx}(\omega) = \sum_{k=-\infty}^{\infty} r_{xx}[k] \exp(-j\omega k) \quad -\pi \le \omega \le \pi \,, \tag{7.1}$$

where $r_{xx}[k]$ is the autocorrelation of $x[n]$ defined as,

$$r_{xx}[k] = E(x[n]x[n-k]) = E(x[n+k]x[n]) \,. \tag{7.2}$$

Note that $P_{xx}(\omega)$ is simply the discrete-time Fourier transform (DTFT) of $r_{xx}[k]$ and that it is a nonrandom continuous function of ω. The PSD can be interpreted as the average power of $x[n]$ in the narrow frequency band from ω to $\omega + d\omega$ [Kay, 1988]. However, the analytic evaluation of $P_{xx}(\omega)$ requires the complete knowledge of the probability density function (PDF) of $x[n]$, which is generally unknown. Therefore, we often have to rely on the estimations of the PSD, from a given finite length observation.

The periodogram, $I_{xx}(\omega)$, of an N point sequence $x[n]$ is defined as

$$I_{xx}(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] \exp(-j\omega n) \right|^2 = \frac{1}{N} \left| X(\omega) \right|^2 . \tag{7.3}$$

Note that the periodogram is a random continuous function of ω and that it is the squared magnitude of the DTFT of the signal, normalized by the signal length. In practice, the periodogram is computed via the discrete Fourier transform (DFT). In this case, only samples of $I_{xx}(\omega)$ at $\omega_k = 2\pi k/N$ are evaluated. The periodogram can also be defined as [Marple, 1987]

$$I_{xx}(\omega) = \sum_{m=-(N-1)}^{N-1} \rho_{xx}[m] \exp(-j\omega m), \quad \text{where } \rho_{xx}[m] = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x[n]\, x[n+|m|] \tag{7.4}$$

The sample autocorrelation $\rho_{xx}[m]$ is an estimate of the true autocorrelation $r_{xx}[k]$, based on a length-N observation of $x[n]$. Due to the similarity in the definitions of $P_{xx}(\omega)$ and $I_{xx}(\omega)$, the periodogram is called a *natural estimate* of the power spectral density. However, even though the periodogram is an asymptotically unbiased estimator of the PSD, it is not consistent in the sense that its variance does not decrease with increasing data record length [Oppenheim *et al*., 1999].

Suppose that $s[n]$ is a real harmonic signal with M as its fundamental period, where M is a positive integer. The signal $s[n]$ is a sum of harmonically related sinusoids at frequencies $f_i = i/M$, or

$$s[n] = \sum_{i=1}^{\left\lceil \frac{M}{2}-1 \right\rceil} A_i \cos(2\pi f_i n + \phi_i), \tag{7.5}$$

where each sinusoid amplitude A_i is a Rayleigh random variable with $E(A_i^2/2) = P_i$. Each phase ϕ_i is uniformly distributed in the interval of $[-\pi, \pi)$. All the random variables are independent under this signal model. Suppose that $\mathbf{x} = [x[0]\ x[1] \ldots x[N-1]]^T$ represents a length-N observation vector of a real discrete-time WSS random process $x[n]$. We want to detect whether $x[n]$ contains $s[n]$:

$$\begin{cases} H_0 : x[n] = w[n] \\ H_1 : x[n] = s[n] + w[n] \quad n = 0, 1, \ldots, N-1, \end{cases} \tag{7.6}$$

where $w[n]$ is a zero-mean additive white Gaussian noise (WGN) with variance σ^2. Thus, under the first hypothesis (H_0), the signal is purely WGN. Under the second hypothesis, the observation is the sum of a harmonic signal and a WGN.

Given the above signal model, Kay [1998] proposed the following generalized likelihood ratio test (GLRT) to detect the presence of a harmonic signal with an unknown fundamental period M:

$$\begin{aligned} T(\mathbf{x}) &= \max_M h(\mathbf{x}, M) \\ &= \max_M \left(\ln \frac{p(\mathbf{x}; \hat{\mathbf{P}}, H_1)}{p(\mathbf{x}; H_0)} - \frac{\left\lceil \frac{M}{2}-1 \right\rceil}{2} \ln N \right) \\ &= \max_M \left(\sum_{i=1,\, I_{xx}(i/M)/\sigma^2 > 1}^{\left\lceil \frac{M}{2}-1 \right\rceil} \left(\frac{I_{xx}(\frac{i}{M})}{\sigma^2} - \ln \frac{I_{xx}(\frac{i}{M})}{\sigma^2} - 1 \right) - \frac{\left\lceil \frac{M}{2}-1 \right\rceil}{2} \ln N \right) \geq \gamma, \end{aligned} \tag{7.7}$$

where $h(\mathbf{x}, M)$ is the raw test statistic and γ is the threshold of the test. The raw test statistic is evaluated for a set of possible fundamental periods and the maximum is selected as the test statistic $T(\mathbf{x})$ of the test. The test statistic is compared with the threshold. If it is smaller than the threshold, H_0 is selected (no harmonic signal).

Otherwise, H_1 is selected (presence of a harmonic signal). Note that in the latter case, the fundamental period of the harmonic signal is determined from the argument of the maximum raw test statistic.

The first term of the raw test statistic is the generalized likelihood ratio (GLR) of the hypothesis test, where $\hat{\mathbf{P}}$ is the maximum likelihood (ML) estimation of unknown parameters with $\mathbf{P} = \left\lfloor P_1 \; P_2 \ldots P_{\lfloor \frac{M}{2} - 1 \rfloor} \right\rfloor^T$. The second term is the penalty factor based on the minimum description length idea [Rissanen, 1978]. $I_{xx}(\cdot)$ is the periodogram of $x[n]$.

A practical concern in computing the test statistic is the evaluation of the periodogram. As discussed before, only samples of the periodogram at frequencies $f_k = k/N$ are available via the N-point DFT of the signal. The periodogram should have a sufficiently dense frequency population so that it can be evaluated at various sample frequencies (i/M). Note that as M increases, its frequency increment ($1/M$) gets smaller. Also increasing M by one results in totally different harmonic frequencies. For a given range of M, N should be sufficiently large to accommodate all possible sampling frequencies.

However, depending on the circumstances, the data length N may not be made arbitrarily large. For example, the signal may have a time-varying nature or the system may need an update at an interval shorter than N. In this case, one solution is to increase the periodogram size (denser frequency population) by zero padding. In zero padding, the periodogram size is increased to L ($L > N$) by padding $L - N$ zeros at the end of the signal and performing L-point DFT. However, zero padding does not improve the frequency resolution. It simply provides us with a method for interpolating the values of the measured spectrum at more frequencies [Proakis *et al*., 2002].

Given M, the periodogram size requirement can be expressed by the following simple formula:

$$i\left(\frac{1}{M} - \frac{1}{M+1}\right) > 2\left(\frac{k+1}{L} - \frac{k}{L}\right), \tag{7.8}$$

where i is the harmonic number, k is the frequency number and L is the length of the periodogram sequence. It simply states the fact that the frequency resolution provided by the periodogram should be smaller than the resolution at harmonic number i. The scaling factor 2 is introduced to allow rounding to the nearest integer for k. The worst case occurs when $i = 1$, i.e. in evaluating the fundamental frequency or period. We rearrange the inequality with $i = 1$, to get the simple result:

$$L > 2M(M+1). \tag{7.9}$$

With L roughly proportional to M^2, as M increases, L has to increase quadratically. In our application, the fundamental period is on the order of 50 with the 120 Hz power coding rate. Thus, L is typically set at 8192.

When the fundamental period M of a harmonic signal $s[n]$ (corrupted by an additive white noise) is available, its ML estimation is given by the following formulae [Wise *et al*., 1976]:

$$\hat{s}[k] = \begin{cases} \dfrac{1}{K+1}\displaystyle\sum_{r=0}^{K} x[k+rM] & 0 \le k \le N - KM - 1 \\ \dfrac{1}{K}\displaystyle\sum_{r=0}^{K-1} x[k+rM] & N - KM \le k \le M - 1 \end{cases}, \tag{7.10}$$

where $K = \left\lfloor \frac{N}{M} \right\rfloor$, i.e. the number of length-M cycles in the length-N observation of $x[n]$. Note that $\hat{s}[k]$ is an M-periodic sequence, i.e. $\hat{s}[k] = \hat{s}[k+rM]$, $r \in \mathbb{Z}$. Thus, the ML estimation of a harmonic signal is simply the periodic-average of an observed waveform. When N is an exact multiple of M, only the second formula is used. Obviously, the quality of estimation increases with K, although depending on the circumstances it can not be made arbitrarily large. In practice, three or four seems to be a reasonable lower bound for K to yield a reliable estimation.

When multiple harmonic signals exist, it is customary to detect and estimate each harmonic signal sequentially. Each harmonic signal is detected, estimated and eliminated from the observation starting from the one with the largest magnitude (strongest signal), until the detection test statistic of the remaining observation becomes smaller than a threshold. A simultaneous detection and estimation of multiple harmonic signals is possible, but is computationally more expensive and requires *a priori* knowledge of the number of harmonic signals, which may change depending on the circumstances. In this chapter, we take the sequential detection and estimation approach.

To eliminate the harmonic signal with a known fundamental frequency from an observation, a certain type of filtering has to be performed. In cochannel speaker separation in speech processing, a comb filter with nulls at the harmonic frequencies is applied to the observation to suppress the harmonic signal, while a comb filter with peaks at the harmonic frequencies is used to enhance the harmonic signal [Morgan *et al*., 1997; Lim *et al*., 1978]. This technique of *harmonic enhancement and suppression* works reasonably well for speech signals, in which only relative strengths of spectral components are important (the absolute strength of an audio signal is easily changed with amplification). However, in our case, we want to conserve the absolute strength of both the harmonic signal and the remaining observation for further analysis. Thus, instead of applying a comb filter to suppress a harmonic signal, we simply subtract its ML estimation from the observation (implicit comb filtering). The difference (or remaining observation) is subject to further harmonic signal detection test.

Figure 7-1 shows the time graphs of the real power P, the reactive power Q and the fifth apparent power A_5 defined as $\sqrt{P_5^2 + Q_5^2}$, i.e. the magnitude of the fifth complex harmonic power. The data were collected at a test building in 160 Sansome St., San Francisco, CA, on 12/04/02, at the 120 Hz preprocessing rate. The building has two

VSDs– one for a supply fan (100 hp) and the other for a return fan (75 hp) in its air handling system. The graphs clearly show periodic peaks, implying the colored nature of the signal generated by the VSDs. A closer examination of the graphs reveals that the period and the phase of the periodic signal components are the same across the three waveforms. However, they differ in magnitudes. The real power shows the largest amplitude, followed by A_5 and Q.

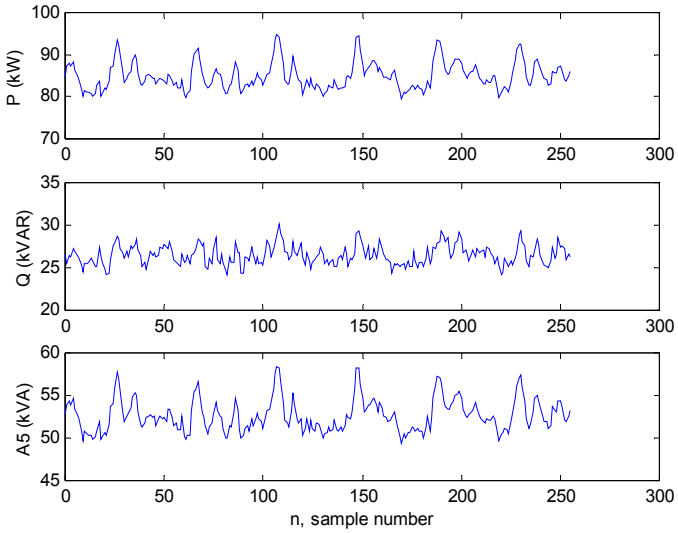


Figure 7-1 Time graphs of power signals.

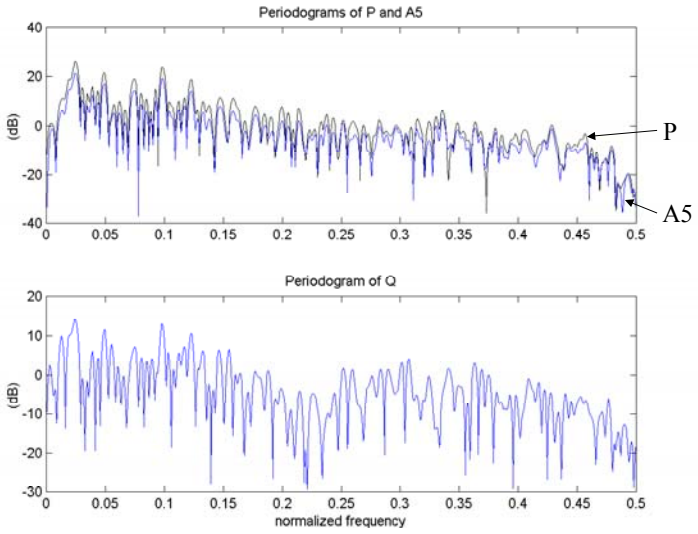


Figure 7-2 Periodograms of power signals in Figure 7-1 after their VSD means removed.

Figure 7-2 shows the periodograms of the signals in Figure 7-1 after their VSD mean values are removed. The periodograms clearly show the harmonic nature of the periodic signals and that the fundamental frequencies are the same across the power signals. However, they differ in harmonic signal magnitudes. P and A_5 differ almost by a constant

amount in the logarithmic magnitude[13] scale, implying a constant (multiplicative) scaling factor in the original magnitude scale. Q has smaller peaks than P or A_5, but has relatively high magnitudes at mid and high frequency region. The figure suggests that we may use either P or A_5 to detect the presence of a harmonic signal across power waveforms (Q has too small magnitudes) and to estimate its fundamental frequency, but that the estimation should be performed separately on each power signal because they contain different amounts (magnitudes) of harmonic signals.

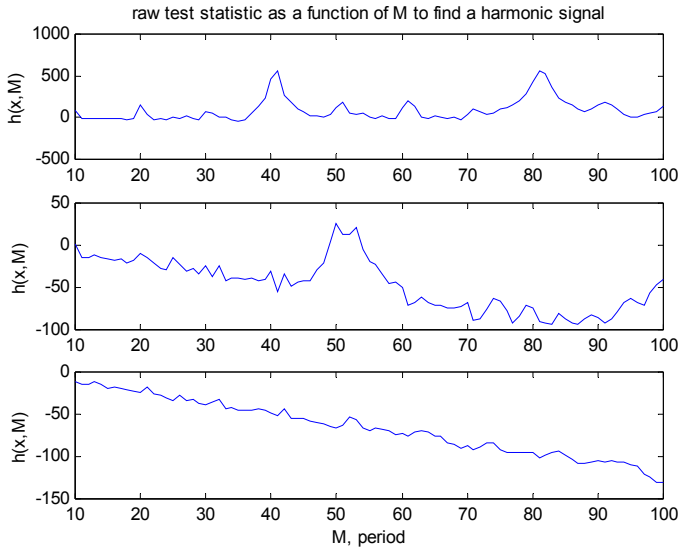


Figure 7-3 Raw test statistics to detect the presence of a harmonic signal.

Figure 7-3 shows the raw test statistic at each different stage of the sequential harmonic signal detection test, using the P in Figure 7-1 as an observation sequence. The top graph shows the raw test statistic as a function of M for the original observation. The graph shows two almost identical-height peaks at $M = 41$ and $M = 82$. However, the peak at $M = 41$ is slightly higher and the test estimates the fundamental period of the first harmonic set as 41, which is indeed correct. Nonetheless, the test may yield $M = 82$ with a slight change in the signal content. *Period doubling error* is often observed and is a weak point of the harmonic signal detection and period estimation method based on the periodogram.

The second graph of Figure 7-3 shows the raw test statistic after the ML estimation of the first harmonic signal is subtracted from the observation. The graph clearly shows the presence of a second harmonic set with $M = 50$, although it is much weaker than the first harmonic signal. The upper graph of Figure 7-4 shows the periodogram of the observation after the first harmonic set is subtracted. The periodogram clearly shows a peak at $f = 1/M = 0.02$. We remember that the building has two VSDs. So, it is natural to expect the presence of two separate harmonic sets. In reality, the first strong signal is related to the supply fan and the second weak signal is related to the return fan. Thus, we can detect and estimate two harmonic signals generated by two different physical loads, which in turn helps to monitor the state of each load.

[13] The logarithmic magnitude of x is defined as $10\log_{10}\|x\|$ (decibel, dB).

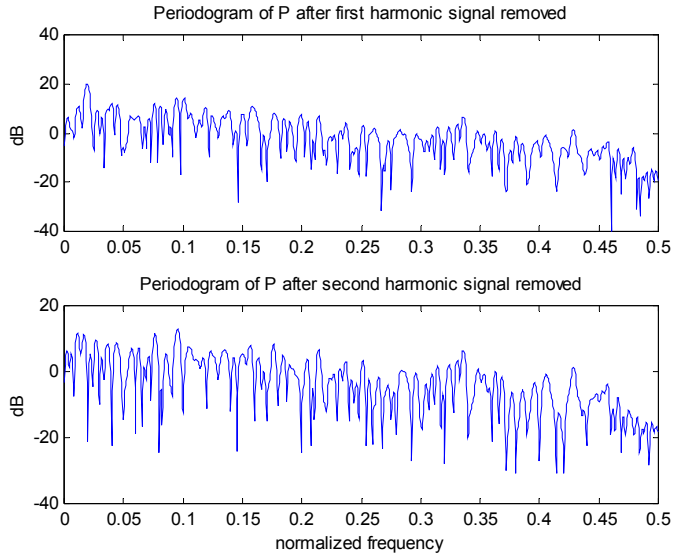


Figure 7-4 Periodograms of observations after harmonic signals are subtracted.

The third graph of Figure 7-3 shows the raw test statistic after the ML estimation of the second harmonic set is subtracted from the observation. It shows no conspicuous peak and decreases almost monotonically. With the threshold of approximately zero, the test will conclude that there exists no more harmonic signal. The lower graph of Figure 7-4 shows the periodogram of the observation after two harmonic sets are subtracted. It shows no outstanding peak, implying the absence of a harmonic signal. Also, we note that the periodogram is skewed rather than flat. Hence, the harmonic signal subtraction is, in effect, a harmonic suppression filter rather than a whitening filter.

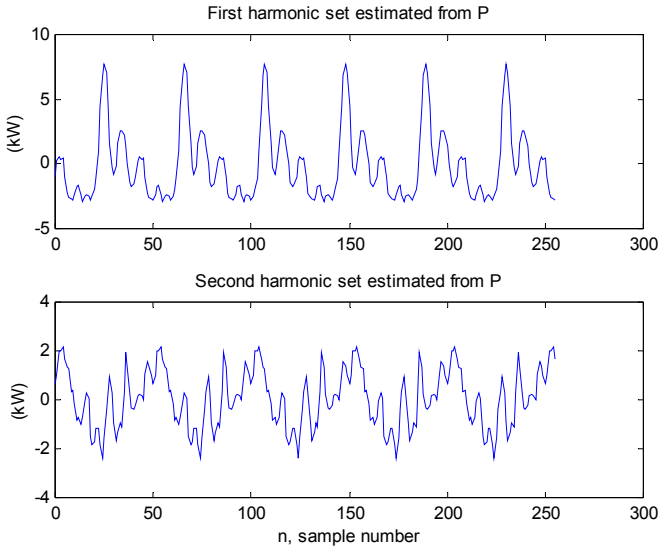


Figure 7-5 ML estimations of harmonic signals from *P*.

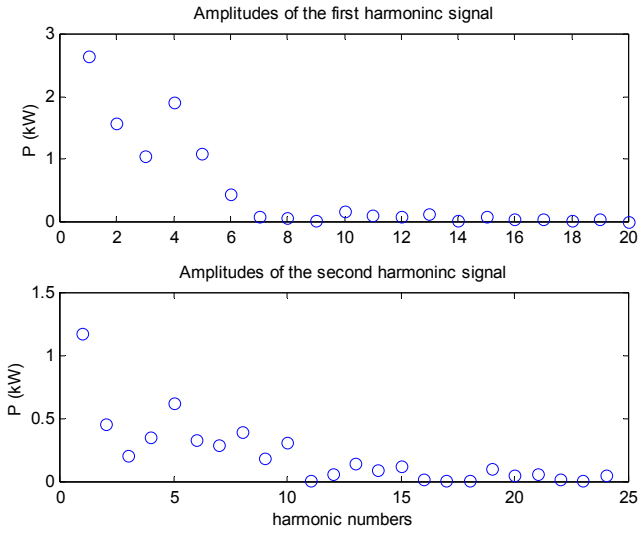


Figure 7-6 Amplitudes of each harmonic set from *P*.

Figure 7-5 shows the results of the harmonic signal ML estimation. The second set is estimated after the first set estimation is subtracted from the observation. The graphs clearly show that the first set has much larger magnitudes than the second set and that the fundamental frequency of each set has the largest amplitude. Figure 7-6 shows the amplitude of each harmonic set at each harmonic number, using the relationship between a sinusoid amplitude and its corresponding periodogram peak height [Kay, 1993]:

$$\hat{A}_k^{\,2} = \frac{4}{N} I_{xx}(f_k), \tag{7.11}$$

where $f_k = k/M$ is the kth harmonic frequency and A_k is the amplitude of the sinusoid with the frequency f_k. N is the length of the observation sequence $x[n]$. The fundamental frequency of each set has the largest amplitude as evidenced in the time waveforms. The amplitudes of the first set roll off quickly after the sixth harmonic. The amplitudes of the second set become negligible after the tenth harmonic.

Figure 7-7 shows the harmonic signals estimated from *Q*, with known fundamental periods. Figure 7-8 shows corresponding harmonic amplitudes. Again the first set has larger magnitudes than the second set. The amplitude at each harmonic frequency and their relative strengths are in general different from those of the harmonic sets from *P*.

This section has applied the method of harmonic signal detection and its fundamental period estimation based on the periodogram and the maximum-likelihood harmonic signal estimation to the power waveforms. The tests show that the ML estimation yields good results, when at least four cycles (of a fundamental period) are included in a single observation sequence. The periodogram-based harmonic signal detection and fundamental period estimation also performs reasonably well. However, it is slightly unsatisfactory in the sense that it often generates period doubling errors and that it requires *a priori* knowledge of the additive white noise covariance, which may change

with time. Also, it is burdensome to have to compute a large size periodogram. The next section introduces a better alternative in detecting a harmonic signal and estimating its fundamental period using *autocorrelation*.

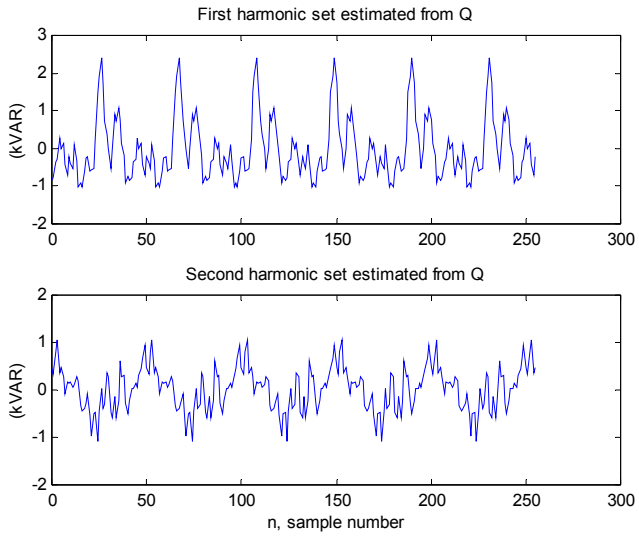


Figure 7-7 ML estimations of harmonic signals from Q.

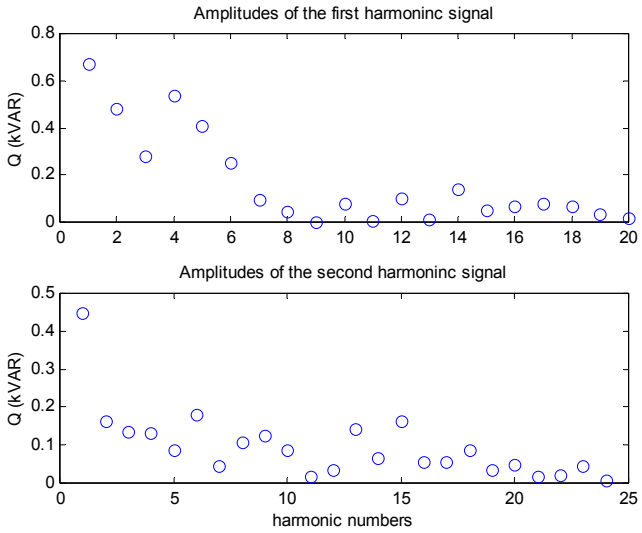


Figure 7-8 Amplitudes of each harmonic set from Q.

7.3 Detection and pitch estimation of harmonic signal using autocorrelation

Detecting and estimating a harmonic signal has been one of the central problems in speech processing, because a (voiced) sound has a harmonic nature due to the periodic pulsation of the glottis (an opening between the vocal folds) and the subsequent excitation (resonance) in the vocal tract [Deller *et al*., 2000]. The fundamental frequency of a voiced sound (hence a harmonic signal) is often called *pitch*. Several pitch detection

and estimation methods are known to the speech processing community, such as the harmonic selection based on the histogram method [Parsons, 1976], the cepstral analysis [Stubbs and Summerfield, 1988], the modified covariance spectrum estimator based on the autoregressive signal model [Naylor and Porter, 1991], the mean-square-error estimator based on the sinusoidal speech model [McAulay and Quatieri, 1990], delay filtering and error minimization in the linear predictive coding (LPC) [Chen *et al*., 1993], etc. However, the ML pitch detector and estimator based on the autocorrelation is widely used and performs well in the presence of additive noise [Morgan *et al*., 1997; Naylor and Boll, 1987]. In this section, we apply the same method to detect the harmonic signal and estimate its pitch from a power waveform.

Given a length-N real sequence $x[n]$, the raw test statistic $E(\mathbf{x}, p)$ to detect the presence of a harmonic signal with a pitch period p is given by [Wise *et al*., 1976]:

$$E(\mathbf{x}, p) = \frac{2p}{N} \sum_{k=1}^{L} R_{xx}[kp], \tag{7.12}$$

where $L = \lfloor (N - p)/p \rfloor$ and $R_{xx}[k]$ is the autocorrelation function, defined as

$$R_{xx}[k] = \sum_{j=0}^{N-1-k} x[j]x[j+k]. \tag{7.13}$$

Note that the autocorrelation function is an unnormalized estimate of the true autocorrelation of a real WSS random process. The test statistic is selected by maximizing $E(\mathbf{x}, p)$ over a range of (integer) p. The ML estimation of the pitch period is simply the argument of the maximum raw test statistic.

$E(\mathbf{x}, p)$ is often interpreted as the inner product of the autocorrelation function and a comb filter $C_p[k]$ with intertooth spacing p, i.e.

$$C_p[k] = \begin{cases} 1 & \text{if } k = lp, \ l = 1, 2, 3, \ldots \\ 0 & \text{otherwise.} \end{cases} \tag{7.14}$$

Then the raw test statistic can be represented as

$$E(\mathbf{x}, p) = \frac{2p}{N} \mathbf{c}_p{}^T \mathbf{r}_{xx}, \tag{7.15}$$

where $\mathbf{r}_{xx} = [R_{xx}[0]\ R_{xx}[1] \ldots R_{xx}[N-1]]^T$ and $\mathbf{c}_p = [C_p[0]\ C_p[1] \ldots C_p[N-1]]^T$. The comb filter is determined by maximizing the correlator output of $C_p[k]$ and $R_{xx}[k]$.

The ML pitch detector and estimator is often susceptible to pitch period doubling and halving errors. To remedy this problem, the comb filter is modified to have teeth with

decreasing amplitude [Martin, 1982]. In practice, a power weighting of $k^{-0.8}$ is applied to each element of the comb filter:

$$E(\mathbf{x}, p) = \frac{2p}{N} \sum_{k=1}^{N-1} k^{-0.8} C_p[k] R_{xx}[k] . \tag{7.16}$$

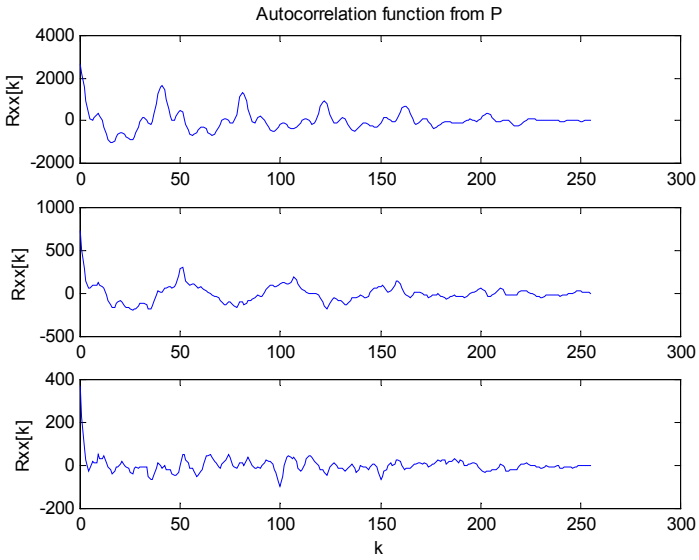


Figure 7-9 Autocorrelation functions computed from *P*.

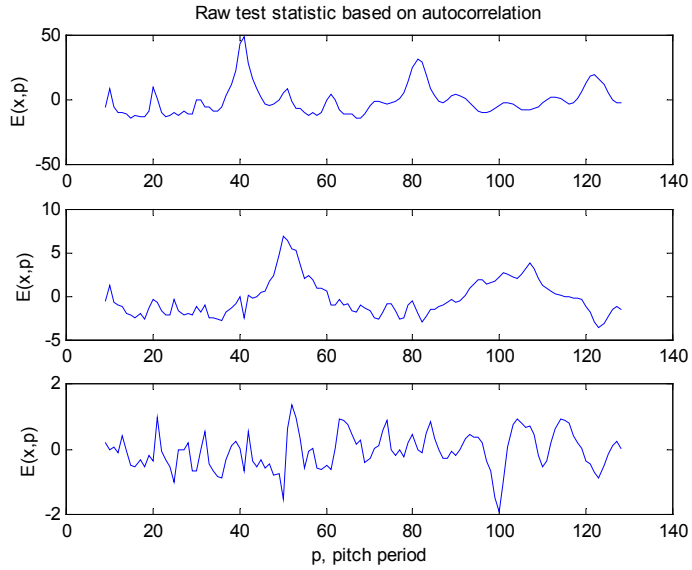


Figure 7-10 Raw test statistics from autocorrelations in Figure 7-9.

The first graph of Figure 7-9 shows the autocorrelation function computed from the real power waveform in Figure 7-1. The autocorrelation clearly shows the periodic nature with decreasing magnitude and its period is equal to the pitch period of the harmonic signal, i.e. $p = 41$. The periodic character of the autocorrelation function is evident from

its definition, where products of observations with a lag k add constructively when the lag is an integer multiple of an underlying pitch period. The first graph of Figure 7-10 shows the corresponding raw test statistic as a function of p. The maximum clearly occurs at the correct pitch period, although multiples of the pitch period also have peaks with substantial but smaller heights. Thus, we verify that the prospect of period doubling is significantly reduced, an advantage over the method based on the periodogram in the previous section.

The first harmonic set is estimated using the ML estimator discussed in the previous section and subtracted from the observation. New autocorrelations are computed from the difference and shown in the second graph of Figure 7-9. The autocorrelation still shows a periodic nature, implying the presence of the second harmonic set. The second graph of Figure 7-10 shows the corresponding raw test statistic. The second harmonic signal with p = 50 is detected, estimated and subtracted from the observation. The third graph of Figure 7-9 shows the autocorrelation of the observation after two harmonic sets are subtracted. It does not show a conspicuous period and its corresponding raw test statistic in the third graph of Figure 7-10 shows no dominant peak.

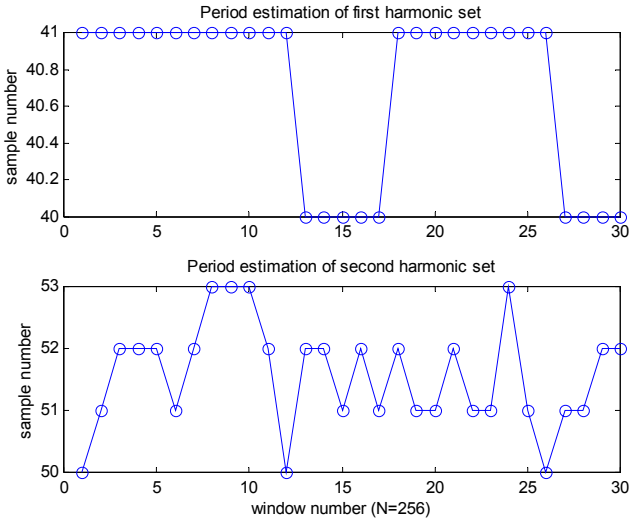


Figure 7-11 Estimated pitch periods for a series of observation.

Figure 7-11 shows the ML pitch estimation for each harmonic set from the data recorded at the 160 Sansome building on 12/04/02. The estimation was conducted for thirty consecutive windows with no overlap, each with a length of N = 256. The first graph clearly shows the oscillatory nature of the pitch period of the first harmonic set. As explained before, the pitch of a harmonic signal is proportional to the rotational speed of a corresponding VSD. Thus, the graph suggests that there is an oscillation in the rotational speed of the supply fan. Because the pressure in the supply air duct regulates the fan speed, this in turn suggests that there exists a pulsation in the duct air pressure. This phenomenon can be explained as a *limit cycle* created by the combination of the physical characteristic of the air duct system and the specific control logic used to control the fan speed based on the duct air pressure, as explained in Appendix B. However, the oscillation in the fan speed is not fixed but varies with time.

The second graph of Figure 7-11 shows the pitch period estimation of the second harmonic set. Unlike the pitch period of the first set, it fluctuates rather randomly between 50 and 53, though its true period seems to lie somewhere between 51 and 52. Because the speed of the return fan (source of the second harmonic signal) is regulated by the static pressure measured at a location in large building space, it should be relatively constant. The unexpected random nature of the second harmonic pitch may have a physical origin, such as a noise or disturbance in the velocity control loop of the return fan.

However, it is more likely that the randomness of the second harmonic pitch is caused by the *spectral leakage* in the process of windowing and harmonic subtraction. Any windowing forces a sharp impulse-like frequency component to be dispersed around its nominal frequency. Also, the first harmonic set is not necessarily confined at its nominal harmonic frequencies for various reasons. Windowing is obviously one of the reasons. The time-varying nature of the signal actually creates a very narrow band of pitch, not necessarily an impulse. Also, the true pitch period of a harmonic signal is not necessarily an integer. Thus, the first harmonic set has bands of harmonic frequencies, and their ML estimations are, in a sense, weighted averages of each band with the constraint that the pitch period is an integer. When the ML estimation of the first harmonic set is subtracted from the observation, not all of its spectral components are eliminated but some of them leak into the remaining observation. The converse is also true. Some of the spectral components of the second harmonic set may be picked up by the first harmonic set estimation if their frequencies are located sufficiently close to the harmonic frequencies of the first set.

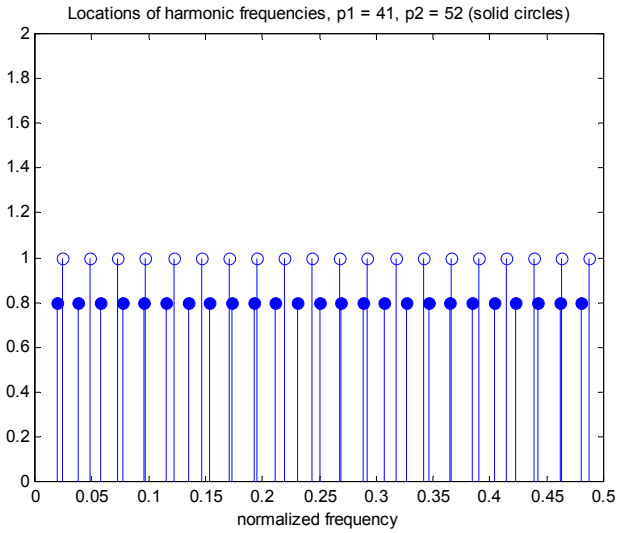


Figure 7-12 Harmonic frequencies of two harmonic sets with pitch periods of 41 and 52.

Figure 7-12 shows the harmonic frequency locations of two harmonic sets when pitch periods are 41 (open circles) and 52 (solid circles). The graph demonstrates that the first and fourth harmonics of the first set, which typically have large amplitudes, are located

very close to the harmonics of the second set. It is inevitable that some elements of the second set are included to the first set estimation and the leakage from the first set changes the spectral content of the second harmonic signal. Depending on the nature of true spectral content (or PSD) of each harmonic set, the degree of the spectral leakage may vary. This in turn creates a distortion in the spectral content of the second harmonic set embedded in the remaining observation and shows up as a fluctuating pitch period estimation. Improvements are possible by a better windowing (Hamming window, etc.), a simultaneous pitch estimation, etc. These issues will be explored in future. The current sequential detection and estimation technique still yields reliable and useful results and the pitch period estimation is usually within 5% of its nominal value.

Figure 7-13 shows the standard deviation of each power waveform after its two harmonic sets and VSD means are removed. The standard deviations exhibit substantial fluctuations partly due to the time varying (noise) nature of the VSDs and mostly due to the incompleteness of the harmonic set removal. As explained before, spectral content of the two harmonic sets leaks into the remaining observation. Also, the harmonic de-noising by subtraction performs poorly towards the boundary of a window. The ML signal estimation tends to align itself to the harmonic waveform in the middle of a window, if the signal has a time varying nature. We also remember that the harmonic subtraction is not a white filtering in a strict sense. A separate white filtering may be performed to reduce the fluctuations in the standard deviations.

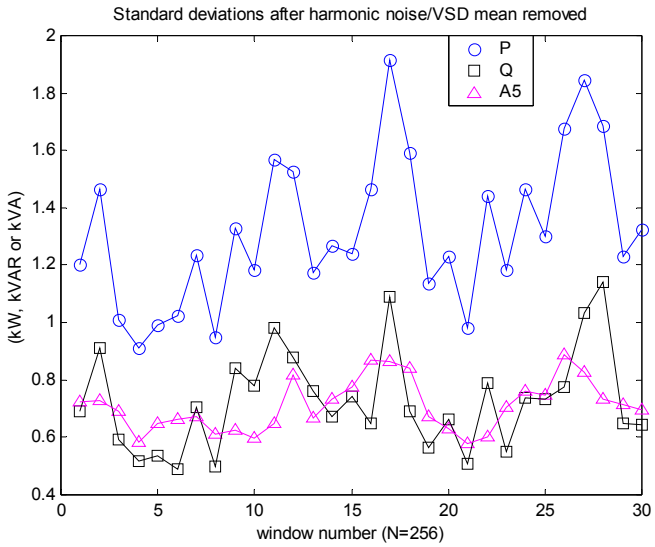


Figure 7-13 Standard deviations of observations after their harmonic signals and VSD means are removed.

Figure 7-14 shows the amplitude at each harmonic number of the first harmonic set estimated from the real power waveform at each window. The amplitude becomes negligible after the sixth harmonic and the first and the fourth have dominant magnitudes. The amplitude distribution is relatively stable over time, implying the consistency of the underlying load imbalance mechanism (or the stable nature of the resonance/excitation mechanism). Figure 7-15 shows the amplitudes of the second harmonic set. The

amplitude distribution is also relatively stable, although the amplitudes at high harmonic numbers are slightly uneven, possibly due to the spectral leakage.

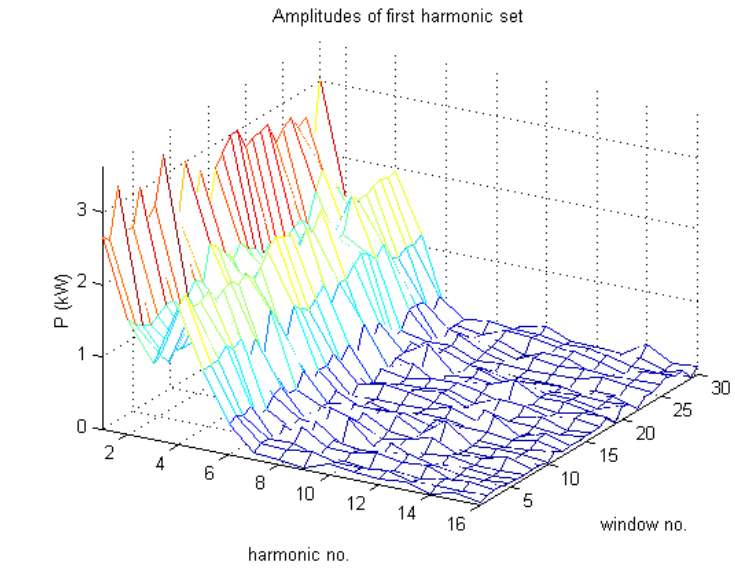


Figure 7-14 Amplitudes of first harmonic set at each window.

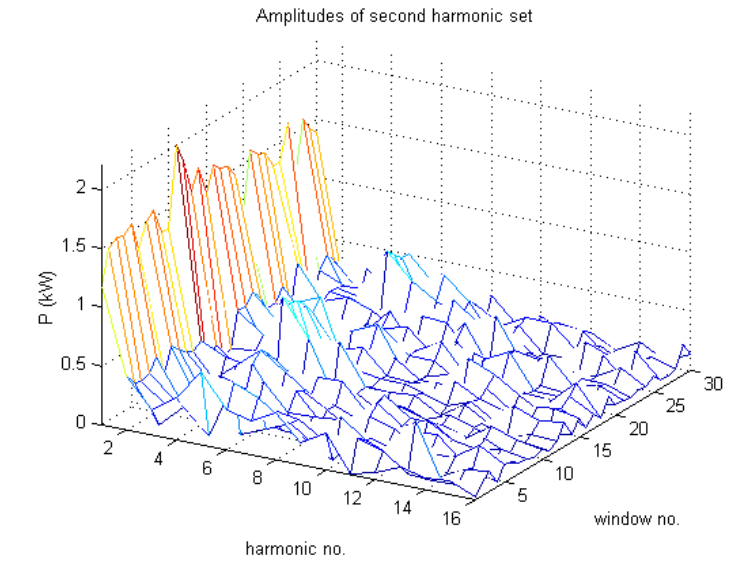


Figure 7-15 Amplitudes of second harmonic set at each window.

Figure 16 shows the total amplitude of each harmonic set defined as:

$$A = \sqrt{\sum_{i=1}^{L} A_i^{\,2}}\,, \tag{7.17}$$

where L is the maximum harmonic number and A_i is the ith amplitude of a harmonic set. Although there exist fluctuations, we observe that the total magnitude of each set follows a certain mean value. The total magnitude is a suitable measure of the extent of the load

imbalance and a monitoring routine can be installed which detects a fault when a low-pass filtered version of the total magnitude exceeds a certain threshold.

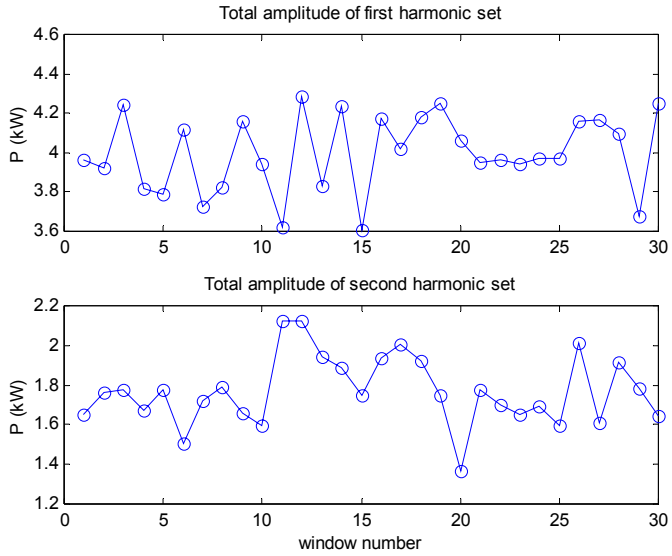


Figure 7-16 Total amplitude of each harmonic set.

In this section, we applied the ML pitch detection and estimation method based on the autocorrelation to detect and estimate the pitch periods of the harmonic signals generated by VSDs. The ML detector performs better than the periodogram based detector in the previous section, in the sense that the former generates less period doubling errors, that *a priori* knowledge of noise variance is not required and that the computational burden is slightly reduced[14]. However, the ML pitch estimation is still unsatisfactory because the pitch period is constrained to be an integer. In the next section, we further refine the method to allow a *fractional* pitch period.

7.4 Fractional pitch period estimation and least squares signal estimation

The ML pitch estimator provides a reasonably accurate pitch period estimation for most applications. However, its performance is limited due to the integer pitch period constraint. A small round up error in the pitch period can result in significant biases in the higher harmonic frequency estimations. Natural processes do not necessarily have integer periods (of sampling intervals), especially when they have time-varying natures.

Fractional pitch (FP) period estimation is often employed to improve the quality of estimation, obviously at the cost of additional computation and complexity. In an FP period estimation, the pitch period is allowed to be an arbitrary positive real number. The

[14] For a length-N sequence, the autocorrelation requires approximately $N(N+1)/2$ multiplications and $N(N+1)/2$ additions. A size L fast Fourier transform (FFT) requires $L\log_2 L$ additions and $\frac{L}{2}\log_2 L$ multiplications. With $N=256$, $N(N+1)/2=32896$. With $L=8192$, $L\log_2 L=106496$.

FP period estimation can be performed directly from the Fourier transform of a signal, such as by *peak picking* [Quatieri and Danisewicz, 1990], or through an additional refinement procedure based on an integer pitch period estimation. The latter includes an interpolation filter [Chen *et al*., 1993], in which the sampling rate is interpolated to match a proposed FP period, and the iterative minimization of the difference between the original spectrum and the estimated spectrum to obtain the optimal estimations of signal parameters (including the FP period) [Griffin and Lim, 1988].

There also exist general-purpose sinusoidal parameter estimation techniques such as Pisarenko's harmonic decomposition method [Pisarenko, 1973], MUSIC [Stoica and Nehorai, 1989] and ESPRIT [Roy and Kailath, 1989]. These methods, which rely on the signal subspace idea, provide high frequency-resolution estimations of multiple sinusoids in an additive white noise. However, they are not specifically designed for a harmonic signal, in which sinusoidal frequencies are harmonically related. They also require very accurate estimations of the autocorrelation matrices for the eigen-decompositions. Thus, they are not good candidates for a harmonic signal with a time-varying nature.

In this section, we apply a variation of the harmonic enhancement and suppression method, in which a comb filter is constructed with a FP period. This method can yield an FP period estimation with a reasonable accuracy without the burden of the iterative minimization [Morgan *et al*., 1997].

For a given fractional pitch period p, the corresponding second-order harmonic suppression filter (HSF) $H_{-}(z)$ is given by the following formula:

$$H_{-}(z) = \frac{(1-\alpha z^{-p})(1-\alpha z^{p})}{(1+\alpha)^{2}}, \tag{7.18}$$

where $0 < \alpha < 1$. The variable α is chosen close to one, and $\alpha = 0.99$ is normally used. On the frequency domain, $H_{-}(z)$ is simply a comb filter with nulls at the harmonic frequencies of $1/p$.

A length-N observation sequence $x[n]$ is filtered through $h_{-}[n]$ to yield a harmonically suppressed signal $s[n]$:

$$s[n] = x[n] \otimes h_{-}[n], \tag{7.19}$$

where $\otimes$ denotes circular convolution. For each FP period from a set of candidates, $s[n]$ is computed and the one that gives minimum signal energy is selected.

The circular convolution can be implemented efficiently in the discrete frequency domain via the DFT,

$$S[k] = X[k]H_{-}[k]. \tag{7.20}$$

The DFT of the HSF is evaluated as (due to the noninteger circular shift property of the DFT):

$$H_{\_}[k] = \begin{cases} H_{\_}(z)\,|_{z=\exp(j2\pi k/N)} & 0 \le k \le \frac{N}{2} \\ H_{\_}(z)\,|_{z=\exp(j2\pi(k-N)/N)} & \frac{N}{2}+1 \le k \le N-1 \end{cases}, \tag{7.21}$$

assuming an even N. The second norm of $S[k]$ is minimized to select the best FP period.

When an integer pitch period estimation p is given, the fractional pitch period candidates p_i can be obtained either to limit the error at the highest harmonic frequency less than one-half bin of an N-point DFT, or with a fixed increment/decrement interval from p. The former is represented by [Morgan *et al.*, 1997]:

$$p_i = \left(1 + \tfrac{2i}{N}\right)p \quad -I \le i \le I\,, \tag{7.22}$$

where i is an integer and I is a bound. The latter is simply given as:

$$p_i = p + i\,\Delta p \quad -I \le i \le I\,, \tag{7.23}$$

where Δp is an increment.

Figure 7-17 shows an example of the HSF result when $p = 41$ and $\Delta p = 0.15$. The norm of the filtered signal shows a minimum at $p_i = 41.45$.

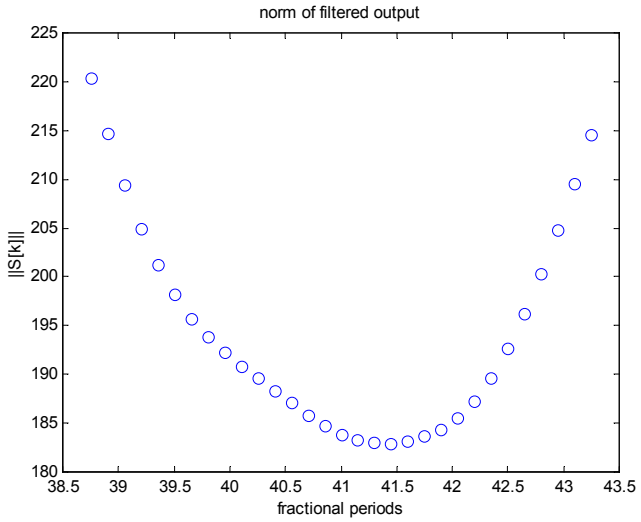


Figure 7-17 Norm of harmonic suppressed signal as a function of fractional pitch period.

Given a FP period estimation, we would like to estimate the corresponding harmonic signal from the observation. However, we can not use the ML signal estimator because it relies on the integer pitch period assumption. The ML estimator may be applicable if the signal is upsampled via interpolation so that the FP period becomes an integer. The

iterative minimization method [Griffin and Lim, 1988] can be applied to estimate a harmonic signal with an appropriate sinusoidal signal model. The iterative minimization can be costly because it tries to seek a global minimum in the difference between the original spectrum and a synthesized spectrum. An alternative is to seek a local minimum at a harmonic frequency via a frequency warping and repeat the procedure for every harmonic frequency [Choi and Youn, 2002].

However, these methods are generally complex and computationally intensive. In this section, we develop a simple least squares (LS) signal estimator based on a harmonic sinusoidal signal model. The LS estimation of a harmonic signal in the frequency domain was investigated by Quatieri and Danisewicz [1990].

Consider a sinusoidal harmonic signal $s[n]$ with a pitch period p:

$$s[n] = \sum_{i=1}^{L} A_i \cos(2\pi f_i n + \phi_i), \tag{7.24}$$

where L is the maximum harmonic number and $f_i = i/p$ is the ith harmonic frequency. Each cosine term can be expanded into the sum of a cosine and a sine using the trigonometric identity:

$$A_i \cos(2\pi f_i n + \phi_i) = a_i \cos 2\pi f_i n + b_i \sin 2\pi f_i n, \tag{7.25}$$

where $a_i = A_i \cos\phi_i$ and $b_i = -A_i \sin\phi_i$. We assume that a length-N observation sequence $x[n]$ is the sum of $s[n]$ and a zero-mean additive noise $w[n]$:

$$x[n] = s[n] + w[n] \quad 0 \le n \le N-1. \tag{7.26}$$

In a vector-matrix format,

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \tag{7.27}$$

where

$$\mathbf{x} = [x[0]\ x[1] \ldots x[N-1]]^T,\ \mathbf{w} = [w[0]\ w[1] \ldots w[N-1]]^T,\ \boldsymbol{\theta} = [a_1\ b_1\ a_2\ b_2 \ldots a_L\ b_L]^T,$$

$$\mathbf{H} = \begin{bmatrix} \cos 2\pi f_1 0 & \sin 2\pi f_1 0 & \cdots & \cos 2\pi f_L 0 & \sin 2\pi f_L 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \cos 2\pi f_1 (N-1) & \sin 2\pi f_1 (N-1) & \cdots & \cos 2\pi f_L (N-1) & \sin 2\pi f_L (N-1) \end{bmatrix}. \tag{7.28}$$

Note that $\mathbf{H}$ is an $N \times 2L$ matrix and has a rank of $2L$ when $2L < N$.

The LS estimation of the parameter vector is given by:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}. \tag{7.29}$$

Then, each harmonic amplitude estimate can be computed from the LS parameters,

$$\hat{A}_i = \sqrt{\hat{a}_i^{\,2} + \hat{b}_i^{\,2}} \,. \tag{7.30}$$

Due to the Gauss-Markov theorem, the LS estimator is the same as the best linear unbiased estimator (BLUE) when the additive noise is zero-mean white Gaussian [Kay, 1993]. However, no optimality can be claimed when $w[n]$ is not white Gaussian.

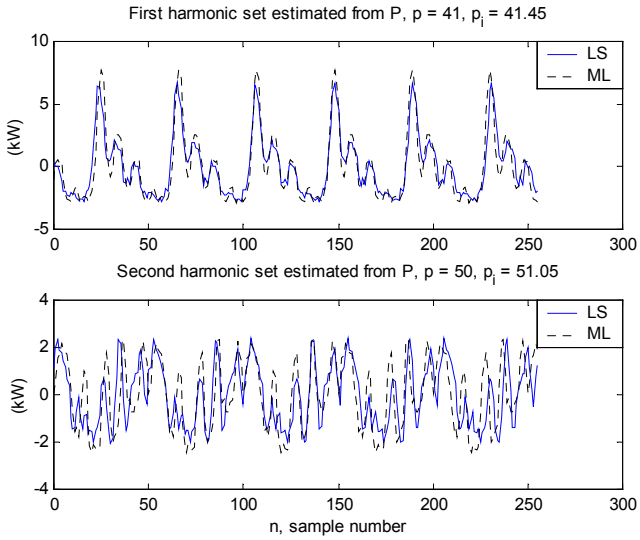


Figure 7-18 Harmonic signals estimated from P (ML estimations same as Figure 7-5).

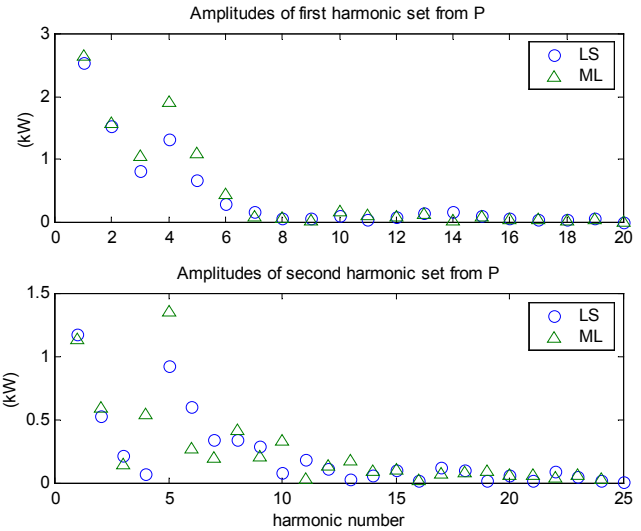


Figure 7-19 Amplitudes of each harmonic set.

Figure 7-18 shows the LS estimations of harmonic signals from the real power in Figure 7-1. The figure also shows the corresponding ML estimations with integer pitch periods (reproduction of Figure 7-5). Each fractional pitch period estimated from the HSF is also

shown in the figure. The graphs suggest that the LS estimator tends to estimate a harmonic signal less aggressively than the ML estimator. We also note that two estimations tend to align in the middle of a window and grow out of phase towards window boundaries. Figure 7-19 shows the amplitudes of each harmonic set, both from the LS and the ML estimation. The graphs verify that the LS estimator underestimates the harmonic signal amplitudes.

Figure 7-20 shows the periodograms of the original observation and the residual observation after the LS harmonic signal estimation is subtracted. We observe that the residual still has relatively large peaks in the vicinity of the third, fourth and fifth harmonic frequencies of the first harmonic signal.

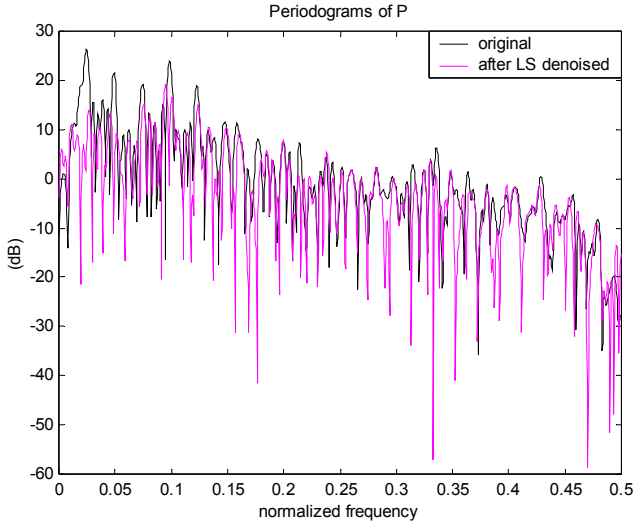


Figure 7-20 Periodograms of real power before and after de-noising.

There can be various reasons for this unsatisfactory result. As discussed before, the harmonic signal actually has a narrow band for each harmonic frequency rather than a fixed impulse-like location due to the time-varying nature. The HSF may yield an FP period estimation well adapted to a few large peaks, but the resulting harmonic frequencies can be off by significant amounts at other peaks. The LS signal estimation is very sensitive to frequency locations, thus a small offset between the center of a frequency band and frequency estimation can yield a significant bias in the signal estimation. Another concern is that a harmonic signal does not necessarily have peaks at exact harmonic locations, often justifying the use of a frequency-warping technique. Also, the harmonic signal in power waveforms has a relatively low signal-to-noise (SNR) ratio (compared to speech signals). The combination of these causes results in a relatively flat band of minimum values in Figure 7-17, reducing the credibility in determining the exact FP period.

The presence of the second harmonic signal can affect the FP period estimation of the first harmonic signal, because their first peaks are located close to each other. The second harmonic signal also acts as a non-white noise in the estimation of the first harmonic signal. A better windowing may be necessary to minimize the signal distortions. Also, the

power signal itself has certain distortions introduced by the power coder, during the course of interpolation and resampling of the raw voltage and current waveforms.

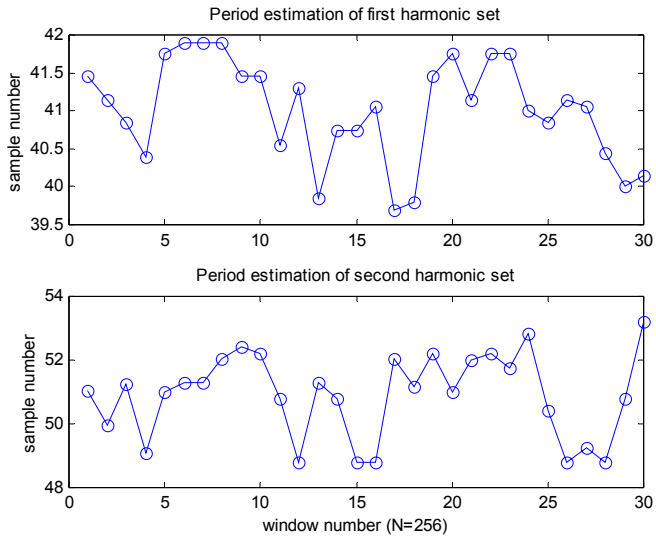


Figure 7-21 Estimated fractional pitch periods for a series of observation.

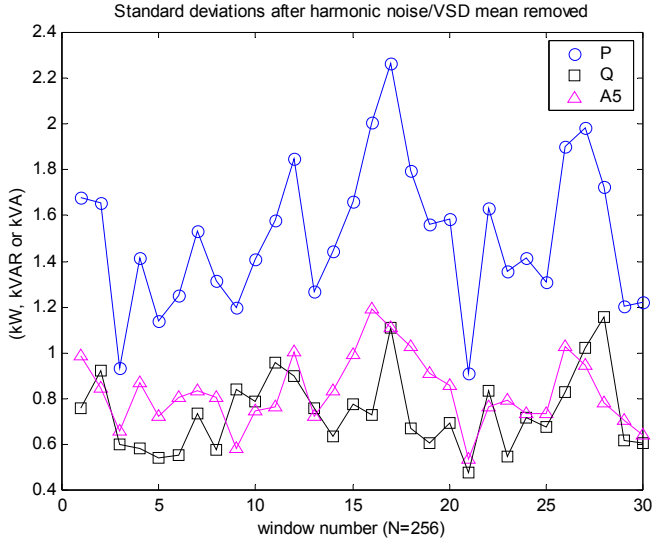


Figure 7-22 Standard deviations of residue observations (FP periods used).

Figure 7-21 shows the FP period estimation of each harmonic set from the same data as in Figure 7-11. In spite of the increased resolution, it is difficult to see any clear periodicity, compared to Figure 7-11. Figure 7-22 shows the standard deviations of the residue observations after the LS estimations of the FP period harmonic signals are subtracted. Compared to Figure 7-13, it generally shows larger standard deviations, implying a poor signal estimation and subtraction result.

A fractional pitch period estimation coupled with an appropriate harmonic signal estimation technique can potentially yield a better estimation of the signal and hence a

better harmonic de-noising of an observation. However, when the signal has a time-varying nature and a low SNR, it is difficult to estimate the FP period accurately and thus the signal estimation and de-noising can not be performed reliably. Also, the power signal has distortions created by the power coding and the method itself introduces various distortions, making the FP period estimation unreliable. The current FP period estimation and harmonic signal estimation technique may be applicable, when the circumstances are more favorable. We may also develop a more sophisticated method in future. Until then, we will use the ML harmonic signal estimator with an integer pitch period estimation, due to its simplicity and robustness.

7.5 Autoregressive modeling of harmonic signal and white filtering

In this section, we take a different approach from the previous sections by treating the harmonic signal present in a power waveform as colored noise, rather than the signal of interest to be estimated. Estimating and eliminating a harmonic signal can be difficult when it has a time-varying nature and/or a low SNR. Sometimes, we may simply want to filter out the harmonic noise before the power signal is fed to a load status analyzer, which relies on the whiteness of the noise present in the observation.

Constructing a whitening filter, nonetheless, requires a certain amount of knowledge about the nature of a targeted harmonic signal. In this sense, the harmonic noise is still a signal of interest. However, the estimation of the signal characteristics is performed implicitly in the context of white filtering. The specific properties of a harmonic signal, such as its pitch and harmonic amplitudes, will not be available in general.

We adopt the linear system model for a real zero-mean discrete-time WSS random process $x[n]$, in which a linear system $G(z)$ generates $x[n]$ from a zero-mean white Gaussian noise (WGN) process input $w[n]$. Figure 7-23 shows the linear system model. $G(z)$ is assumed to be time-invariant.

Figure 7-23 Linear system model of a discrete-time WSS process.

With this linear model, the PSD of $x[n]$ is determined solely by the PSD of $w[n]$ and $G(z)$ [Papoulis and Pillai, 2002]:

$$P_{xx}(z) = G(z)G(1/z)P_{ww}(z) = \sigma^2 G(z)G(1/z), \tag{7.31}$$

where σ^2 is the PSD of $w[n]$. Thus the PSD of $x[n]$ is completely determined by $G(z)$ within a scaling factor. In this sense, $G(z)$ is called the *shaping filter* of $x[n]$. For a given

PSD, infinitely many choices of $G(z)$ are generally possible. However, we are interested in a *unique* $G(z)$ that is causal and stable, and has a causal and stable inverse.

We can also conceive an inverse linear system in which $x[n]$ is filtered to $w[n]$ as shown in Figure 7-24. It can be shown that $H(z) = 1/G(z)$, where $G(z)$ is the unique one so that $H(z)$ is causal and stable. $H(z)$ is called the *whitening filter* of $x[n]$ and $w[n]$ is often referred to as the *innovations* of $x[n]$.

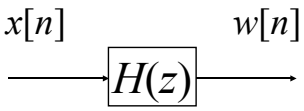


Figure 7-24 Inverse linear system to whiten $x[n]$.

Because $H(z)$ is the reciprocal of $G(z)$, constructing a whitening filter for $x[n]$ is equivalent to finding a unique $G(z)$, via *spectral factorization*, to represent the PSD of $x[n]$. Because the true PSD of a random process is often unknown, designing a whitening filter is intimately tied to the estimation of its PSD.

There exists a variety of PSD estimation methods, such as the nonparametric methods based on periodograms, the minimum-variance spectral estimation methods and the parametric methods [Kay, 1988; Kay and Marple, 1981]. Among these, the parametric methods usually give the best result [Proakis *et al*., 2002] and suit our need because we want to construct a whitening filter explicitly with the obtained parameters. The eigenanalysis algorithms, such as ESPRIT and MUSIC, are often considered as PSD estimation methods, although they can be more appropriately termed as sinusoidal parameter estimation methods. They also do not provide the PSD in a parameterized form, making it difficult to construct a corresponding whitening filter explicitly.

Three types of the parametric model of a random process are used, namely autoregressive (AR), moving-average (MA) and autoregressive moving-average (ARMA). The AR model is good at representing spectral peaks, while the MA model is good at spectral valleys. The ARMA model is well adapted when the PSD of a process has both peaks and valleys, and in general can yield a PSD estimation with a fewer model order (sum of its AR order and MA order) than the AR or MA model. However, the estimation of an ARMA model is more costly than the AR model of an equivalent order, because its MA part is emulated by a larger order AR process to estimate its parameters [Kay, 1988]. Also, any MA or ARMA process can be represented by an AR process of possibly infinite order [Wold, 1954]. For these reasons, the AR model is the most widely used in estimating the PSD of a random process and is appropriate for our need to model the harmonic signal that has a peaky spectrum.

Given an order p AR model, AR(p), the random process $x[n]$ is described by the following time series equation:

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + w[n], \tag{7.32}$$

where a_k is the kth coefficient of the AR(p) model. Its shaping filter is all-pole:

$$G(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{7.33}$$

Note that $H(z)$ is equal to $A(z)$ and that $H(z)$ is all-zero. Thus, the whitening filter is simply an MA process with the same coefficients:

$$w[n] = x[n] + \sum_{k=1}^{p} a_k x[n-k]. \tag{7.34}$$

When the coefficients are known, the PSD of $x[n]$ is given by:

$$P_{xx}(e^{j\omega}) = \frac{\sigma^2}{\left|A(e^{j\omega})\right|^2}. \tag{7.35}$$

Under the AR(p) model, the goal is to select the model order p and the coefficients a_k to best represent the true PSD of $x[n]$, which in turn allows the construction of the whitening filter of $x[n]$. The AR or all-pole model is widely used in speech processing to model a sound [Lim and Oppenheim, 1978] and in geophysical signal processing to model a reverberated seismic source [McClellan, 1988].

If we multiply both sides of the autoregressive equation by $x[n - l]$ ($l \geq 0$) and take expectations, we obtain the following equation of correlations:

$$r_{xx}[l] = -\sum_{k=1}^{p} a_k r_{xx}[l-k] + r_{wx}[l], \tag{7.36}$$

where $r_{xx}[l] = E\big(x[n]x[n-l]\big)$ is the autocorrelation of the process at lag l. The equation can be further simplified by evaluating the crosscorrelation:

$$\begin{aligned} r_{wx}[l] &= E\big(w[n+l]x[n]\big) = E\left(w[n+l]\sum_{k=-\infty}^{\infty} g[n-k]w[k]\right) \\ &= \sum_{k=-\infty}^{\infty} g[n-k]E\big(w[n+l]w[k]\big) = \sum_{k=-\infty}^{\infty} g[n-k]\sigma^2\delta[n+l-k] \\ &= \sigma^2 g[-l]. \end{aligned} \tag{7.37}$$

Because $g[k]$ is a causal filter ($g[k] = 0$, $k < 0$), the crosscorrelation is non-zero only when $l = 0$. Combining the results, we obtain the celebrated Yule-Walker equations:

$$r_{xx}[l] = \begin{cases} -\sum_{k=1}^{p} a_k r_{xx}[l-k] + \sigma^2 & l = 0 \\ -\sum_{k=1}^{p} a_k r_{xx}[l-k] & l > 0 \\ r_{xx}[-l] & l < 0\ , \end{cases} \tag{7.38}$$

where the last equality follows from the symmetric property of a real autocorrelation. In matrix format,

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[-1] & \dots & r_{xx}[-p+1] \\ r_{xx}[1] & r_{xx}[0] & \dots & r_{xx}[-p+2] \\ \vdots & \vdots & & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \dots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix} \text{ and } \sigma^2 = r_{xx}[0] + \sum_{k=1}^{p} a_k r_{xx}[k]. \tag{7.39}$$

Because the correlation matrix is Toeplitz, it can be efficiently inverted by the Levinson algorithm [McClellan, 1988] to find a_k. The Levinson recursion is summarized below:

$$\begin{aligned} a_{11} &= -\frac{r_{xx}[1]}{r_{xx}[0]} \\ \sigma_1{}^2 &= (1 - a_{11}{}^2) r_{xx}[0]. \end{aligned} \tag{7.40}$$

For $k = 2, 3, \dots, p$,

$$\begin{aligned} a_{kk} &= -\frac{r_{xx}[k] + \sum_{l=1}^{k-1} a_{k-1,l} r_{xx}[k-l]}{\sigma_{k-1}{}^2} \\ a_{ki} &= a_{k-1,i} + a_{kk} a_{k-1,k-i} \quad i = 1, 2, \dots, k-1 \\ \sigma_k{}^2 &= (1 - a_{kk}{}^2) \sigma_{k-1}{}^2, \end{aligned} \tag{7.41}$$

where $\sigma_k{}^2$ is the estimation of σ^2 and a_{ki} is the estimation of a_i at the kth iteration. The recursion yields the solution of the Yule-Walker (YW) equations when k reaches p.

A practical concern is the estimation of $r_{xx}[k]$, because the true autocorrelation is usually unknown. It can be shown that the ML estimation of the AR parameters can be found by using the following biased sample autocorrelation function [Kay, 1993]:

$$\hat{r}_{xx}[k] = \begin{cases} \frac{1}{N} \sum_{n=0}^{N-1-|k|} x[n]x[n+|k|] & |k| \le N-1 \\ 0 & |k| \ge N. \end{cases} \tag{7.42}$$

The resulting AR model is stable.

Other AR parameter estimation methods are also available, such as the Burg method (or maximum entropy spectral estimation) and the unconstrained least-squares method [Proakis *et al*., 2002; Ulrych and Clayton, 1976]. The Burg method usually yields a better frequency resolution, at a slightly higher computational cost, than the YW method, especially when the data length is short. However, the Burg method exhibits spectral line splitting at high SNR and spurious peaks at high order models [Proakis *et al*., 2002]. The unconstrained LS method generally yields a better PSD estimation with sharp spectral peaks than the YW and Burg methods, and has a computationally efficient algorithm comparable to the Levinson recursion [Marple, 1980]. Unlike the Burg and YW methods, the unconstrained LS method does not guarantee the stability of the AR model.

When the stability of the AR model is not a concern, the unconstrained LS method is often the best choice. However, experiences show that when the signal has a combination of a few large spectral peaks and many small peaks, the unconstrained LS method with a moderate model order tends to focus on the former, creating very sharp peaks, while ignoring the latter. Because the harmonic signal present in power waveforms has numerous small peaks at higher frequencies, the resulting whitening filter constructed from the unconstrained LS method often over-compensates low frequency components while under-compensating high frequency components. Also, the underlying time-varying nature of the harmonic signal creates relatively broad peaks at harmonic frequencies, negating the necessity of sharp peak estimations. Thus, the unconstrained LS method is better suited for a stationary signal with a few large amplitude sinusoids. For the harmonic signal we observe in the power waveforms, the YW method or the Burg method is a better candidate. Because we can select the data length of a reasonable size ($N \sim 200$), we adopt the YW method for the remainder of this section.

Another aspect in the use of the AR model is the selection of the model order p. The model order can be determined by trial and error, but it is helpful to have an order selection criterion. It also is necessary to check the model order occasionally in real time when the signal has a dramatic time-varying character, i.e. when the harmonic amplitudes and the number of harmonic sets change with time. The *Akaike information criterion* (AIC) is one of the well-known AR model order selection criterions [Akaike, 1974]. It minimizes

$$\mathrm{AIC}(p) = \ln \hat{\sigma}^2(p) + \frac{2p}{N}, \tag{7.43}$$

where $\hat{\sigma}^2(p)$ is the estimation of σ^2 at model order p. As p increases, $\hat{\sigma}^2(p)$ decreases rapidly, reducing the overall AIC. However, as p passes a certain threshold value, the

decrease in the noise variance becomes negligible, increasing the AIC. Hence, the AIC reaches a minimum at a certain p, which is selected as the AR model order. Other criteria are also available, such as the minimum description length (MDL) criterion [Rissanen, 1983] and the autoregressive transfer (CAT) function [Parzen, 1974].

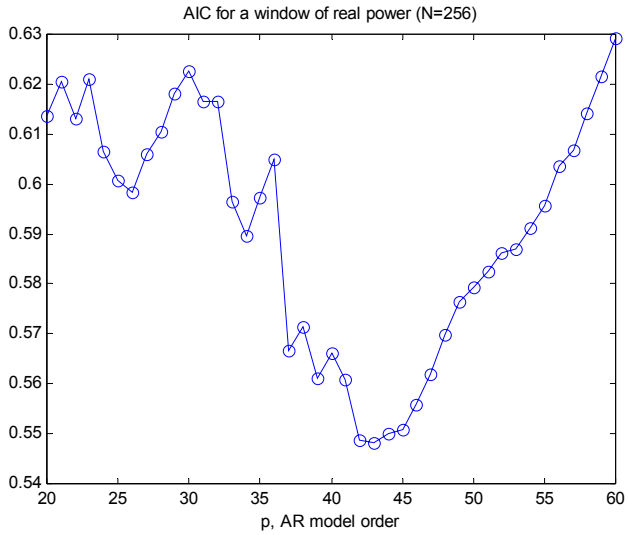


Figure 7-25 Akaike information criterion of a 256-point window of real power containing two harmonic sets.

Figure 7-25 shows the AIC applied to a window of real power waveform holding two harmonic signals (same data as in Figure 7-1). The AIC becomes minimum at $p = 43$. Tests show that $p = 40$ is the minimal AR order to estimate reliably the PSD of the power waveform containing two nominal harmonic sets. In this chapter, $p = 50$ is typically employed for white filtering.

Because both the real and the reactive power require white filtering, we have to construct two filters. We can either construct a single whitening filter, e.g. from A_5, and use it for both channels, or construct two filters separately based on a separate PSD estimation of each channel. The former is attractive when computational budget is tight. However, its white filtering result is slightly inferior to that of the latter, because P and Q have different harmonic amplitude distributions although they share the same harmonic frequencies. In this chapter, we employ the separate estimation and filtering approach for the real and reactive power channels.

When the AR parameter estimation and white-filtering is performed on a continuous stream of data, the data are divided into size-N windows, with or without overlap, and the estimation and filtering is executed on a block-by-block basis. The block processing method is simple to implement and yields satisfactory results in most cases. However due to the discontinuity of AR parameters at block boundaries, the innovations typically show larger variances in the beginning of a window. Sequential estimation methods (and subsequent white filtering) also exist, which update AR model parameters sequentially as a new data point becomes available. They rely on the lattice filter realization of an all-

pole system and the recursive estimation of the filter parameters (or prediction coefficients) [Proakis *et al*., 2002]. In this chapter, we use the block processing method for its simplicity.

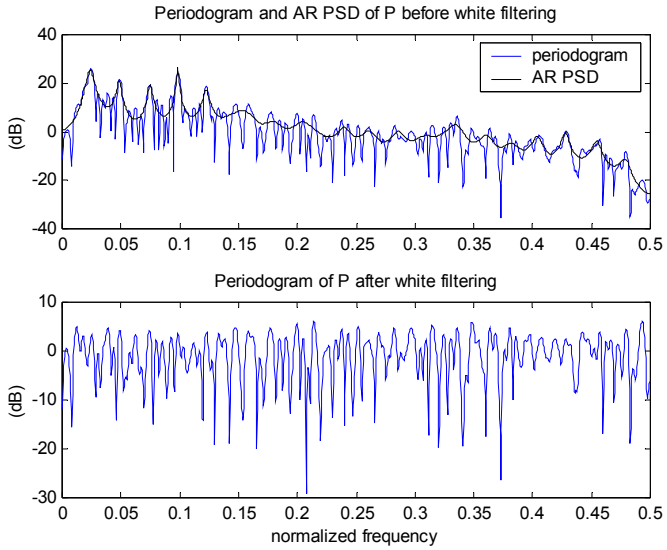


Figure 7-26 PSD estimation of real power based on all-pole model and periodograms before and after white filtering.

Figure 7-26 shows the result of the PSD estimation for a segment of real power (Figure 7-1 data) based on the AR model of order 50, along with the periodograms before and after white filtering. The PSD conforms very well to the first five large peaks of the periodogram (top graph), without creating too sharp or too high peaks, and generally follows the contour of the small peaks at high frequencies. The comparison of the PSD and the periodogram hints that the AR(50) model combined with the YW method is amenable to estimating the true PSD of the real power. The bottom graph shows the periodogram of the real power after white filtering. The periodogram is relatively flat at 4 dB (whitened), implying a good white filtering result.

Figure 7-27 shows the results for a window of reactive power (Figure 7-1 data). Again, the PSD estimated from the AR(50) model via the YW method conforms the shape of the periodogram very well. We also notice the difference in the PSDs between the real power and the reactive power, demonstrating the necessity to construct two separate white filters. The bottom graph shows the periodogram after white filtering. It is flat at 0 dB.

Figure 7-28 shows the standard deviations of the channels after white filtering. Thirty windows of length-256 are used without overlap. Compared to the case where the harmonic signals are estimated and subtracted, e.g. Figure 7-13, the standard deviations are substantially reduced and show much less variations over windows. Thus, we verify that the white filtering based on an AR signal model is a better option to condition the observation for a further analysis, when the embedded harmonic signals do not need to be estimated.

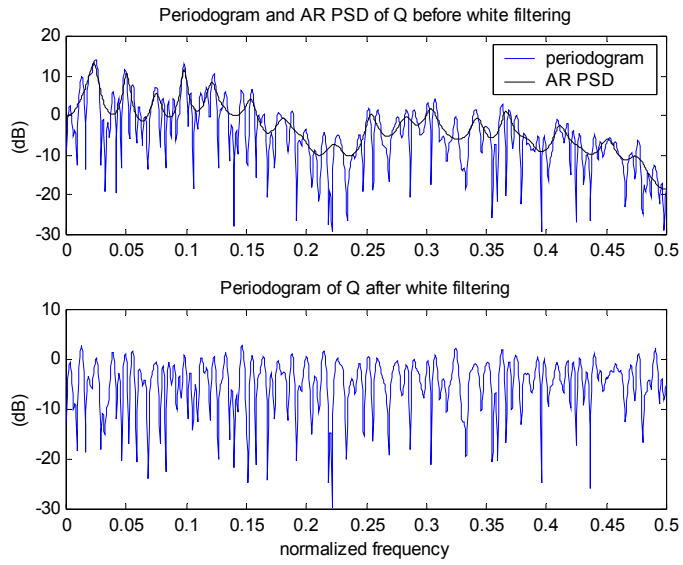


Figure 7-27 PSD estimation of reactive power based on all-pole model and periodograms before and after white filtering.

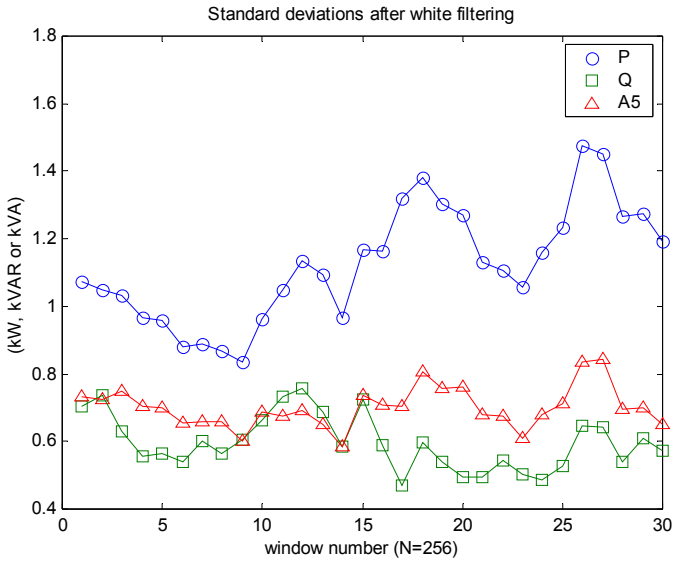


Figure 7-28 Standard deviations after white filtering.

7.6 Chapter summary

In this chapter, four different methods were investigated to eliminate the harmonic signals embedded in the power waveforms. The first three methods treat the harmonic signal as the signal of interest and rely on its explicit estimation and subtraction from the observations. The last method treats it as a colored noise and constructs the whitening filter based on the explicit estimation of the PSD with a parameterized model. All four methods achieve the common goal of suppressing the harmonic noise, which was often viewed as a nuisance in processing power signals for load monitoring.

Because the first three methods estimate the harmonic signal explicitly, they are well adapted for load monitoring and fault detection, even though they perform poorly from the viewpoint of white filtering. Due to the physical origin of a harmonic signal, i.e. the load imbalance of a rotating machine and accompanying resonance mechanism, it can be used to monitor the operational status of its source. Its pitch estimation can be used as a virtual tachometer and the amplitude estimation can be employed to assess the severity of the imbalance. With an appropriate postprocessing decision-making algorithm, the knowledge gained from the harmonic signal estimation can play a significant role in the overall electric load information system based on non-intrusive power monitoring, for the objectives of load status identification, fault detection, optimization, etc.

An interesting observation is the analogy between the harmonic signal in the power waveforms and human speech. Although they have remarkably different physical origins, rotating machine vs. vocal tract, they share the common mechanism of pulsation and resonance. Speech is used to convey an extraordinary amount of information, and thus has a very sophisticated structure, e.g. pitch, formant, voiced/unvoiced, word, sentence, etc. The harmonic signal is not comparable to speech in terms of complexity, but nonetheless can carry a certain amount of low-level information. Strictly speaking, the harmonic signal has the quality of pitch and formant found in the speech. Humans can distinguish different speakers based on their distinctive pitches and formants. Harmonic sets can be identified by their characteristic fundamental frequencies and amplitude distributions. One can perceive that a person is in a different mental or physical state based on his or her abnormal pitch and formant. The electric load information system can monitor the health of a load by analyzing the harmonic signal generated by the load.

The first method introduced a harmonic signal detector and its integer pitch period estimator based on the periodogram of the observation. The signal was retrieved by the ML estimator. The harmonic detector performs reasonably well, but it is cumbersome because it requires *a priori* knowledge of the additive white noise covariance and a large frequency population in computing the periodograms. Also, it often shows period doubling errors, making the pitch estimation completely unreliable.

To overcome the problems associated with the harmonic detector based on periodograms, the second method employed the ML pitch detector based on autocorrelations. The tests show that the ML detector is less prone to the period doubling error. The ML detector is also more convenient to use, because it does not require the variance of the additive noise and is computationally more efficient. A sequential harmonic detection and estimation strategy was adapted to handle multiple harmonic sets present in an observation. Interference in estimating two harmonic sets was also observed. Spectral leakage occurs due to windowing, spectral proximity between the signals and their time-varying nature. A better windowing can reduce the spectral leakage, but can not eliminate it completely. We observed the fluctuations in the pitch estimation of the second harmonic set, due to the spectral leakage in estimating two harmonic sets sequentially.

The third pitch period estimator was introduced to allow a fractional pitch period estimation, based on the harmonic suppression filtering. A simple least squares harmonic signal estimator was also developed, because the ML estimator can not handle a fractional pitch period (without interpolation). Although the FP period estimation and signal estimation is an improvement compared to the former method, the test results are not satisfactory in the sense that the standard deviations after the subtraction are increased and that the FP period estimations showed large variations. The LS signal estimator can be partially blamed because it is not optimal and typically underestimates signal amplitudes. However, the ambiguity of the FP pitch estimation is more liable to the underperformance, caused by the combination of the time-varying nature, low SNR, windowing, interference between harmonic sets and signal distortion introduced by the power coder. These issues need to be investigated in depth to develop a better FP period estimation and harmonic signal estimation technique. Until then, the ML detector and signal estimator based on an integer pitch period assumption is the best tool to capture the harmonic signal present in power waveforms.

The fourth method changed the perspective and treated the harmonic signal as colored noise to be removed by white filtering with a filter constructed from the parametric estimation of its PSD. The signal was modeled as the output of an all-pole system excited by white noise and the Yule-Walker method was used to obtain the AR parameters via the Levinson recursion. The model order of 50 was chosen from the evaluation of the Akaike information criterion and experience. The resulting PSD estimation shows a good conformity to the periodogram, without creating too sharp or too high peaks. When the observation is filtered by the whitening filter constructed from the AR parameter estimations, its spectrum is relatively flat, indicating the high performance of the white filter. We also note that the white filtering removes the formant (overall PSD contour) present in an observation, which is impossible in the former three methods because they can not reshape the PSD over the whole spectrum but only comb at a few harmonic frequencies.

Another advantage of the white filtering is that it can be used irrespective of the presence of a harmonic signal. Thus, it does not require a harmonic detector and can be placed in front of a load status analyzer to whiten the incoming signal at all conditions (possibly with a real-time model order selector). The white filtering based on the AR PSD estimation yields smaller and more stable standard deviations of the filtered signal compared to the harmonic signal estimation and subtraction methods. Thus, it is a better alternative to eliminate the harmonic noise when the signal does not need to be estimated. It also is possible to combine the white filtering and one of the harmonic signal estimation and subtraction methods. For example, the residual signal from the second method may be white filtered, to remove the formant and eliminate any harmonic spectral content leaked into the residue.

Chapter 8

Test Results, Conclusion and Future Work

8.1 Test results

A real-time constant load disaggregator was developed, which can handle multi-load events. The disaggregator displays current power consumptions graphically on its web page, and identified load events can be viewed from a linked web page. Its software architecture is explained in Appendix D. The disaggregator was tested using a set of offline data, recorded from 160 Sansome St., San Francisco, CA. Figure 8-1 shows an example of its event history page. The list of the monitored loads is given in the top portion of the page. Each identified load event is explained both by a graph and a tag. The graph displays an event by a red line and a matched exemplar by green dots. The tag is shown to the right of a graph, including the number, identification and time of the event. The load status after the event is also displayed. Figure 8-2 shows another example where a multi-event was successfully captured along with a series of other events.

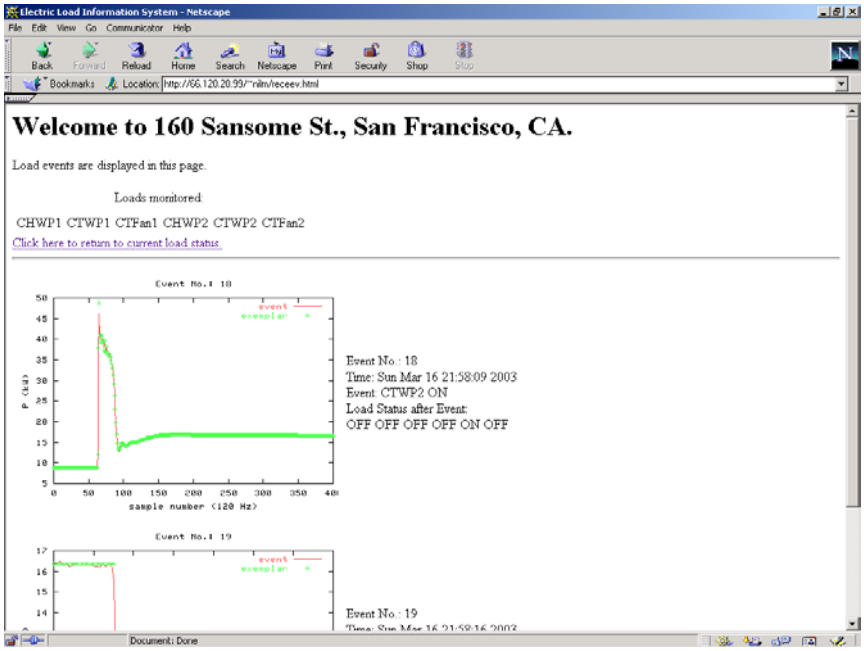


Figure 8-1 Events identified by constant load disaggregator.

On the other hand, an NIPM is installed at the laundry room of Next House, an MIT undergraduate dormitory located at 500 Memorial Drive, Cambridge, MA. The laundry room has a three-phase 208 V service to power up its ten dryers. A dryer has a 120 V single-phase motor (phase-to-neutral) and a 240 V heater (phase-to-phase). The NIPM currently samples the phase-A current only and nine dryers were observable from the power signal. Among those, three dryers have their motors connected to the phase-A. The rest six dryers appear as pure heaters in the power signal.

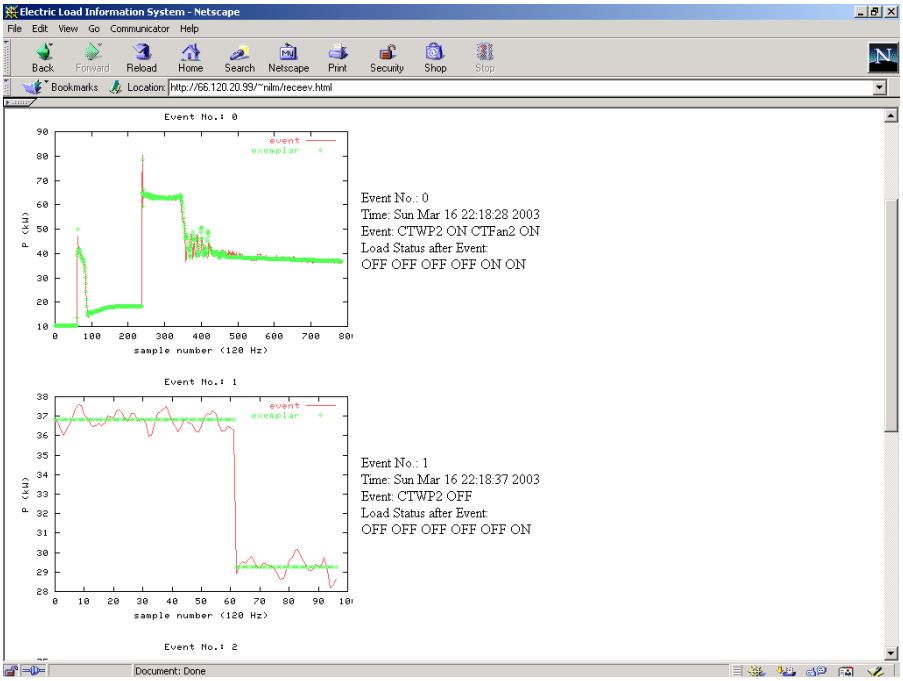


Figure 8-2 Multi-event identified by constant load disaggregator.

The dryer is an exemplary multi-stage load (OFF, Motor and Heater ON, Motor ON), and its heater goes through a series of ON-OFF cycles, based on its thermostat setting, while its motor runs continuously. Also, although the dryers can be grouped based on their steady-state and transient responses, they typically show almost identical responses within a group. Thus, to identify the load status of a place like the Next House laundry room, the load disaggregator should be able to handle the multi-stage and multiple identical loads. The load disaggregator developed in this thesis currently can not handle these situations.

However, to test the capability of the load disaggregator in a real situation, it is modified to identify load events only without an underlying load state model, so that it can be used as an event tracker at Next House. It can recognize a dryer ON or OFF event, but it can not answer how many dryers are on at a given moment. Because it does not have a state model, the ML classifier based on the Gaussian random process model is not applicable. Instead, a simple Euclidean distance classifier is used to classify step change observations. The transient pattern classifier based on the LS error criterion is still applicable.

Figure 8-3 shows an event history page of the event tracker. The first event (event no. 365) is a dryer ON event, whose motor is connected to the phase-A. The next event is another dryer ON event, but because its motor is not connected to the phase-A, it is simply reported as a heater ON event. An event tag shows the event number, time, description and step magnitudes. Figure 8-4 shows the power consumption behavior of last twenty hours. It illustrates that Next House students are active in laundry after midnight, in the morning and in the afternoon (on Sunday).

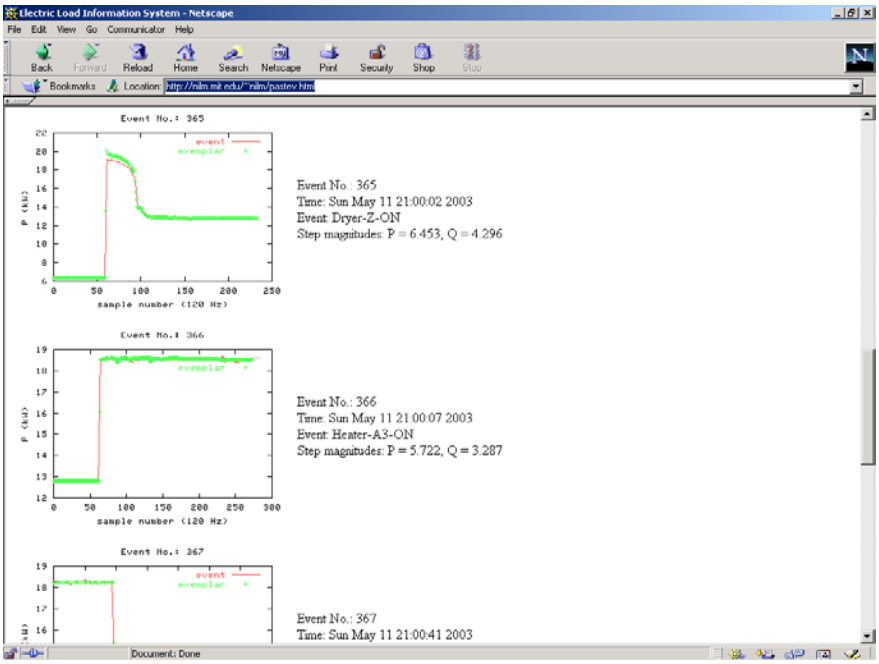


Figure 8-3 Load events identified by event tracker.

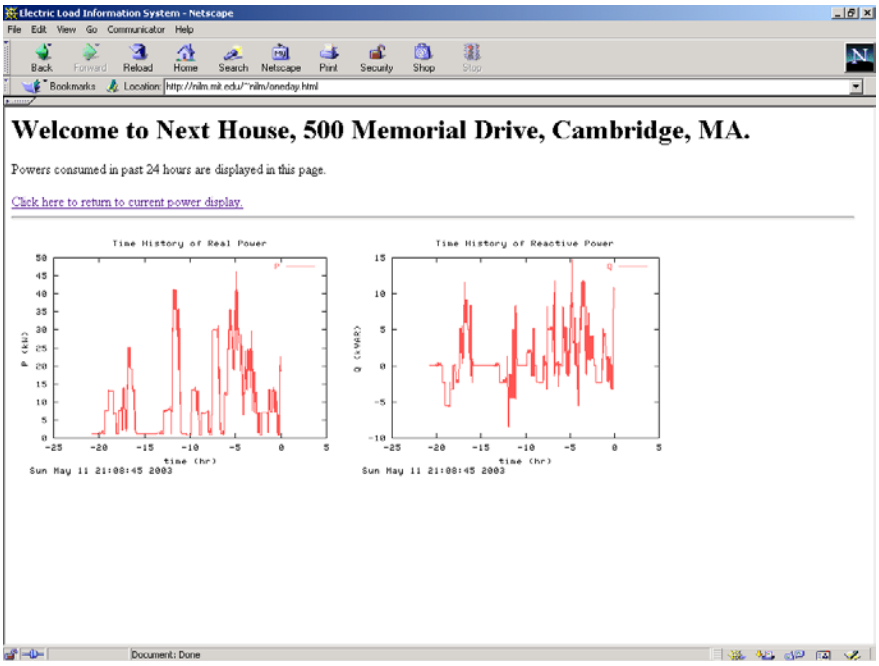


Figure 8-4 Event tracker history page.

A real-time web based VSD load tracker was developed and successfully tested at the Sansome building. Its software architecture is also explained in Appendix D. Figure 8-5 shows its main page. The bottom graph shows the fifth apparent power in red line. Its first order mean estimation is shown in solid green line. The red line in the upper left graph is the real power and the solid green line is the estimation of the real power consumed by the VSD obtained from the A_5 mean estimation. The real power consumed by non-VSD loads (background load in this case) is shown in blue line. Note that the

non-VSD power is white filtered, based on the autoregressive parameter estimation of the PSD of A_5. The upper right graph shows the VSD tracking result of the reactive power.

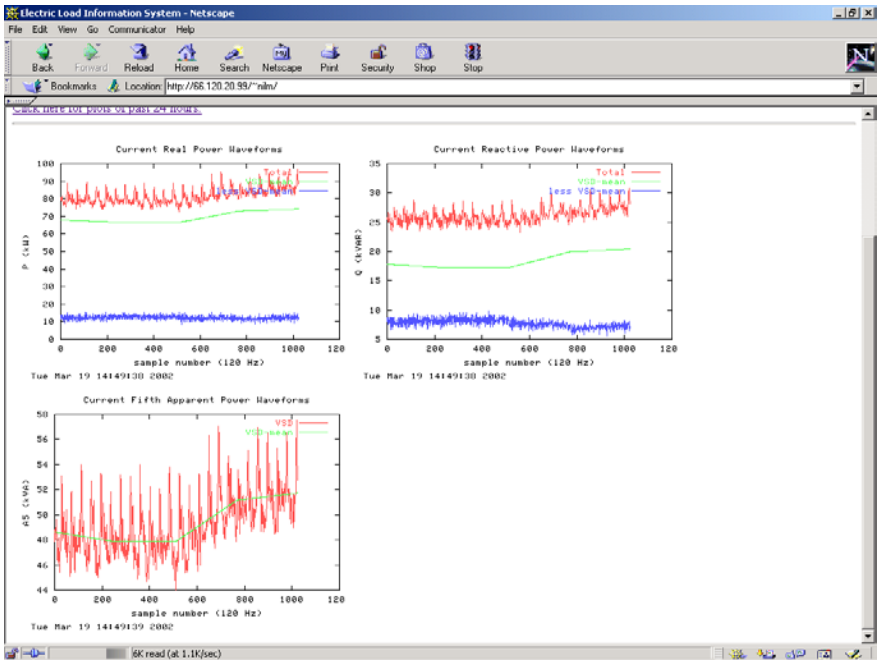


Figure 8-5 VSD tracker main page (18:25 4/2 2003 EST).

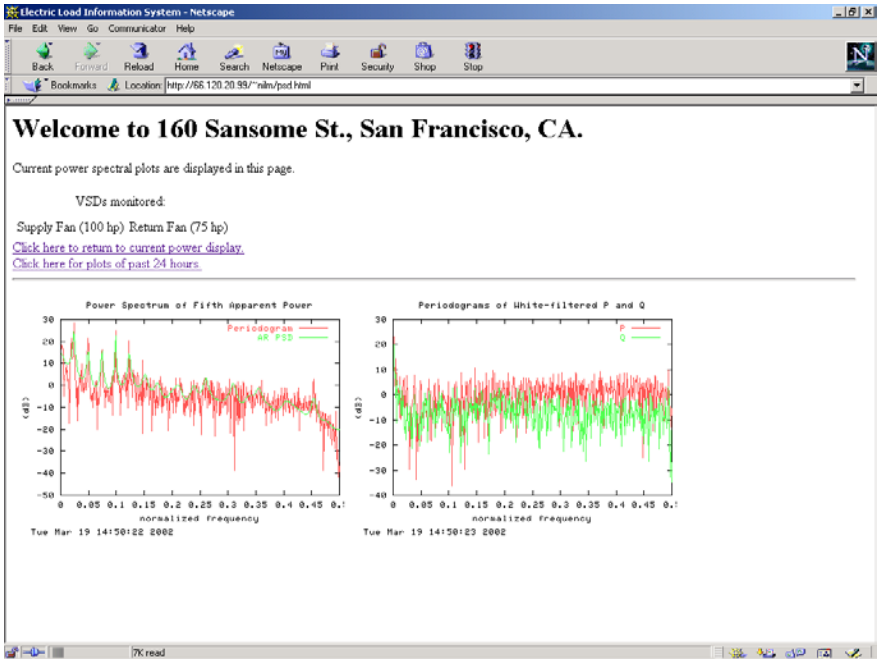


Figure 8-6 VSD tracker PSD estimation page (18:25 4/2 2003 EST).

Each graph is based on 1024 data points, and with the 120 Hz power coding rate, the main page is updated at every 8.53 second. The time server of the computer has a problem, and as a result, the time stamps in the graphs were not correct. Correct observation time is given in the caption of the figure.

Figure 8-6 shows the PSD estimation page of the VSD tracker. The left graph shows the periodogram of A_5 in red line obtained from its 1024 data points. The parameters of the autoregressive model of order 50 are estimated from A_5 and the resulting PSD estimation is shown in solid green line. The periodogram clearly shows six main harmonic peaks, and the PSD estimation follows the contour reasonably well. The right graph shows the periodograms of the white filtered P and Q (blue lines in Figure 8-5). The parameters obtained from the A_5 PSD estimation are used to construct a whitening filter. As a result, the periodograms shows slight height changes, rather than being flat, because of the differences in the formant of A_5 and those of P and Q. However, we note that the harmonic peaks are completely removed.

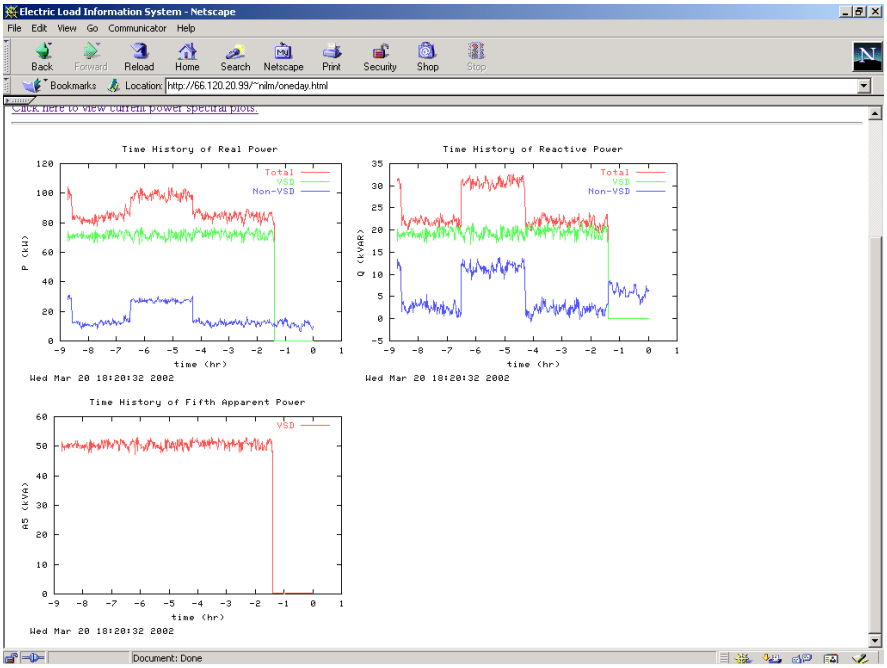


Figure 8-7 VSD tracker history page (21:57 4/3 2003 EST).

Figure 8-7 shows a history page of the VSD tracker. Samples are taken in one minute intervals (as the average of 1024 points). Due to large downsampling and averaging, the fifth apparent power seems to be almost constant. The graphs show that there was an activity of a constant load, evidenced by the step changes in the real and reactive power waveforms. However, the VSD fundamental power estimations were unchanged, because the fifth apparent power was unaffected. As a result, the step changes are conserved in the non-VSD leftover power signals. This example clearly shows that the VSD tracker can separate the fundamental powers consumed by VSDs and non-VSD loads.

Figure 8-8 shows another history page. This figure is interesting in many aspects. A large step change set in the afternoon is a chiller activity. The 160 Sansome building has two large chillers to meet its cooling load. Each chiller has a constant speed compressor, but adjusting its inlet vane position can vary its capacity. Thus, the chiller can be categorized as a non-VSD variable load. Unlike constant loads, non-VSD variable loads can draw higher harmonic powers, as evidenced by a dip in the A_5 history graph. However, the

higher harmonic power consumed by a non-VSD variable load is almost constant during its operation, although its fundamental power consumption can vary. Figure 8-9 shows this behavior more clearly.

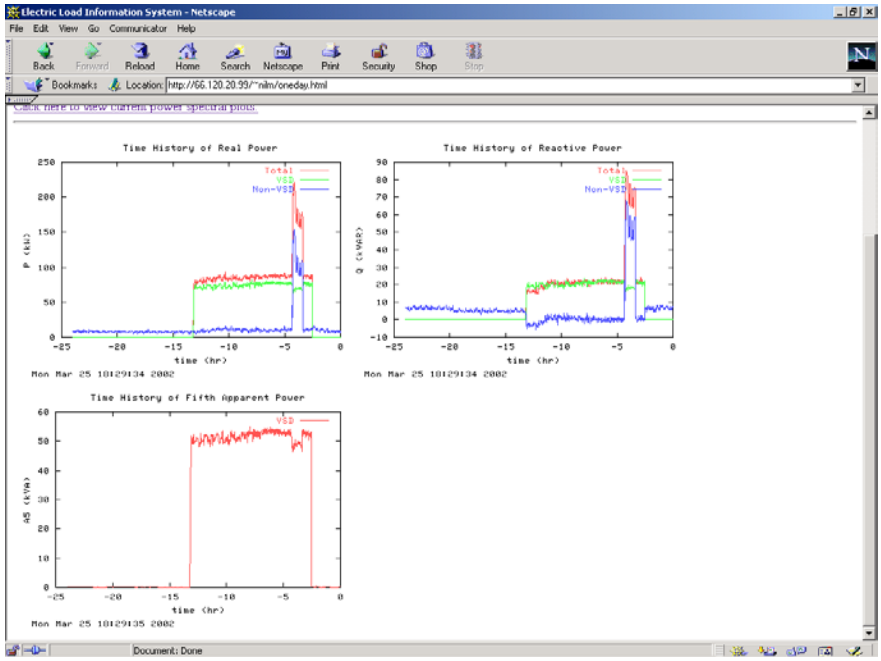


Figure 8-8 VSD tracker history page (23:05 4/8 2003 EDT).

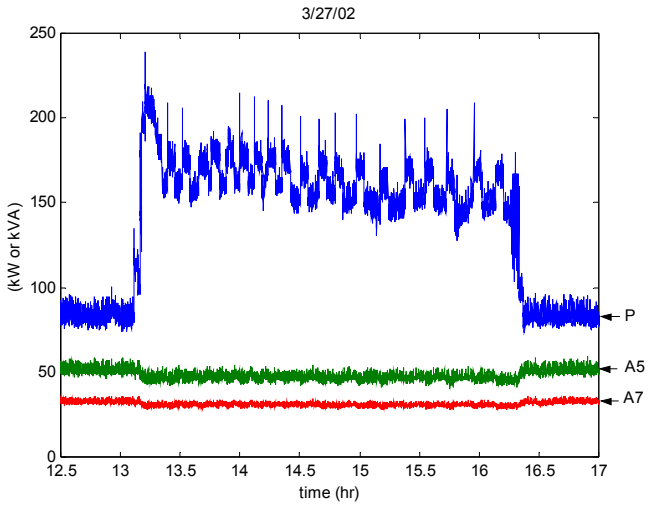


Figure 8-9 Change in higher harmonic powers when chiller is on.

In order to estimate the fundamental powers consumed by VSDs reliably, any (constant) amount of higher harmonic power consumed by non-VSD variable loads needs to be detected and compensated. Figure 8-8 shows that there are sudden drops in the VSD fundamental power estimations when the chiller is on, because the compensation was not performed. The compensation will require a type of change-of-mean detection and estimation in the stream of A_5. The general change-of-mean/variance detector developed in Appendix A may be employed for this purpose.

Another observation is that the reactive power of the VSDs is slightly overestimated. This is evident from the fact the leftover non-VSD reactive power is almost zero, whereas at nights the reactive power is of substantial amounts. There may have been some changes since the correlation graph was measured. Updating the correlation graph will partially solve the problem. Another observation is that the reactive power is more overestimated in the morning. The HVAC load in the early morning is different from that in the afternoon, i.e. when the building is fully occupied and the ambient temperature is high. Thus, the VSD set will see different load conditions. It may be that the VSD set has two different correlation curves and as the loading condition changes, the correlation moves from one curve to the other. Thus the solution to this problem will be a set of correlation curves and educated decisions about which correlation curve the VSD set is currently on, based on various observations.

8.2 Conclusion

This thesis is a great accomplishment, because it has shown that information about the status of electric loads can be obtained from the non-intrusive power signals and because several signal models and signal processing and information handling algorithms were successfully developed to extract the information. The idea of monitoring electric loads non-intrusively has been around for more than two decades. But it has not been widely received by the academic or technical community, because it has lacked a concrete scientific foundation and because the benefit or usefulness of the information offered by a non-intrusive monitor has not been clear. Its practicality was also a lingering question. Even when the monitor was employed as a residential load disaggregator, its performance was limited because of its rudimentary steady-state load model. It could not handle overlapping load events and lacked a state estimation capability. The variable loads were declared intractable. This thesis is a partial answer to these challenges.

Traditionally, only a few power terms, such as the real and reactive powers, were measured to monitor the electric loads. The current harmonic powers, defined in this thesis, are a more complete and systematic tool to characterize a current waveform that contains rich information about the electric loads. They have a solid mathematical background of Fourier analysis and modern discrete-time signal processing. They can be viewed as an efficient coding process to capture most of the information present in a current waveform with a few selected Fourier bases. Analyzing current harmonic powers is more economical than analyzing raw current waveforms, due to the reduction in the number of samples, and looking at the behaviors of the electric loads from the harmonic dimension provides an insight and a mathematical basis to characterize them.

The concept of voltage harmonic intensity was also introduced, to quantify the harmonic content of a voltage waveform. It can be computed using the same method of the current harmonic power, and its time series will possibly provide additional information about the electric loads. The current harmonic power and the voltage harmonic intensity can be

used in a complementary manner to better characterize electric loads and to extract relevant information.

Earlier works considered electric loads to be more or less deterministic, as evidenced from the approaches to track constant loads based on the steady-state power level changes, although people agreed that there can be variations in their observations. This thesis introduced a stochastic model to describe a power signal generated by an electric load or a collection of them. When the power signal of an electric load is modeled as a random process, it has both the property of a mean to describe its steady-state power consumption and the property of higher order moments, such as variance, to describe its noise behavior. Thus, each load becomes more distinct when modeled by a random process, than when modeled by a deterministic process. The stochastic process has a rich mathematical background and makes it possible to describe the situations, such as the temporal overlap of multiple events and the steady-state of multiple loads, in clear mathematical terms. The mathematical description in turn allows developing various detection, estimation and classification algorithms to identify those situations.

The harmonic powers generated by a constant load at its steady state were modeled as a Gaussian random process. The model is not perfect in the sense that a real statistical distribution is not necessarily a strict Gaussian and even so, its characterizing moments, i.e. mean and variance, can change with time. However, this model is crucial in expressing the observations in various situations in mathematically tractable terms. Based on this model, event observations (step changes) and steady-state observations can be explained as linear combinations of individual Gaussian processes, which in turn follow Gaussian distributions. Based on these Gaussian distribution models of events and steady-states, the uni-event classifier, multi-event classifier and load state estimator were successfully developed.

The transient response of a load is another asset often overlooked in the past. Although a few researchers acknowledged its potential value, no significant effort has been made to exploit it systematically and in a complimentary manner to the steady-state step changes. The transient response is not explicitly modeled as a stochastic process, due to the difficulty in modeling and in obtaining its parameters. Rather, a few representative responses are used as exemplars or templates, and the classifiers based on the method of least squares are developed. The advantage of the least squares method is that it relies on an implicit statistical noise model and that when the noise is white Gaussian it is the same as the optimal estimator based on an explicit noise model. The transient pattern classifiers were developed to handle both single and simultaneous load events.

By utilizing two different aspects of a constant load, steady state and transient, in a constructive and complementary manner, it was possible to design a more accurate load disaggregator, which can handle single load events more reliably. It can also handle multi-load events and identify load states without load events, both of which were impossible before. Also, the disaggregation is performed locally in real time, as compared to the previous central and batch processing. The load disaggregation scheme proposed in this thesis is a quantum leap compared to the past practices.

The VSDs were previously considered intractable, because they do not have clearly defined steady states. However, by watching the behavior of a VSD from the whole dimension of current harmonic powers, it is possible to track its fundamental power consumptions. There exists a correlation between a higher harmonic power and a fundamental power, and the former is unobstructed by the activity of other constant linear loads. The VSD is still modeled as a stochastic process, but its mean and covariance change with time and its noise is not white Gaussian. To simplify the model, the covariance matrix was assumed to be proportional to the square of the real mean power, and the harmonic mean powers were considered as correlated. The correlations were obtained from the observations and they were successfully used to estimate the fundamental mean power consumptions from the estimation of a harmonic mean power.

The noise generated by a VSD was modeled as the sum of a white Gaussian source and colored noise. The colored noise was removed by white filtering. Once the mean and colored noise of a VSD are removed from the fundamental power observation, the left-over is simply the fundamental power consumed by non-VSD loads plus the white noise of the VSD. Subsequent (constant) load disaggregation can be performed on this non-VSD signal, because it now follows a Gaussian model, with the residual white VSD noise taken into account. The VSD tracking scheme proposed in this thesis is a first practical solution to estimate the fundamental VSD powers non-intrusively, and is compatible with the constant load disaggregation scheme proposed in this thesis.

The estimation and elimination of the harmonic signal observable on the power waveforms was first started as a byproduct of the VSD tracking research, because it was the colored noise to be removed. However, it has the physical origin of the mechanical load imbalance of the output shaft of a VSD and thus carries the information about the status of the drive. The estimation of the harmonic signal allows us to quantify the severity of the imbalance from its amplitude and monitor the shaft rotation speed from its fundamental frequency. Its elimination makes it possible to apply the various constant load disaggregation techniques on the leftover signals.

The harmonic signal was also modeled as the output of a stochastic source. Once this model was adopted, numerous estimation and filtering techniques were readily available because the estimation of a harmonic signal and its PSD has been one of the central problems in signal processing. Three different methods to estimate a harmonic signal were presented: periodogram, autocorrelation and fractional pitch period. The periodogram-based method works reasonably well, but it suffers period doubling errors and its computational cost is rather high. The autocorrelation method is the best choice and is also widely used in speech processing. The fractional pitch period estimation coupled with the least squares signal estimation has the potential to improve the overall quality of the harmonic signal estimation. However, its test result in this thesis is unsatisfactory, possibly because of the time-varying nature of the signal and the biases introduced by the power coder. One explicit white filtering method was also introduced, which relies on the autoregressive signal shaping model and the estimation of its parameters using the Yule-Walker equations to construct the inverse whitening filter.

Various accomplishments have been made in this thesis, such as putting the current harmonic powers on a sound mathematical basis, characterizing electric load behaviors in the harmonic dimension, adopting random process models to explain steady-state power observations, combining transient and steady-state responses for the common goal of load identification, etc. Specifically, a real-time constant load disaggregator was developed which can handle single load events, multiple load events and state estimations without load events. A VSD load tracker was developed for the first time in the history of NIPM. Various methods to estimate and eliminate harmonic signals were also developed.

Yet, there still are numerous unsolved problems to develop a commercial-grade ELIS based on the NIPM. The issues presented in this thesis are the most important ones, selectively chosen by the author, so that their solutions can advance the *status quo* of the NIPM research significantly. These solutions lay a solid foundation, based on which the solutions of other unresolved problems shall be developed by future researchers. This thesis will also inspire and motivate them to come up with new approaches for coming challenges and better methods for the problems already presented in this thesis. The information has been and will be the key in developing various algorithms and will remain to be the key motivation in developing NIPM-based applications.

8.3 Future work

Many issues need to be addressed in coming years to make the ELIS based on the NIPM more complete and practical. Some of the issues were conceived in the beginning of the research but the author did not have enough time and resources to handle them, and the rest emerged along the course of the research. This section briefly outlines these issues and proposes preliminary solutions or research directions where possible.

The current harmonic power is a great idea in the sense that it can compress most of the information present in a current waveform to a few selected power terms. However, it is blind to the information that may be embedded in a voltage waveform. It is true that the current waveform carries most of the information about the electric loads, but the voltage waveform also transmits a substantial amount of information, especially about the transient behavior of a load. Voltage waveforms were monitored traditionally from the power quality viewpoint, i.e. to monitor voltage dip, swell, transient, interruption, etc. The voltage harmonic intensity can be used to quantify these power quality indices and to monitor electric loads. Thus a future ELIS will extract richer information from both the voltage harmonic intensity and the current harmonic power generated by the NIPM, and make better decisions concerning the attributes of the monitored electric loads.

The current harmonic powers are computed via the DFT of current waveform samples, following the method presented in Chapter 2. As discussed, it generally is difficult to obtain a consistent number of evenly spaced samples per cycle, due to the generator-sampler synchronization problem. It is currently solved by oversampling, followed by filtering and resampling. However, this approach inevitably introduces errors, such as the

phase biases of the harmonic powers due to the sampling position uncertainties. The current harmonic powers will continue to be computed from the DFT. Thus, it is necessary to develop a more effective solution to this synchronization problem.

The Gaussian random process model was introduced in this thesis to describe the steady-state power observations of constant loads in clear mathematical terms. The Gaussian model was decisive in developing various estimation routines for constant load disaggregation, and the resulting estimators were very effective. However, it is still limited in the sense that a constant load can have slightly different means at different instances and its noise characteristic can change with time. A better stochastic model will be required to address this time-varying nature of a constant load. Also, the steady-state power model does not necessarily have to be stochastic. Someone may come up with a better non-stochastic model in describing the steady-state power observations.

The transient response of a constant load was the other half in developing the load disaggregator. However, the statistical properties of responses were not derived explicitly and only the simple least squares estimator was used to classify patterns. It obviously is true that a better classifier can be designed if the statistical distributions of the transient patterns are known. Future classifiers may have a training period to collect enough information about the pattern statistics and, based on the statistical characterization, come up with a classification metric, such as the maximum-likelihood.

A multi-event handler was developed in this thesis based on the event space search and the transient pattern composition. The event space search can not be said to be optimal, because it relies on the heuristic search of a binary tree to visit possible state candidates. Nonetheless, it usually does not impose too much a computational burden and thus is acceptable. However, the transient pattern composition is not a polynomial algorithm, because of the permutation of the pattern sequences, although dynamic programming was used to reduce the computational complexity significantly. Various heuristic methods will have to be developed to further reduce the complexity. Alternatively, the problem may be formed from a totally different viewpoint, which may do away with the non-polynomial algorithm.

The constant load disaggregation scheme based on both steady state and transient was very effective in distinguishing similar loads, i.e. 10 hp water pumps. However, there are situations where two loads are exactly identical and where similar loads have to be treated as identical for practical reasons, i.e. when there are too many individual loads but they can be categorized to several distinct groups. How to design a constant load disaggregator under this situation is an unresolved challenging but interesting problem.

A constant load has only two states: ON and OFF. However, there are loads that consume a fixed amount of power at a state but the number of states is more than two. Dishwashers and laundry machines are good examples. Each unit has an assembly of smaller electric devices, such as motors, heaters, valves, etc. Depending on which device is on during the course of its programmed use, it consumes different amounts of power at different stages while it is ON. They can be called multi-stage loads. A multi-stage model is required to

classify load events, which can update the stages of relevant loads and thus determine the states of electric loads (and observation).

A very simple VSD tracker was implemented in this thesis, using the correlations between the fundamental powers and higher harmonic powers. Although effective, it sometimes shows a poor estimation result (within about 10 % of true value), due to the time varying nature of the correlations. Future improvements may include collecting different correlation curves and making the tracker to make an educated decision which curve to choose based on observations. The VSD power estimation from higher harmonics can become more accurate by employing a time-varying representation of correlations and an *in situ* selection of the most likely correlation.

The proposed VSD tracker assumes that there exists only a single set of VSDs. It is an interesting problem how to develop a power estimator for multiple VSD sets that can estimate powers either collectively or separately for each set. When multiple VSDs are running, there exist certain constraints for the composition of the overall harmonic powers due to the characteristic of each VSD. The observed harmonic powers may be decomposed, using the constraints, into individual harmonic powers generated by each VSD and thus the corresponding fundamental powers may be estimated.

The feasibility of the multi-VSD power tracking method will depend in large part on the availability of a new power coder with a sufficiently small phase bias. The proposed constraints in harmonic powers will dictate necessary phase precision. On the other hand, the phase accuracy is necessary even to observe the presence of the constraints and to estimate them reliably.

How to handle non-VSD variable loads is also an interesting problem. They may have signatures in a higher harmonic domain, although constant, as was the chiller at the Sansome building. These signatures may be used for detection purposes. If no good method is available to track their individual power consumptions, their collective power consumption, at least, can be estimated by subtracting the powers consumed by the constant loads and VSDs from the observations.

Various estimators and classifiers require *a priori* information of the signal, such as mean and variance, to make necessary decisions. In this thesis, the statistical parameters were obtained separately from manual analyses. However, a commercial grade ELIS will be required to collect *a priori* information by itself from observations, to minimize labor-intensive human interventions.

This thesis has concentrated on the load disaggregation aspect of the ELIS based on the NIPM. Other possibilities were not specifically investigated, such as the diagnosis, power quality monitoring, optimization and power management. To make the NIPM a more competent and diverse tool, the researches to materialize these applications should be pursued in future. With relentless endeavors, the NIPM will soon become a ubiquitous electric load information system.

Appendix A

General Change-of-mean/variance Detector

A change-of-mean detector identifies the presence of a change-of-mean event and estimates the location of the event and the mean and variance after the event (and before the event if necessary), from a window of data. The detector may be able to identify multiple events or a single event depending on the construction. The change-of-mean detector performs an abstraction on the incoming data and groups them as segments, each of which is identified by a unique mean and variance.

In the context of the steady state load classification, the change-of-mean detection can be performed on a single stream (e.g. P or Q only), multiple streams or a composite quantity of the current harmonic powers. However, in most cases, it is suitable to use the P stream only, because virtually all loads consume real power and because any turn-on or turn-off event will accompany a change in the real power level.

The design of a change-of-mean detector depends mostly on whether the location of change and the mean and variance of the segment before the change are known. It is natural to assume that the location of change is not known, and furthermore that whether there is a change-of-mean event also has to be tested.

A simple binary hypothesis test based on a generalized likelihood ratio (GLR) can be designed to detect the presence of a change-of-mean event, when the mean and variance of a prior-event segment is known and the variance does not change after the event [D. Luo, 2001; Basseville and Nikiforov, 1993]. The prior-event mean and variance need to be estimated from a *pre-event* window, and thus they depend on the size and location of the window.

If one tries to detect a steady-state event by suppressing transients with adequate downsampling and filtering, the aforementioned GLR test can be used with a careful choice of pre-event and detection window lengths and mean and variance estimation techniques. However, it is restrictive in the sense that it can detect only a single change-of-mean event, that the mean and variance before the event should be known, and that it can not detect a change-of-variance event.

In this chapter, a general change-of-mean/variance detector is developed. The detector is general in the sense that it can detect an arbitrary number of mean and/or variance changes from a given data window (with an upper limit) and that it does not require any prior information (i.e. means and variances) to perform its job.

The following composite multiple hypothesis test can be conceived for an N-point sequence $x[n]$:

$$
\begin{aligned}
&H_0 : x[n] = A_0 + w[n] \qquad n = 0, \ldots, N-1-q \qquad \mathrm{var}(w[n]) = \sigma_0{}^2 \\
&H_1 : x[n] = \begin{cases} A_0 + w[n] & n = 0, \ldots, n_0 - 1 - q \quad \mathrm{var}(w[n]) = \sigma_0{}^2 \\ A_1 + w[n] & n = n_0, \ldots, N-1-q \quad \mathrm{var}(w[n]) = \sigma_1{}^2 \end{cases}. \\
&\quad \vdots \\
&H_k : x[n] = \begin{cases} A_0 + w[n] & n = 0, \ldots, n_0 - 1 - q \quad \mathrm{var}(w[n]) = \sigma_0{}^2 \\ A_1 + w[n] & n = n_0, \ldots, n_1 - 1 - q \quad \mathrm{var}(w[n]) = \sigma_1{}^2 \\ \quad \vdots & \\ A_k + w[n] & n = n_{k-1}, \ldots, N-1-q \quad \mathrm{var}(w[n]) = \sigma_k{}^2 \end{cases} \\
&\quad \vdots \\
&H_{k_{\max}} ,
\end{aligned}
\tag{A.1}
$$

where k_{max} is the maximum number of mean/variance changes tested for the given size N window and q ($q \geq 0$) is the number of transition points to be ignored between two segments. p is the minimum length of a segment ($p \geq 2$). In general, if j parameters are to be estimated from a segment, $p \geq j$. p and q should satisfy the following relationship: $(k_{max} + 1)(p + q) \leq N$. A_i is an unknown but constant mean of segment i. The zero-mean additive white Gaussian noise (WGN) $w[n]$ has an unknown but constant variance σ_i^2 at the segment i. $n_{k\text{-}1}$ is the starting index of the segment k (with $n_{\text{-}1} = 0$ and $n_{k_{\max}} = N$).

For a hypothesis H_i, unknown parameters are

$$
\mathbf{A}_i = [A_0 \; A_1 \ldots A_i]^T, \; \mathbf{n}_i = [n_0 \; n_1 \ldots n_{i-1}]^T \text{ and } \boldsymbol{\lambda}_i = [\lambda_0 \; \lambda_1 \ldots \lambda_i]^T, \tag{A.2}
$$

where $\lambda_j = \sigma_j^2$. A single parameter vector can represent these parameters:

$$
\boldsymbol{\theta}_i = \left[\mathbf{A}_i{}^T \quad \mathbf{n}_i{}^T \quad \boldsymbol{\lambda}_i{}^T \right]^T. \tag{A.3}
$$

If l_i denotes the number of unknown parameters for H_i (or the length of $\boldsymbol{\theta}_i$), $l_i = 3i + 2$.

The probability density function (PDF) of $\mathbf{x}$ given H_i is denoted by $p(\mathbf{x}; \boldsymbol{\theta}_i \mid H_i)$. The PDF can not be evaluated due to the lack of the knowledge of $\boldsymbol{\theta}_i$.

A typical multiple hypothesis test chooses the hypothesis for which the conditional density $p(\mathbf{x} \mid H_i)$ is maximum, which is known as the maximum likelihood (ML) rule. However, it generally is impossible to obtain the conditional density exactly

$$
p(\mathbf{x} \mid H_i) = \int p(\mathbf{x} \mid \boldsymbol{\theta}_i, H_i) \, p(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i , \tag{A.4}
$$

because the integration is usually difficult to perform and requires the knowledge of $\boldsymbol{\theta}_i$. Instead, an approximation to the original ML rule is widely used [Kay, 1998]:

$$\begin{aligned} &\hat{H}_k = \arg\max_{H_i} \xi_i , \\ &\text{where } \xi_i = \ln p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i \mid H_i) - \tfrac{1}{2} \ln \det\left(\mathbf{I}(\hat{\boldsymbol{\theta}}_i)\right). \end{aligned} \tag{A.5}$$

$\hat{\boldsymbol{\theta}}_i$ is the ML estimation of $\boldsymbol{\theta}_i$ assuming H_i is true. $\mathbf{I}(\cdot)$ is the Fisher information matrix [Papoulis and Pillai, 2002]. The above criterion is often referred to as the *generalized ML rule*.

Because the evaluation of the Fisher information is rather involved, the test statistic ξ_i of the generalized ML rule is approximated by the minimum description length (MDL) approach [Rissanen, 1978]:

$$\xi_i \approx \ln p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i \mid H_i) - \tfrac{1}{2} l_i \ln N. \tag{A.6}$$

The first term of ξ_i is an estimated likelihood of the data given H_i and the second term is the penalty factor for having to estimate unknown parameters. The penalty factor increases linearly with the number of unknown parameters and logarithmically with the data length. When more parameters are to be estimated from a given data, the *credibility* of the likelihood decreases, the penalty factor punishes this tendency.

The q points per segment to be ignored are treated as unknown *deterministic* quantities, and thus are eliminated in computing the likelihood. In turn, l_i is increased to accommodate the number of these points

$$l_i = 3i + 2 + (i+1)q. \tag{A.7}$$

Due to the assumption of additive WGN (AWGN), $p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i \mid H_i)$ is easily expressed as a Gaussian PDF:

$$p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i \mid H_i) = \frac{1}{(2\pi)^{\frac{N-(i+1)q}{2}} \left(\det(\boldsymbol{\Lambda}_i)\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \boldsymbol{\Lambda}_i^{-1} (\mathbf{x} - \mathbf{m}_i) \right), \tag{A.8}$$

where $m_i = \left[\hat{A}_0[\underbrace{1\,1 \ldots 1}_{n_0 - q}] \quad \hat{A}_1[\underbrace{1\,1 \ldots 1}_{n_1 - n_0 - q}] \quad \ldots \quad \hat{A}_i[\underbrace{1\,1 \ldots 1}_{N - n_{i-1} - q}] \right]^T$ and

$\mathbf{\Lambda}_i = \mathrm{diag}\left(\left[\hat{\lambda}_0 [\underbrace{1\,1\,\ldots\,1}_{n_0-q}] \quad \hat{\lambda}_1 [\underbrace{1\,1\,\ldots\,1}_{n_1-n_0-q}] \quad \ldots \quad \hat{\lambda}_i [\underbrace{1\,1\,\ldots\,1}_{N-n_{i-1}-q}] \right]^T \right)$. diag(**a**) is a diagonal matrix with the diagonal of **a** and zero off-diagonal terms. $\hat{A}_j$ and $\hat{\lambda}_j$ are the ML estimations of the mean and variance of the segment *j* [Kay, 1993], i.e.

$$\hat{A}_j = \frac{1}{n_j - n_{j-1} - q} \sum_{n=n_{j-1}}^{n_j-1-q} x[n] \text{ and } \hat{\lambda}_j = \frac{1}{n_j - n_{j-1} - q} \sum_{n=n_{j-1}}^{n_j-1-q} \left(x[n] - \hat{A}_j \right)^2. \tag{A.9}$$

The test statistic (A.6) becomes

$$\begin{aligned} \xi_i = & -\frac{N-(i+1)q}{2} \ln 2\pi \; - \; \frac{1}{2}\left((n_0 - q) \ln \hat{\lambda}_0 + (n_1 - n_0 - q) \ln \hat{\lambda}_1 + \ldots + (N - n_{i-1} - q) \ln \hat{\lambda}_i \right) \\ & - \frac{1}{2}\left(\frac{1}{\hat{\lambda}_0} \sum_{n=0}^{n_0-1-q} (x[n] - \hat{A}_0)^2 + \frac{1}{\hat{\lambda}_1} \sum_{n=n_0}^{n_1-1-q} (x[n] - \hat{A}_1)^2 + \ldots + \frac{1}{\hat{\lambda}_i} \sum_{n=n_{i-1}}^{N-1-q} (x[n] - \hat{A}_i)^2 \right) \; - \; \frac{1}{2} l_i \ln N. \end{aligned} \tag{A.10}$$

Maximizing ξ_i is equal to minimizing $\xi_i' = -2\xi_i$ and now (A.5) becomes, with the new test statistic ξ_i',

$$\hat{H}_k = \arg\min_{H_i} \xi_i', \tag{A.11}$$

where

$$\begin{aligned} \xi_i' = & \left(N - (i+1)q\right) \ln 2\pi + \left((n_0 - q) \ln \hat{\lambda}_0 + (n_1 - n_0 - q) \ln \hat{\lambda}_1 + \ldots + (N - n_{i-1} - q) \ln \hat{\lambda}_i \right) \\ & + \left(\frac{1}{\hat{\lambda}_0} \sum_{n=0}^{n_0-1-q} (x[n] - \hat{A}_0)^2 + \frac{1}{\hat{\lambda}_1} \sum_{n=n_0}^{n_1-1-q} (x[n] - \hat{A}_1)^2 + \ldots + \frac{1}{\hat{\lambda}_i} \sum_{n=n_{i-1}}^{N-1-q} (x[n] - \hat{A}_i)^2 \right) + l_i \ln N. \end{aligned} \tag{A.12}$$

Because $\mathbf{n}_i$ is unknown, the test statistic can not be readily evaluated. Because there exists no explicit formula to determine ML indices, each n_j needs to be found by enumerating all possible change indices. It generally requires $O(N^i)$ evaluations of ξ_i' under H_i. Due to the large computational load, the enumeration scheme is not feasible except for a small *i*.

One approach to reduce the computational load in finding change times is to use the technique of *dynamic programming* (DP) [Kay, 1998]. DP in general can reduce the computational complexity down to $O(iN)$, i.e. an exponential complexity is changed to a polynomial complexity.

We note that the first and last terms of the test statistic do not depend on change times. Thus the ML change times can be found by minimizing the following cost function:

$$J(\hat{\mathbf{A}}_{i,},\hat{\boldsymbol{\lambda}}_i,\mathbf{n}_i)=\Big((n_0-q)\ln\hat{\lambda}_0+(n_1-n_0-q)\ln\hat{\lambda}_1+\ldots+(N-n_{i-1}-q)\ln\hat{\lambda}_i\Big)$$
$$+\left(\frac{1}{\hat{\lambda}_0}\sum_{n=0}^{n_o-1-q}(x[n]-\hat{A}_0)^2+\frac{1}{\hat{\lambda}_1}\sum_{n=n_0}^{n_1-1-q}(x[n]-\hat{A}_1)^2+\ldots+\frac{1}{\hat{\lambda}_i}\sum_{n=n_{i-1}}^{N-1-q}(x[n]-\hat{A}_i)^2\right), \tag{A.13}$$

i.e. $\hat{\mathbf{n}}_i=\arg\min_{\mathbf{n}_i}J(\hat{\mathbf{A}}_{i,},\hat{\boldsymbol{\lambda}}_i,\mathbf{n}_i)$.

The following additive cost function is defined:

$$\Delta_j[n_{j-1},n_j-1-q]\;=\;(n_j-n_{j-1}-q)\ln\hat{\lambda}_j+\frac{1}{\hat{\lambda}_j}\sum_{n=n_{j-1}}^{n_j-1-q}(x[n]-\hat{A}_j)^2. \tag{A.14}$$

The overall cost function is simply the sum of additive costs

$$J(\hat{\mathbf{A}}_{i,},\hat{\boldsymbol{\lambda}}_i,\mathbf{n}_i)=\sum_{j=0}^{i}\Delta_j[n_{j-1},n_j-1-q]. \tag{A.15}$$

The following intermediate optimal cost up to the segment k and index L is defined:

$$I_k[L]=\min_{\substack{n_0,n_1,\ldots,n_{k-1}\\ n_{-1}=0,\,n_k=L+1+q}}\sum_{j=0}^{k}\Delta_j[n_{j-1},n_j-1-q], \tag{A.16}$$

where $n_{k\text{-}1}+(p-1)\le L$. Note that the minimum of the overall cost function is $I_i[N-1-q]$. It can be shown that the intermediate optimal cost function can be written in a recursive form [Bertsekas, 1987]

$$I_k[L]=\min_{n_{k-1}}\big(I_{k-1}[n_{k-1}-1-q]+\Delta_k[n_{k-1},L]\big). \tag{A.17}$$

The equation (A.17) says that the intermediate optimal cost up to the segment k and index L is the minimum of the sum of the intermediate cost up to the segment $k-1$ and index $n_{k\text{-}1}-1-q$ and the additive cost between the indices $n_{k\text{-}1}$ and L. $\hat{n}_{k-1}$, the starting index of the segment k at which $I_k[L]$ becomes a minimum, is found during the course of the minimum search.

With the aid of the dynamic programming, ξ_i', the test statistic for H_i, is computed via the following procedure:

1. Let $k=0$ and $I_{-1}[\cdot]=0$. For $L=(p-1),(p-1)+1,\ldots,(N-1)-i(p+q)$, compute

$$I_0[L]=\Delta_0[n_{-1}=0,L]=(L+1)\ln\hat{\lambda}_0+\frac{1}{\hat{\lambda}_0}\sum_{n=0}^{L}(x[n]-\hat{A}_0)^2,$$

where $\hat{A}_0 = \frac{1}{L+1}\sum_{n=0}^{L} x[n]$ and $\hat{\lambda}_0 = \frac{1}{L+1}\sum_{n=0}^{L}(x[n]-\hat{A}_0)^2$.

2. Let $k = 1$. For $L = (p-1) + k(p+q), \ldots, (N-1) - i(p+q) + k(p+q)$, compute

$$I_k[L] = \min_{(p-1)+(k-1)(p+q)+(q+1) \le n_{k-1} \le L-(p-1)} \left(I_{k-1}[n_{k-1}-1-q] + \Delta_k[n_{k-1}, L]\right).$$

Also, for each L, determine $\hat{n}_{k-1}$ at which $I_k[L]$ becomes minimum and call it $n_{k-1}[L]$.

3. Let $k \leftarrow k+1$ and iterate until $k = i$.

4. Let $k = i$. For $L = (N-1) - q$, compute

$$I_k[L] = \min_{(p-1)+(k-1)(p+q)+(q+1) \le n_{k-1} \le L-(p-1)} \left(I_{k-1}[n_{k-1}-1-q] + \Delta_k[n_{k-1}, L]\right).$$

Also, find $\hat{n}_{i-1} = n_{i-1}[N-1-q]$, i.e. the start index of the last segment.

Now $I_i[N-1-q]$ is the desired minimum of the overall cost function and ξ_i' can be computed by adding two remaining terms, i.e.

$$\xi_i' = I_i[N-1-q] + \left(N-(i+1)q\right)\ln 2\pi + l_i \ln N. \tag{A.18}$$

The change times are found by the following backward recursion:

$$\begin{aligned} \hat{n}_{i-1} &= n_{i-1}[N-1-q] \\ \hat{n}_{i-2} &= n_{i-2}[\hat{n}_{i-1}-1-q] \\ &\vdots \\ \hat{n}_0 &= n_0[\hat{n}_1-1-q]. \end{aligned} \tag{A.19}$$

Once the change times are found, the ML estimations of the mean and variance of each segment are easily obtained by using (A.9).

The test statistic ξ_i' is computed for each i, $i \in \{0, 1, \ldots, k_{max}\}$, and an H_i which gives a minimum test statistic is selected from the composite multiple hypothesis test. Note H_i indicates that there exist i mean/variance changes for the given data of length N and $(i+1)$ constant mean/variance segments. The start index, mean estimation and variance estimation of each segment are obtained as byproducts of the hypothesis test.

The computational time to complete the hypothesis test increases substantially with k_{max} ($\sim O\left(N^2 \sum_{k=0}^{k_{\max}} k\right)$). Thus a reasonable number should be chosen for k_{max}, considering the length of the data record and the relative frequency and nature of load activities.

As an example, consider the test signal shown in Figure A-1 with $N = 100$. The original signal $s[n]$ is composed of three segments with start indices at 0, 20 and 40. There exist two transition points between the segments. $s[n]$ is corrupted with an additive zero-mean WGN whose variance is 1 and changes to 16 starting at $n = 80$. Thus the resulting signal $x[n] = s[n] + w[n]$ has four segments, the last of which is a pure variance change. For the test, following parameters are used: $k_{max} = 4$, $p = 5$ and $q = 2$.

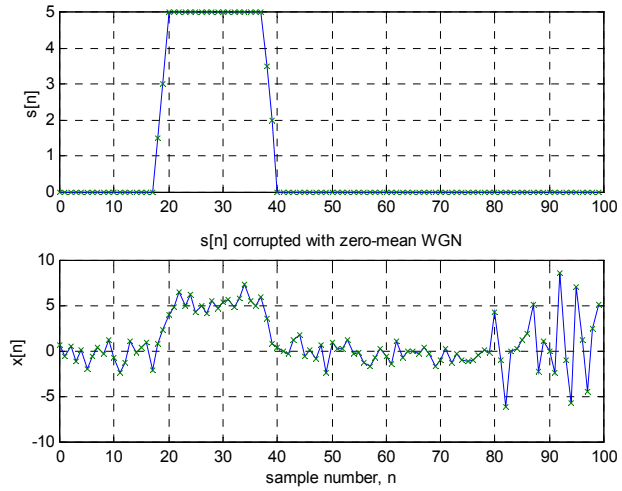


Figure A-1 Test signal for change-of-mean detector (Top: original signal, Bottom: corrupted by zero-mean WGN).

Figure A-2 shows the test statistic computed for each hypothesis. The test statistic clearly becomes minimum at $i = 3$, and it successfully detected three mean/variance changes present in the test signal.

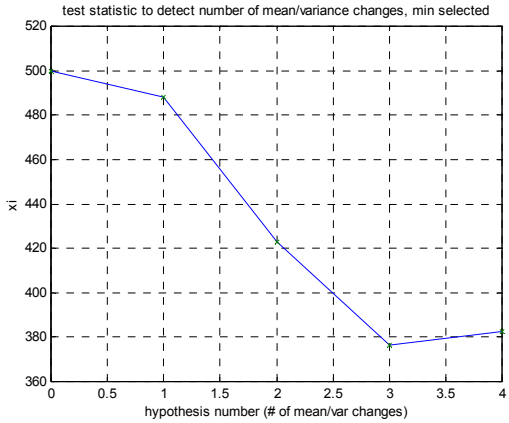


Figure A-2 Test statistic for each hypothesis with the signal of Figure A-1.

The following is the screen dump of the change-of-mean detector output:

```
n estimation:
    21    40    82   100
A estimation:
 -2.4928e-001  5.4169e+000 -2.4563e-001  3.0583e-001
var estimation:
  1.1539e+000  6.0716e-001  7.6817e-001  1.6012e+001
```

The change time estimates are reasonably accurate considering two points are dismissed between the segments. The resulting mean and variance estimations are also acceptable.

When the change-of-mean/variance detector is used in the context of the steady-state load classification, it has to be applied to the continuous stream of a selected power. A window of size N is applied to the stream to identify change-of-mean/variance events, then the window is shifted by L and the process continues. Considering the fact that the boundary of a window can be placed arbitrarily with respect to a turn-on/off response, it is necessary to allow a small amount of overlap and adjust the overlap size, if necessary, while keeping L fixed.

The general change-of-mean/variance detector developed in this chapter can detect an arbitrary number of mean/variance changes for given data. Except for test parameters, such as the maximum number of changes, the minimum length of a segment and the number of transition points to be ignored, it does not require any prior information about the data. It detects the number of changes, and estimates the change index, mean and variance of each segment. It converts the stream of a power into the series of constant mean/variance segments.

The computational load was reduced substantially by the introduction of dynamic programming. However, computations can still be a problem when a large k_{max} is tested. If k_{max} is set to one, the detector functions similar to the conventional GLR detector, and also with a comparable computational cost.

Appendix B

Air Handling Unit Modeling

An air handling unit (AHU) is designed to provide thermal comfort and ventilation to the occupant in a building. The AHU consists of fans to induce airflow, heating and cooling coils to deliver thermal energy, dampers to regulate air flow, thermostats to control temperature, ducts to provide air passage and building space. There exist many varieties of AHU systems. However, they can be classified mainly either as a constant volume (CV) system or a variable-air-volume (VAV) system. A CV system typically has a single supply fan, whereas a VAV system has both a supply and a return fan to meet the varying air flow load [Graf and Lee, 1998]. Because most commercial buildings have VAV systems, we will concentrate on their modeling and analysis.

B.1 VAV model

Figure B-1 shows the schematic of the VAV system. The combination of heating and cooling coils transfers energy to the airflow required to meet the thermal load of a building. The supply fan delivers the conditioned air to each VAV terminal box through the supply duct. Each VAV box regulates the airflow, by varying its damper position, to maintain a constant set temperature of the space or zone being served. The return fan returns air from the space to the mixing zone.

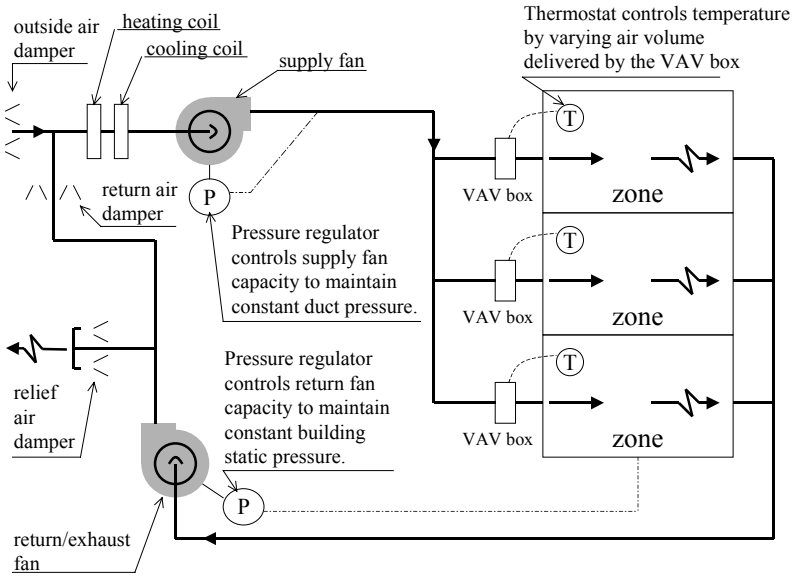
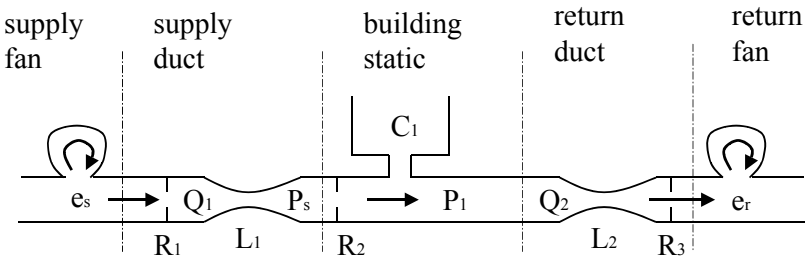


Figure B-1 Schematic of variable-air-volume system.

Depending on the position of the three dampers– the outside air, return air and relief air dampers– 0 to 100 % of the return air can be recirculated to the supply fan. For example, if the outside air and relief air dampers close fully and the return air damper opens fully,

the system recirculates 100 % the return air. If the dampers are in the opposite positions, the return air is fully vented to the environment, and 100 % outside air is supplied to the mixing zone. The fan capacity is controlled to maintain a constant pressure at a designated position, either by varying speed or varying inlet damper position. For example, both fans can be controlled to maintain a constant supply duct pressure, or the supply fan maintains a supply duct pressure and the return fan maintains a building static pressure.

We present a very simple model of the VAV AHU system. Figure B-2 shows the schematic of the model.



e_s: pressure created by supply fan	L_1: inertance (inductance) of supply duct
e_r: pressure created by return fan	L_2: inertance (inductance) of return duct
P_s: supply duct pressure	C_1: capacitance of building
P_1: building static pressure	R_1: resistance of supply duct
Q_1: flow rate along supply duct	R_2: resistance of building space
Q_2: flow rate along return duct	R_3: resistance of return duct

Figure B-2 Simple model of AHU system.

A set of assumptions is made in developing the model:

1. The main characteristic of the air duct is described by the inductance-resistance (*L*-*R*) model, while the building space is represented by the resistance-capacitance (*R*-*C*) model. Because a duct has a very long, yet confined air passage, it is difficult to develop flow rate (inertial effect) whereas it is relatively easy to build up pressure. It is observed in buildings that sudden rises in pressure often damage duct systems, when fans are accelerated too rapidly. A gradual ramp up of a variable speed fan, during start up, is required to prevent an abrupt pressure rise and hence duct damage. On the other hand, the building space has a much larger volume. Thus it is relatively easy to establish flow rate, but it is difficult to increase the space pressure (capacitance effect). This capacitance effect is the result of the inherent air compressibility, though minor. Of course, the duct has a certain amount of capacitance and the building space has an inertance. But they are not the dominant characteristic of each system and ignored in this model. On the other hand, resistance is present in every system.

2. The presence of dampers (relief air, outside air and return air) is ignored for simplicity. In this configuration, the supply and return fans are connected directly to the environment. They are open to atmospheric pressure. This is the same as the 100 % outside air supply case (return damper closes; relief and outside damper open). To add the dampers to the model, the supply and return fans should be connected with a variable resistance and with two tees, with their own varying resistances, to the atmosphere. The influence of the dampers will be considered in future models. Because the reference pressure is atmospheric, all the pressures in this model are gauge pressures (negative for vacuum).

3. The time-varying nature of the system is ignored during the development. The system has an inherent time-varying nature. For example, the resistance of the building static varies depending on loads. If there are more thermal, or ventilation, loads, VAV boxes will move dampers towards the fully open positions, hence reducing the resistance to the air flow. Or, a branch of the supply duct may be closed if the corresponding space of a building is not in use, increasing the duct resistance. The inertance and the capacitance of the system are believed to remain relatively unchanged. We describe the system by first assuming a time-invariant nature. Later, we will discuss the effect of the time variance.

4. The duct and building space network are simplified. The duct system is not a single air stream passage. It typically divides into many branches, each of which serves a different portion of a building. The return duct joins many branches and forms a trunk before it reaches the return fan. The same is true in the building space. It is divided into many floors and zones, providing multiple passages for the air flow. However, the model describes the network in an *average* sense.

5. Dynamics of the fans and motors are ignored. The duct and space air passage system has more inertance and hence a slower time constant than the fans and motors. The relatively fast dynamics of these devices are ignored for simplicity. Fans are treated as effort, or pressure, sources that instantly create pressure rises. A more sophisticated model may include the dynamics of fans and possibly of motors.

6. Air infiltration is ignored. In the model, the air delivered by the supply fan, minus any amount stored by the building capacitance, is received by the return fan without any loss. In real buildings, depending on the pressure difference, there exists an air leak to or infiltration from the environment. The effect of infiltration can be modeled by adding a tee with a resistance to the building static.

The bond graph of the described VAV AHU model is shown in Figure B-3. Bond graphs are often used to describe and model complex dynamic systems. They have the advantage of showing the inherent nature of a system, such as power flow, causality, state equation, etc., in a simple compact graphical format [Karnopp *et al*., 1990].

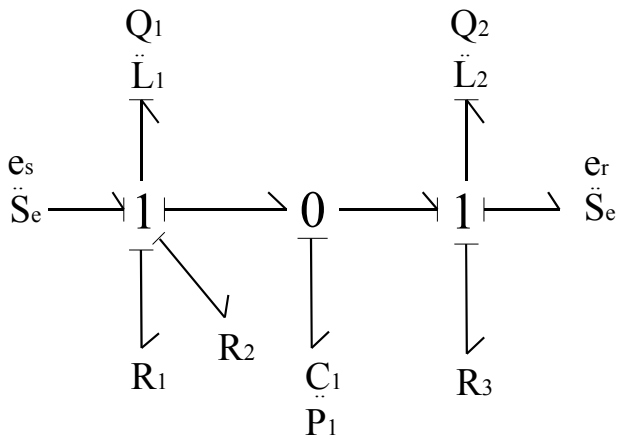


Figure B-3 Bond graph of AHU system model.

A state space equation describing the system can be obtained from the bond graph. In bond graph terms, each energy storage element, L or C, with an integral causality (which simply means that it maintains its own unique characteristic even interacting with other elements) contributes a state variable. Because we have three energy storage elements with integral causality in the system, we have three state variables. Each variable, namely P_1, Q_1 and Q_2, is related to its corresponding energy storage element. The system is described by the following state space equation:

$$\begin{bmatrix} \dot{P}_1 \\ \dot{Q}_1 \\ \dot{Q}_2 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{C_1} & -\frac{1}{C_2} \\ -\frac{1}{L_1} & -\frac{R_1+R_2}{L_1} & 0 \\ \frac{1}{L_2} & 0 & -\frac{R_3}{L_2} \end{bmatrix} \begin{bmatrix} P_1 \\ Q_1 \\ Q_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \frac{1}{L_1} & 0 \\ 0 & \frac{1}{L_2} \end{bmatrix} \begin{bmatrix} e_s \\ -e_r \end{bmatrix}. \tag{B.1}$$

e_r is the pressure created by the return fan (thus negative). $-e_r$ gives a positive vacuum pressure. However the above equation is not convenient in the sense that the flow rate is not readily measurable. Thus, we introduce the supply duct pressure P_s in place of Q_1 using the relationship: $P_s = R_2Q_1 + P_1$. The new state space equation is given by:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}, \tag{B.2}$$

where $\mathbf{C} = \mathbf{I}$ ($\mathbf{I}$ is the identity matrix) and $\mathbf{y}$ is the output of the system. The state vector, input vector and coefficient matrices are given by the following formulae:

$$\mathbf{x} = \begin{bmatrix} P_s & P_1 & Q_2 \end{bmatrix}^T, \quad \mathbf{u} = \begin{bmatrix} e_s & -e_r \end{bmatrix}^T,$$

$$\mathbf{A} = \begin{bmatrix} \left(\frac{1}{C_1 R_2} - \frac{R_1 + R_2}{L_1}\right) & \left(\frac{R_1}{L_1} - \frac{1}{C_1 R_2}\right) & -\frac{1}{C_1} \\ \frac{1}{C_1 R_2} & -\frac{1}{C_1 R_2} & -\frac{1}{C_1} \\ 0 & \frac{1}{L_2} & -\frac{R_3}{L_2} \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} \frac{R_2}{L_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{L_2} \end{bmatrix}. \tag{B.3}$$

The characteristics of the system can be examined by obtaining the open-loop response of the system. Define the open-loop transfer function $\mathbf{G}(s) = \mathbf{Y}(s)/\mathbf{U}(s)$, where $\mathbf{Y}(s)$ and $\mathbf{U}(s)$ are the Laplace transforms of **y** and **u**, respectively. $\mathbf{G}(s)$ is obtained by the matrix equation, $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, where s is the Laplace operator. After working out all the algebraic details, $\mathbf{G}(s)$ is given by:

$$\begin{aligned} \mathbf{G}(s) &= \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \\ G_{31}(s) & G_{32}(s) \end{bmatrix} = \begin{bmatrix} \frac{P_s(s)}{e_s(s)} & \frac{P_s(s)}{-e_r(s)} \\ \frac{P_1(s)}{e_s(s)} & \frac{P_1(s)}{-e_r(s)} \\ \frac{Q_2(s)}{e_s(s)} & \frac{Q_2(s)}{-e_r(s)} \end{bmatrix} \\ &= \frac{1}{\det(s\mathbf{I} - \mathbf{A})} \begin{bmatrix} \frac{R_2}{L_1}\left(s + \frac{1}{C_1 R_2}\right)\left(s + \frac{R_3}{L_2}\right) + \frac{R_2}{L_1}\frac{1}{L_2 C_1} & -\frac{1}{L_2}\left(\frac{1}{C_1}s + \frac{R_1}{C_1 L_1}\right) \\ \frac{R_2}{L_1}\frac{1}{C_1 R_2}\left(s + \frac{R_3}{L_2}\right) & -\frac{1}{L_2 C_1}\left(s + \frac{R_1 + R_2}{L_1}\right) \\ \frac{R_2}{L_1}\frac{1}{C_1 R_2 L_2} & \frac{1}{L_2}\left(s^2 + \frac{R_1 + R_2}{L_1}s + \frac{1}{C_1 L_1}\right) \end{bmatrix}. \end{aligned} \tag{B.4}$$

The determinant is given by the following formula:

$$\det(s\mathbf{I} - \mathbf{A}) = s^3 + \left(\frac{R_3}{L_2} + \frac{R_1 + R_2}{L_1}\right)s^2 + \left(\frac{(R_1 + R_2)R_3}{L_1 L_2} + \frac{1}{C_1 L_1} + \frac{1}{C_1 L_2}\right)s + \frac{R_1 + R_2 + R_3}{C_1 L_1 L_2}. \tag{B.5}$$

The closed-loop response depends on what type of control logic is used. For simplicity, we assume a simple proportional control with an offset. Specifically, we control the supply-fan pressure to maintain a constant duct pressure and the return-fan pressure to maintain a constant building static pressure.

$$\mathbf{u} = \mathbf{K}(\mathbf{r} - \mathbf{x}) + \mathbf{m} = \begin{bmatrix} k_1 & 0 & 0 \\ 0 & -k_2 & 0 \end{bmatrix} \begin{bmatrix} P_{sr} - P_s \\ P_{1r} - P_1 \\ Q_{2r} - Q_2 \end{bmatrix} + \begin{bmatrix} m_s \\ m_r \\ 0 \end{bmatrix}, \tag{B.6}$$

where **r** is the reference, or control set point, and **m** is the offset vector. **K** is the gain matrix. In this case we do not feed back the flow rate, which is difficult to measure, to control the fan pressures. Figure B-4 shows the block diagram of the control system.

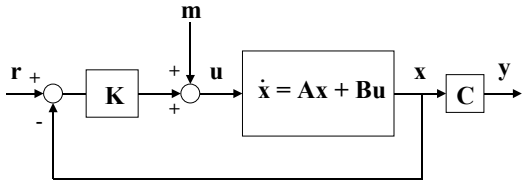


Figure B-4 Block diagram of AHU control system.

The closed-loop system response is given by:

$$\mathbf{Y}(s) = \mathbf{G}_\mathbf{c}(s)\mathbf{R}(s) + \mathbf{H}(s)\mathbf{M}(s), \tag{B.7}$$

where $\mathbf{G}_\mathbf{c}(s) = \dfrac{\mathbf{Y}(s)}{\mathbf{R}(s)} = \mathbf{C}[s\mathbf{I} - (\mathbf{A} - \mathbf{BK})]^{-1}\mathbf{BK}$ and $\mathbf{H}(s) = \dfrac{\mathbf{Y}(s)}{\mathbf{M}(s)} = \mathbf{C}[s\mathbf{I} - (\mathbf{A} - \mathbf{BK})]^{-1}\mathbf{B}$.
$\mathbf{Y}(s)$, $\mathbf{R}(s)$ and $\mathbf{M}(s)$ are the Laplace transforms of **y**, **r** and **m**, respectively.

With this type of control logic, the open-loop transfer functions $G_{11}(s)$ and $G_{22}(s)$ play important roles in determining the closed-loop response of the system. Each transfer function describes the behavior of the state variable as a function of the corresponding control input, fan pressure, which in turn is used to drive the variable to the desired reference point.

The 160 Sansome building AHU is controlled in almost the same way as we described the closed loop control system, i.e. the supply-fan speed is controlled to maintain a constant supply duct static pressure and the return-fan speed is controlled to a constant building static pressure. The following is the control logic of the building.

$$\begin{cases} u_s = 10 \times (P_{sr} - P_s) + 19.25 \\ u_r = -24 \times (P_{1r} - P_1) + 18.3 \end{cases}, \tag{B.8}$$

where u_s and u_r are the velocity command inputs to the supply and the return fan VFD, respectively (4 – 20 mA). Pressures are measured in inch H_2O. The presence of large offsets, close to the upper limit, is noteworthy. These offsets serve as the open loop inputs to the system, deriving the pressures, hence the state variables, to certain DC values, which are usually close to the desired set points. The proportional control with a feedback is added to the control input to regulate the pressure to a desired set point. However, the control logic of the building is not the same as the control law of the model (B.6). In the model, **u** is the pressure input vector to the system. Thus appropriate gains, and the transfer functions of the fan system should be multiplied to the control logic to obtain an equation applicable to the model.

B.2 Sustained oscillation- limit cycle

It will be helpful to glimpse the behavior of the variables of the real AHU system, from the 160 Sansome Energy Monitoring and Optimization (EMO) website. Data are displayed in real time using the Electric Eye software. The minimum reliable refreshing, or sampling, interval is 5 second.

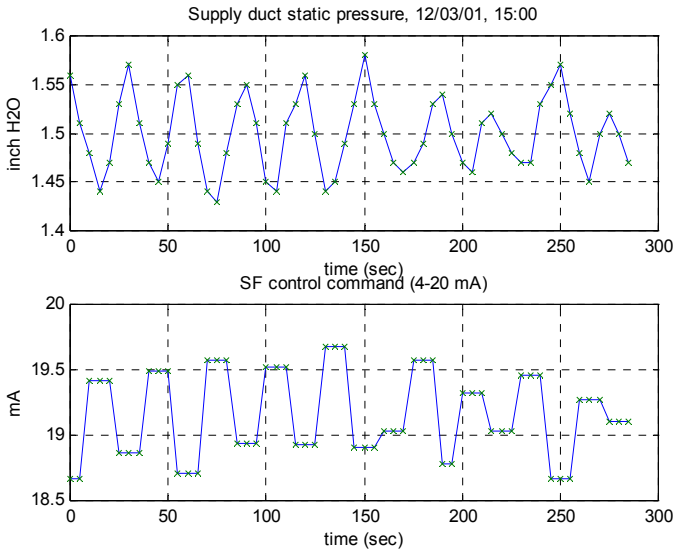


Figure B-5 Supply duct static pressure and SF control command.

Figure B-5 shows an example of the behavior of the state and control variables of the AHU, recorded from the website. From the figure, it is clear that the supply duct pressure has a steady oscillation with a period of 30 sec, the same fan oscillation period observable from the *P* stream. Also we can infer that the supply duct pressure set point is 1.5 inch H_2O, because it is the average value of the measurement. The SF control signal is 180° out of phase with the duct pressure (the pressure is multiplied by –1 in the control logic).

The presence of a sustained oscillation is an interesting problem. In linear system theory, a feedback system can show an oscillation at steady state if it has poles located on the imaginary axis in the complex plane. However, if there is a slight disturbance, the pole

can move to the right half complex plane, meaning that the system becomes unstable (oscillation keeps growing in amplitude). Thus, there is no safety net to allow a sustained oscillation in a linear system. On the other hand, it is unreasonable to believe that the control engineer places the system poles exactly on the imaginary axis after many tuning attempts, even accidentally. Thus, linear system theory alone can not explain the sustained oscillation behavior of the feedback-controlled system such as the AHU system.

Something is nonlinear. Otherwise, the behavior can not be explained. There can be many sources of nonlinearity. A duct may have an inertance that is a function of flow rate and pressure, rather than a constant. There possibly can be a transport delay in the control loop. But, we eye on the actuators– the SF and the RF. They can *saturate*.

AHU fans are typically selected, in terms of capacity, based on their steady state load. They are seldom selected to meet the transient load, or acceleration. In a sense, AHUs are underpowered. It is quite possible that if there is a changing demand, the actuators become saturated because controllers demand more than they can produce, hence entering a nonlinear region. (In modern computer controlled systems, output from a control logic is already limited before it is given to an actuator. In this case, it is the saturation of the control logic, but the result is the same). This is plausible for the 160 Sansome AHU system, because both fans, especially the SF, are operated close to their upper limits.

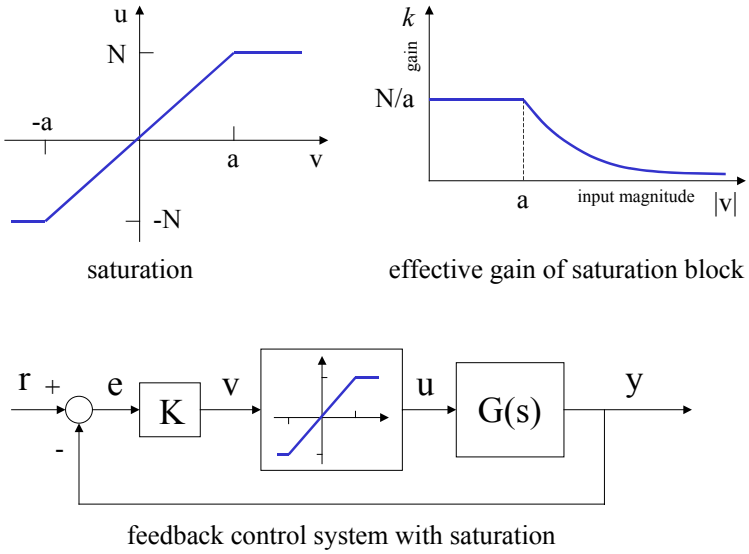
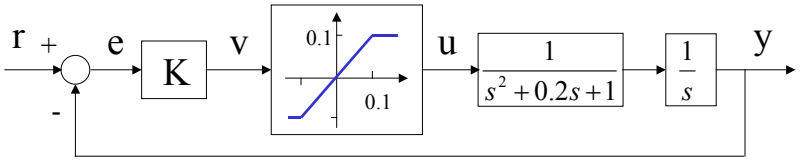


Figure B-6 Effect of saturation in feedback control system.

Figure B-6 shows the effect of saturation on a feedback control system [Franklin *et al.*, 1994]. When input to the saturation block is less than a threshold value, the block gives an output proportional to the input (gain k = N/a, slope of the saturation curve). Note the overall gain of the system is Kk. However, if the input magnitude exceeds the threshold value, the output is no longer proportional to the input. The output is the constant limit

value and the effective gain is reduced to N/|v|. As the magnitude of input grows, the effective gain diminishes. The behavior of the effective gain with varying input magnitude is depicted in the graph. The feedback control system has a varying gain that depends on the magnitude of the error signal (e). When the error signal is small, the system has a constant gain. But as the magnitude of the error signal grows the effective gain of the system is reduced. (For example, the system shows a high gain at its steady state, but a low gain during acceleration or deceleration.)

Figure B-7 shows an example of system with oscillation due to saturation. The root locus of the linear portion of the system is also shown in the picture. If we set the gain K = 0.5, the root locus shows that the system has two unstable closed-loop poles in the right half plane (RHP). Thus the system should be unstable and the output y will diverge to infinity. However, the system shows an oscillatory behavior in the limit. Figure B-8 shows the result of the unit step response simulation of the system including the effect of saturation. After the initial ramp up, it starts to show an oscillatory behavior. The oscillation builds up to a fixed amplitude and stops growing. The frequency is roughly 1 rad/sec (cross-over points at the root locus are $\pm 0.991j$).



Feedback system with an oscillatory mode due to saturation

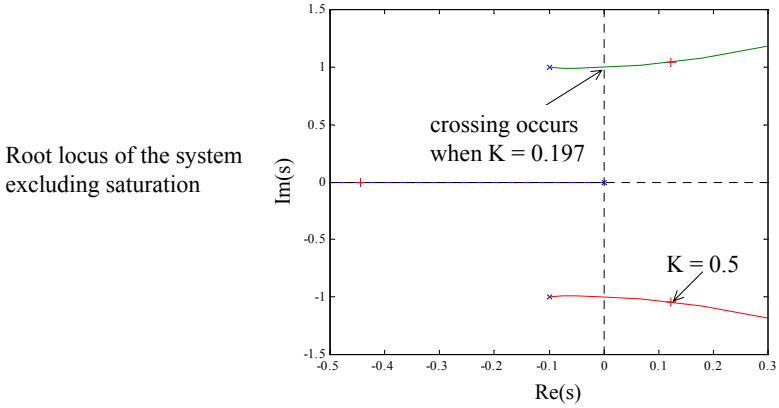


Figure B-7 System with sustained oscillation due to saturation.

Although the system has a gain of 0.5, at the beginning of the step command the controller is saturated due to the large amount of error. Thus, initially the system has a very low effective gain and has poles on the left half plane (LHP). As system builds up output, the error reduces and the degree of saturation diminishes, increasing the effective gain of the system. The system poles start to move towards the RHP. As they near the crossing points on the imaginary axis, the system starts to show oscillatory behavior (or they can actually jump into the RHP). The oscillation builds up until the poles reach the imaginary axis. However, once they move into the RHP, the output amplitude starts to increase due to instability. But on the other hand, the increased output reduces the system

gain due to the saturation of controller. The saturation moves the poles back toward the LHP. Eventually, they will settle exactly on the imaginary axis, resulting in a stable, sustained, and fixed frequency and amplitude oscillation, called a *limit cycle* [Gibson, 1963].

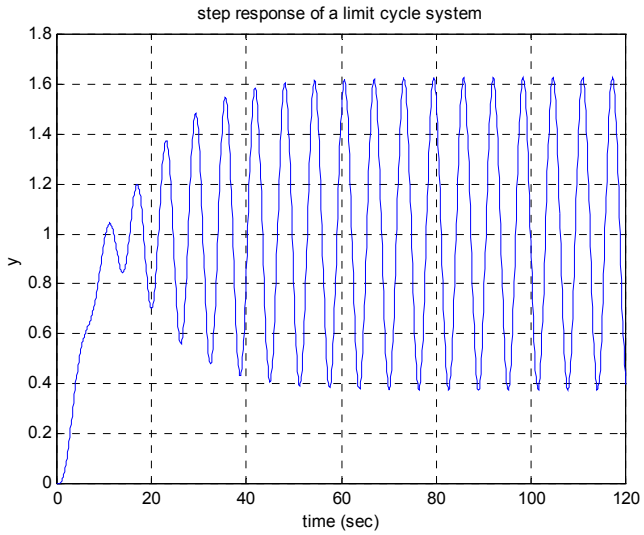


Figure B-8 Unit step response of system showing limit cycle.

The limit cycle is a stable equilibrium of the system response. A disturbance may cause the system poles to move away from the imaginary axis, but eventually they come back to the equilibrium position. In the example, two poles on the imaginary axis contribute to the sustained oscillation, while one pole on the real axis is responsible for the DC value of the step response.

However, the argument in the previous paragraph is based on the heuristic explanation of the system behavior. We need a simple analytic tool to describe the nonlinear behavior of a system. Consider the following single-input single-output system:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}u \\ y &= \mathbf{C}\mathbf{x} \end{aligned} \tag{B.9}$$

with an output nonlinearity $u = \psi(y)$. The linear part of the system is controllable and observable. The linear system Laplace transform function

$$G(s) \cong \left.\frac{Y(s)}{U(s)}\right|_{zeroICs} = \mathbf{C}[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B} = \frac{Q(s)}{P(s)} \tag{B.10}$$

is strictly proper, i.e. $P(s)$ is a polynomial of higher order than the polynomial $Q(s)$. If $y(t)$ is periodic with period T, the following mth order *harmonic balance* condition is satisfied [Vincent and Grantham, 1997]:

$$a_k - G(jk\omega)c_k = 0, \qquad k = 0,1,\ldots,m, \tag{B.11}$$

where a_k is a coefficient of the Fourier series of y and c_k is a Fourier series coefficient of u. They are formally defined as,

$$\begin{aligned} y(t) &= \sum_{k=-\infty}^{\infty} a_k e^{jk\omega t}, \qquad a_k = \frac{1}{T}\int_0^T y(\tau)e^{-jk\omega\tau}d\tau \\ u(t) &= \Psi(y(t)) = \sum_{k=-\infty}^{\infty} c_k e^{jk\omega t}, \qquad c_k = \frac{1}{T}\int_0^T \Psi(y(\tau))e^{-jk\omega\tau}d\tau\,. \end{aligned} \tag{B.12}$$

The harmonic balance equation approximates the output and the nonlinearity with an mth order Fourier series. The simplest approximation is to take m = 1, which yields the describing function method [Truxall, 1955]. The method is widely used to describe the nonlinearities present in otherwise linear system.

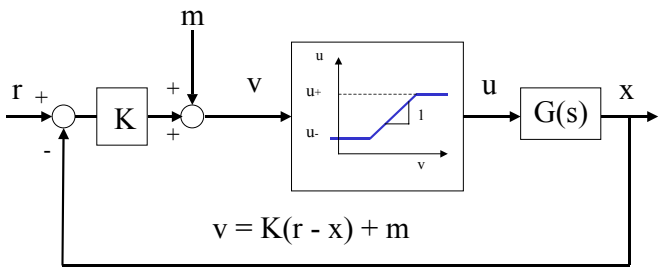


Figure B-9 Feedback control system with saturation nonlinearity

Let's consider the system shown in Figure B-9. The system has an upper saturation limit u_+ and a lower saturation limit u_- for an input signal u. Both of the limits are positive. We assume the existence of a sinusoidal plus DC level signal for x.

$$x = a_0 + a\sin\omega t = \sum a_k e^{jk\omega t} = a_0 + \frac{a}{2j}(e^{j\omega t} - e^{-j\omega t}). \tag{B.13}$$

Thus the output x is described by the first order Fourier series and $a_1 = a/2j$. Then the input signal v to the saturation block is given by

$$v = K(r_0 - a_0) + m - Ka\sin\omega t\,, \tag{B.14}$$

where r_0 is the desired constant set point. The output from the saturation block u depends on the magnitude of v, which in turn depends on K, r and x. There can be many

possibilities for the shape of the u waveform. However we consider two extreme cases for simplicity.

Case 1) $K(r_0 - a_0) + m - Ka \geq u_+$

In this case, the input to the saturation block is always greater than the upper limit of saturation. Thus the output is constant u_+

$$u = \sum c_k e^{jk\omega t} = u_+ \quad \Rightarrow c_0 = u_+, c_k = 0 \text{ for } k = \pm 1, \pm 2, \ldots \quad . \tag{B.15}$$

Case 2) $K(r_0 - a_0) + m - Ka \geq u_-$ and $K(r_0 - a_0) + m + Ka \leq u_+$

The input to the saturation block is in the linear region. Thus the output of the block is the same as the input.

$$\begin{aligned} u &= \sum c_k e^{jk\omega t} = K(r_0 - a_0) + m - Ka\sin\omega t \\ &\Rightarrow c_0 = K(r_0 - a_0) + m, c_1 = \frac{-Ka}{2j}, c_{-1} = \frac{Ka}{2j}, c_k = 0 \text{ for } k = \pm 2, \ldots \quad . \end{aligned} \tag{B.16}$$

In Case 1, u is described by the *zero*th-order Fourier series. In Case 2, a first-order Fourier series is enough to describe u. In an intermediate case, the input signal v is partially saturated and the output has a clipped sinusoidal waveform, which generally requires more coefficients in higher orders to describe with a Fourier series.

In Case 2, the harmonic balance equation with $k = 1$ becomes

$$\frac{a}{2j} - G(j\omega)\cdot\frac{-Ka}{2j} = 0 \quad \Rightarrow \quad a(1 + KG(j\omega)) = 0. \tag{B.17}$$

Because a is nonzero (a is the amplitude of oscillation), the term inside the parenthesis should be zero. But readers will recognize the equation as nothing but the root locus formula (a root locus is a series of points which satisfies $1 + KG(s) = 0$), when the poles are located on the imaginary axis ($s = j\omega$). The assumption of the existence of a sinusoidal oscillation led to the root locus equation on the imaginary axis. In other words, if there is a sustained oscillation within the saturation limits, the system poles are located on the imaginary axis, and the oscillation frequency, in rad/sec, itself is the closed-loop system *pole*.

If K is real, $1 + KG(j\omega) = 0$ can be rewritten as,

$$1 + K\text{Re}[G(j\omega)] + j\text{Im}[G(j\omega)] = 0. \tag{B.18}$$

Thus,

$$1 + K\mathrm{Re}[G(j\omega)] = 0 \text{ and } \mathrm{Im}[G(j\omega)] = 0. \tag{B.19}$$

The second equation can be used to find the frequency of oscillation and the first equation to find the gain K, or an unknown of $G(s)$ if the gain is known.

With $k = 0$, the Case 2 harmonic balance equation gives

$$\begin{aligned} a_0 - G(0)c_0 &= 0 \\ G(0) = \frac{a_0}{c_0} &\approx \frac{a_0}{m}. \end{aligned} \tag{B.20}$$

Thus if the DC value of the signal is near the set point ($r_0 \approx a_0$), $G(0)$ is approximated by the ratio of the signal DC value to the input offset.

We have shown heuristically that a sustained oscillation can develop and exist stably in a feedback system with saturation by means of the varying effective gain. We have also shown that if there is a continued oscillation, the closed-loop poles are located on the imaginary axis and the frequency is the magnitude of the pole.

B.3 Conditions to exhibit limit cycle

So far we have developed the VAV AHU model, and explained the nature of limit cycle, which can be present in a feedback control system with saturation. To explain the oscillatory behavior of the 160 Sansome AHU system, we have to show that there exists a limit cycle for at least one of the transfer functions of $\mathbf{G}(s)$ (remember the matrix $\mathbf{G}(s)$ is composed of six scalar transfer functions). However, this is an extremely difficult task, because we do not know the numerical values of the parameters. We will simply show and explain under which conditions a limit cycle may develop.

We choose $G_{11}(s) = P_s(s)/e_s(s)$, which describes the portion of supply duct pressure generated by the SF pressure input, as a candidate to show a sustained oscillation. Supply fan pressure is the main driver of the AHU system and 'pushes' air to the duct and space network. In comparison, the return fan 'pulls' air from it. A *pull* system is inherently more stable than a *push* system (If you tow a trailer, you know it is much easier to haul than back up). $G_{11}(s)$ is a *push* system. The algebraic form of $G_{11}(s)$ is repeated here again,

$$G_{11}(s) = \frac{P_s(s)}{e_s(s)} = \frac{\dfrac{R_2}{L_1}\left(s + \dfrac{1}{C_1R_2}\right)\left(s + \dfrac{R_3}{L_2}\right) + \dfrac{R_2}{L_1}\dfrac{1}{L_2C_1}}{s^3 + \left(\dfrac{R_3}{L_2} + \dfrac{R_1 + R_2}{L_1}\right)s^2 + \left(\dfrac{(R_1 + R_2)R_3}{L_1L_2} + \dfrac{1}{C_1L_1} + \dfrac{1}{C_1L_2}\right)s + \dfrac{R_1 + R_2 + R_3}{C_1L_1L_2}}. \tag{B.21}$$

When the loop is closed by feeding P_s back to the controller with a gain k, the poles of the closed loop system are located on $1 + kG_{11}(s) = 0$, i.e. on the root locus plot. $G_{11}(s)$

has two zeros and three poles. The locations of poles and zeros depend on the numerical values of the parameters. However, to exhibit a limit cycle, the root locus should cross the imaginary axis. The only possible configuration for a two-zero/three-pole system is when zeros are located close to two symmetric poles but farther away from the real axis. Figure B-10 shows the root locus of such a system, as an example. The system has open loop poles at –0.1 and –0.2 ± *j*0.5 and zeros at –0.2 ± *j*2.

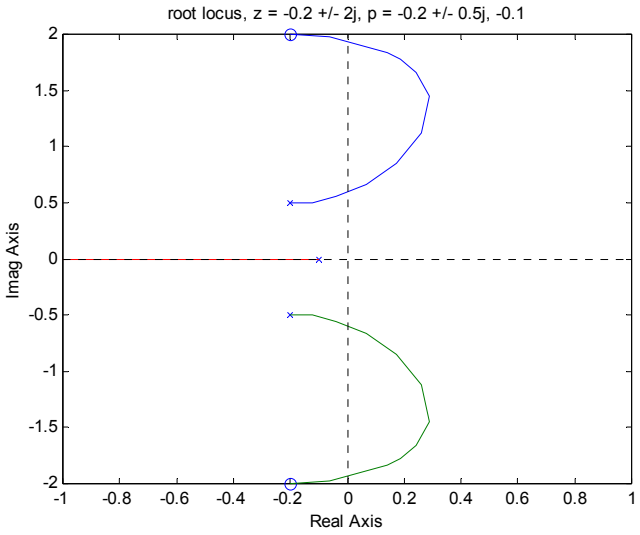


Figure B-10 Root locus of third-order system with crossover points.

Zeros of $G_{11}(s)$ are easily obtained,

$$z = -\frac{1}{2}\left(\frac{1}{C_1 R_2} + \frac{R_3}{L_2}\right) \pm \sqrt{\frac{1}{4}\left(\frac{1}{C_1 R_2} - \frac{R_3}{L_2}\right)^2 - \frac{1}{L_2 C_1}}\,. \tag{B.22}$$

Zeros will have imaginary components if the term inside the square root is negative. The readers will recognize $1/L_2C_1$ as the square of a natural frequency of oscillation[15]. Thus $G_{11}(s)$ has a potential to show a limit cycle if the frequency due to the reactive components (L_2 and C_1) is greater than the one due to resistive components (R_2 and R_3). Note the components involved are those of the system *push*ed by $G_{11}(s)$ (past the supply duct).

Finding poles of $G_{11}(s)$ is a much more difficult task unless numerical values are given, because the third order denominator is not easily decomposed into simple factors. Instead of trying to find poles, we use the equation, $\mathrm{Im}[G_{11}(j\omega)] = 0$, to see under which condition the solutions exist (poles are on the imaginary axis).

[15] In a simple *L*-*R*-*C* circuit, the complex poles, if they exist, are given by $s = -a \pm j\omega$, where $a = R/2L$, $\omega^2 = \omega_n^2 - a^2$ and $\omega_n^2 = 1/LC$. ω_n is called *natural frequency* of oscillation and ω is *damped frequency* of oscillation.

We rewrite $G_{11}(s)$ as,

$$G_{11}(s) = \frac{\alpha(s^2 + Ds + E)}{s^3 + As^2 + Bs + C}. \tag{B.23}$$

After working out all the algebraic details,

$$\text{Im}[G_{11}(j\omega)] = -\omega(\omega^4 - (B + E - AD)\omega^2 + (CD - BE)) = 0. \tag{B.24}$$

$\omega = 0$ is a trivial solution (one system pole is always on the real axis). To show a limit cycle, there should be four crossover points on the root locus. Thus the term inside the parenthesis should have four real roots, i.e.,

$$\omega^2 = \frac{(B + E - AD) \pm \sqrt{(B + E - AD)^2 - 4(CD - BE)}}{2} \tag{B.25}$$

should have two real positive solutions. It is possible when

$$B + E - AD > 0 \text{ and } (B + E - AD)^2 - 4(CD - BE) \geq 0. \tag{B.26}$$

The first condition is worth examining.

$$B + E - AD = \frac{1}{L_2 C_1}\left(\frac{L_1 + L_2}{L_1} + 1\right) - \left(\left(\frac{R_3}{L_2}\right)^2 + \frac{R_1 + R_2}{C_1 L_1 R_2} + \frac{R_3(R_1 + R_2)}{L_1 L_2}\right). \tag{B.27}$$

The first term of the equation is the frequency squared due to the reactive, energy storage, components of the system. The second is the frequency squared due to the resistive, energy dissipative, components. (both in an aggregate sense.) Thus the system can exhibit a limit cycle if the aggregate reactive frequency is greater than the aggregate resistive frequency.

The second condition is difficult to analyze due to the square term involved. However, we surely know that the condition is satisfied if $CD - BE$ is nonpositive.

$$CD - BE = \frac{1}{C_1 L_1 L_2}\left(\frac{R_1 + R_2 + R_3}{C_1 R_2} + \frac{R_3(R_1 + R_2 + R_3)}{L_2}\right) - \frac{1}{C_1 L_1 L_2}\frac{L_1 + L_2}{C_1 L_2}\left(1 + \frac{R_3}{R_2}\right). \tag{B.28}$$

The condition is satisfied when the second term, reactive, is greater than the first term, resistive.

Finally, we look at the harmonic balance equation with $k = 0$:

$$
\begin{aligned}
& a_0 - G_{11}(0)c_{01} - G_{12}(0)c_{02} = 0 \\
& a_0 = \frac{R_2 + R_3}{R_1 + R_2 + R_3} c_{01} - \frac{R_1}{R_1 + R_2 + R_3} c_{02} \\
& a_0 \approx \frac{R_2 + R_3}{R_1 + R_2 + R_3} m_s - \frac{R_1}{R_1 + R_2 + R_3} m_r .
\end{aligned}
\tag{B.29}
$$

We need two transfer functions in this equation, because the supply duct pressure P_s is contributed by two inputs. a_0 is the DC value of P_s, and c_{01} and c_{01} are the *zero*th Fourier series coefficient of e_s, the first component of **u** vector, and $-e_r$, the second component of **u** vector, respectively. Thus we obtain an equation about the resistive components of system, if we know a_0, m_s, and m_r, which are easy to measure. (a_0 is obtained by averaging the supply duct pressure measurements obtained for a period of time. m_s and m_r are obtained by driving the SF/RF VFDs in open loop mode with the commands of fixed offset values (19.25mA and 18.3 mA) and measuring the pressures right after the fans.)

We have developed a very simple model of VAV AHU system, and tried to explain the sustained oscillation behavior as a limit cycle due to the saturation of actuators. We have not proven that there exists a limit cycle, because the numerical values of the system are unknown at this point. However, we have showed a variety of conditions that lead to the development of limit cycle. Also a few equations are derived, which in turn can be used to find system parameters. For example, we know that $\omega = 2\pi/30$ rad/sec is the solution to $\mathrm{Im}[G_{11}(j\omega)]$, because it is the limit cycle frequency. We also know the relationship between the resistive components, the signal DC value and the input offset. Further efforts are required to prove the model, such as finding numerical parameter values and/or improving models to include the effect of dampers, etc. However, once proved, the model can generally be applied to any type of VAV AHU system and in turn can be used to find a few essential system parameters.

Having gone through all the gory details of modeling and analysis, a lingering question is how we can eliminate the oscillation, if necessary[16]. One obvious solution is to reduce the gain so that the closed system poles never cross the imaginary axis. However, this can lead to a poor system performance and essentially is no different from running the system in an open loop mode. Another option is to design a more sophisticated controller, so that the closed loop poles will never have a chance to cross the imaginary axis. This is possible when the model is proved and the necessary parameters are obtained.

[16] We also haven't asked the question that if the system is described in such a way, which is the *best* way to control the system. This is the issue of designing the optimal control logic for an AHU system.

Appendix C

Estimation of First Order Polynomial Mean of a Random Process

This short appendix develops a simple method to estimate the time varying mean of a random process, as a first order polynomial of time. The random process is first modeled as white Gaussian and the maximum-likelihood (ML) estimator is derived. It is shown that the least-squares error criterion leads to the same formula and that the ML estimator can also be used with a non-Gaussian additive noise in a least-squared error sense. The first order polynomial mean estimator is mostly used to estimate the mean of a time series of a higher harmonic power, in order to estimate the mean of the fundamental power consumed by a variable speed drive through a correlation between two powers.

Suppose that **x** is a length-N observation of a random process, e.g. an apparent power (A_k),

$$\mathbf{x} = \begin{bmatrix} x[0] & x[1] & \dots & x[N-1] \end{bmatrix}^T \tag{C.1}$$

We assume that $x[n]$ is the sum of a first order polynomial mean and a white noise:

$$x[n] = m_x[n] + w[n] \quad 0 \le n \le N-1, \tag{C.2}$$

where $m_x[n] = \alpha n + \beta$ and $\text{var}(w[n]) = \sigma^2$. With this signal model, the observation vector **x** follows a Gaussian distribution with the probability density function (PDF) of:

$$\begin{aligned} p(\mathbf{x} \mid \alpha, \beta) &= \frac{1}{(2\pi)^{\frac{N}{2}} \det^{\frac{1}{2}}(\mathbf{\Lambda_x})} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m_x})^T \mathbf{\Lambda_x}^{-1} (\mathbf{x} - \mathbf{m_x}) \right) \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - \alpha n - \beta \right)^2 \right), \end{aligned} \tag{C.3}$$

where α and β are the parameters of the distribution. The maximum likelihood (ML) estimations of these parameters are found by maximizing the logarithm of the PDF with respect to α and β. Because the logarithm of the PDF is a quadratic function of α and β, the maximum occurs where the first derivatives are zero.

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x} \mid \alpha, \beta)}{\partial \alpha} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} n(x[n] - \alpha n - \beta) = 0 \\ \frac{\partial \ln p(\mathbf{x} \mid \alpha, \beta)}{\partial \beta} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - \alpha n - \beta) = 0. \end{aligned} \tag{C.4}$$

By combining these two equations, we obtain the following single matrix equation:

$$\mathbf{A}\mathbf{a} = \mathbf{b}\,, \tag{C.5}$$

where

$$\mathbf{A} = \begin{bmatrix} \sum_{n=0}^{N-1} n^2 & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} 1 \end{bmatrix} = \begin{bmatrix} \frac{N(N-1)(2N-1)}{6} & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & N \end{bmatrix}, \mathbf{a} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{b} = \begin{bmatrix} \sum_{n=0}^{N-1} n x[n] \\ \sum_{n=0}^{N-1} x[n] \end{bmatrix}. \tag{C.6}$$

The matrix **A** is invertible whenever $N \geq 2$ and does not depend on $x[n]$. Thus the ML estimation of **a** can be computed easily by inverting **A**, which is a function of N only.

The same formula can be derived by using the least-squares (LS) error criterion, i.e.

$$\mathbf{x} \approx \alpha\mathbf{n} + \beta\mathbf{1} = \mathbf{H}\mathbf{a}\,, \tag{C.7}$$

where $\mathbf{H} = [\mathbf{n}\ \mathbf{1}]$ and $\mathbf{n} = [0\,1\,\ldots\,N-1]^T$. The LS estimation of **a** is given by:

$$\hat{\mathbf{a}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\,. \tag{C.8}$$

Readers can easily verify that $\mathbf{H}^T\mathbf{H} = \mathbf{A}$ and $\mathbf{H}^T\mathbf{x} = \mathbf{b}$. This is not a coincidence because the LS estimator is the best linear unbiased estimator (BLUE) when the additive noise is zero-mean white Gaussian due to the Gauss-Markov theorem [Kay, 1993]. The BLUE is a minimum variance unbiased (MVU) estimator constrained to be linear. On the other hand, an ML estimator is approximately the MVU estimator when its record size is sufficiently large.

Thus when the additive noise is white Gaussian, we have an ML first order polynomial mean estimator, which is approximately the MVU estimator. When the noise is not white, the same mean estimator is still valid in the LS error sense. The latter point of view is especially helpful because power signals typically carry colored noises, such as those generated by variable speed drives.

When the first order mean estimation is performed on each succeeding window, we may impose an additional constraint that the mean estimation is continuous across window boundaries. Suppose that $\mathbf{m}_{\mathbf{x},j}$ is the length-N mean vector of the jth window and that $\mathbf{m}_{\mathbf{x},j\text{-}1}$ is the length-N mean vector of the preceding (j-1)th window. The constraint can be stated as $m_{x,j\text{-}1}[N] = m_{x,j}[0]$. In terms of parameters:

$$\alpha_{j-1}N + \beta_{j-1} = \beta_j\,. \tag{C.9}$$

Thus when the continuity constraint is applied, the parameter β of a current window is determined solely by the parameters of its preceding window. Then the remaining parameter α_j is obtained by solving the first derivative condition,

$$\alpha_j = \frac{\sum_{n=0}^{N-1} nx[n] - \beta_j \sum_{n=0}^{N-1} n}{\sum_{n=0}^{N-1} n^2} = \frac{\sum_{n=0}^{N-1} nx[n] - \beta_j \frac{N(N-1)}{2}}{\frac{N(N-1)(2N-1)}{6}}. \quad \text{(C.10)}$$

Appendix D

ELIS Software Architecture

This appendix gives a brief overview of the ELIS software architecture at block diagram level. The ELIS software performs signal analysis and information extraction on the non-intrusive power signals generated by a power coder. It also displays acquired information to users on hyper text markup language (HTML) pages that are accessible via the internet. The ELIS is currently written with C++ and executed on a Linux platform.

The power coder performs sampling of raw voltage and current waveforms and computation of current harmonic powers. It currently is configured to output eight power terms (P, Q, P_3, Q_3, P_5, Q_5, P_7, Q_7) at 120 Hz. The power terms become input to the ELIS via a *pipe*[17]. The pipe also allows opening a prerecorded data file and supplying the readings as input to the ELIS, which is useful to test ELIS programs offline.

A real-time data handling software typically consists of three main components– initialization, main loop and termination. The main loop is called repeatedly either at a fixed frequency or with the occurrence of a triggering event. The ELIS also has these three components and its main loop is called with a fixed interval. Figure D-1 shows the block diagram of the ELIS.

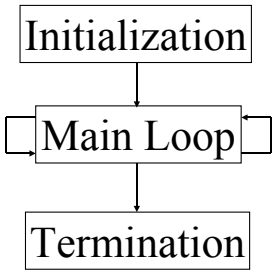


Figure D-1 Three main components of ELIS.

The initialization block reads load database and process parameters (classification methods, thresholds, options, etc.), assigns memory and creates variables. It also initializes HTML pages for webcasting. The HTML pages are generated as text files and interpreted by web browsers. Figure D-2 shows the block diagram of the ELIS initialization.

Currently, the constant load disaggregator and the VSD tracker are implemented as separate entities, although they will be merged as a single ELIS program in future. The main loop of the load disaggregator performs data acquisition, event detection and classification and HTML page update. The load disaggregator has three HTML pages–

[17] Pipe is a file-handling interface in Linux. It simply connects one process's output to another process's input. '|' is the shell command used to create a pipe.

main, event and history. Figure D-3 shows the block diagram of the main loop of the constant load disaggregator.

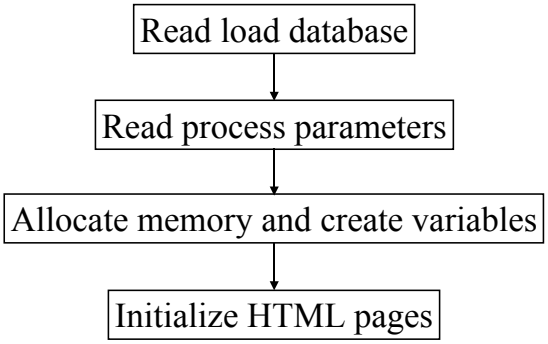


Figure D-2 Initialization block of ELIS.

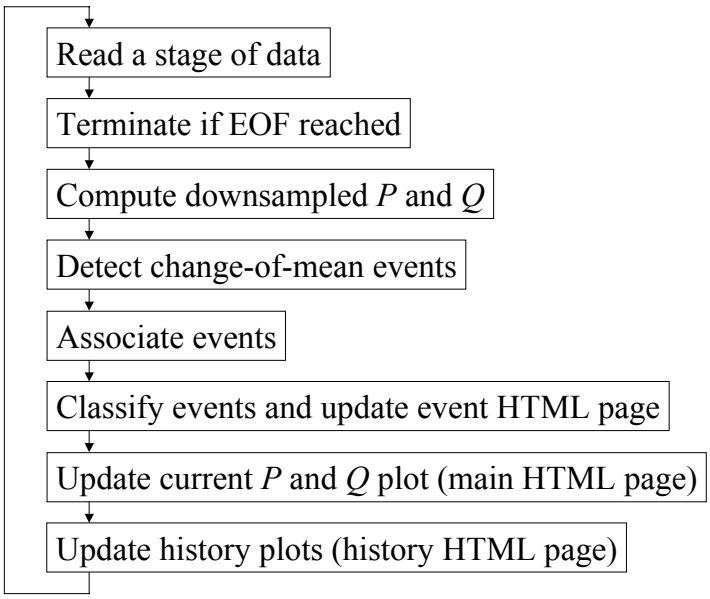


Figure D-3 Main loop of constant load diaggregator.

The main loop first reads a stage of data. A stage typically consists of 1024 or 1200 rows of power terms. Thus, with the 120 Hz power coding rate, the main block is executed at every 8.53 or 10 sec. If an EOF (end of file) is reached during readout, the main block is terminated immediately and the program control is forwarded to the termination block. After a stage of data is successfully retrieved, *P* and *Q* streams are downsampled for the change-of-mean detection.

The change-of-mean detector examines the incoming *P* stream and converts it to a series of constant-mean segments. The event associator analyzes the series and locates load events. The load events are forwarded to the event handler that classifies the events and posts the results on the event HTML page. An event figure is created by *gnuplot* and its link and event tag are inserted to the HTML page. Also, the plots of real and reactive

powers are generated for the main HTML page and the plots of the time history of various powers are created for the history HTML page. Finally, the main block goes back to read a stage of data and the process is repeated.

The VSD tracker has three HTML pages– main, PSD estimation and history. Its main block is generally similar to that of the constant load disaggregator, but has different signal analysis routines. Figure D-4 shows the block diagram of the VSD tracker main loop.

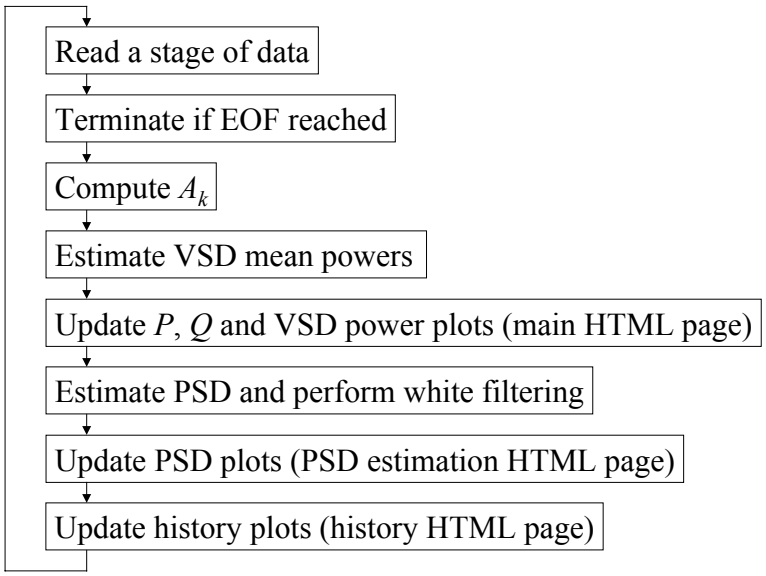


Figure D-4 Main loop of VSD tracker.

The VSD tracker main loop first reads a stage of data and checks if an EOF is reached. Then, it computes necessary higher harmonic apparent powers from the stage of data. First order mean estimation is performed on a selected apparent power, and the mean fundamental power consumed by VSDs is estimated using a known correlation. The estimation results are displayed as figures on the main HTML page. The AR PSD of current stage signal is estimated and its parameters are used to construct a whitening filter for leftover real and reactive powers after their VSD means are compensated. The PSD estimation and white filtering results are again displayed as figures on the PSD estimation HTML page. Finally, history plots are updated for the history HTML page.

The termination blocks of both the constant load disaggregator and the VSD tracker simply release dynamically allocated memory to prevent memory leak.

Bibliography

R. A. Abdoo, "Industry Restructuring for Deregulation and Competition," *IEEE Power Engineering Review*, pp. 16 – 18, July 1999.

H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716 – 723, 1974.

J. Arrillaga, M. H. J. Bollen and N. R. Watson, "Power Quality following Deregulation," *Proceedings of the IEEE*, Vol. 88, No. 2, pp. 246 – 261, February 2000.

B. R. Barkovich and D. V. Hawk, "Charting a New Course in California," *IEEE Spectrum*, pp. 26 – 31, July 1996.

M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.

H. M. Beides and G. T. Heydt, "Dynamic State Estimation of Power System Harmonics using Kalman Filter Methodology," *IEEE Transactions on Power Delivery*, Vol. 6, No. 4, pp. 1663 – 1670, October 1991.

D. P. Bertsekas, *Dynamic Programming*, Prentice Hall, Englewood Cliffs, NJ, 1987.

J. A. C. Bingham, *The Theory and Practice of Modem Design*, Wiley, New York, 1988.

B. K. Bose, "Variable Frequency Drives- Technology and Applications", Conference Proceedings of *the IEEE International Symposium on Industrial Electronics*, ISIE'93, Budapest, pp.1 – 18, 1993.

G. Bucci, E. Fiorucci and C. Landi, "Digital Measurement Station for Power Quality Analysis in Distributed Environments," *IEEE Transactions on Instrumentation and Measurement*, Vol. 52, No. 1, pp. 75 – 84, February 2003.

V. S. Budhraja, "California's Electricity Crisis," *IEEE Power Engineering Review*, pp. 6 – 14, August 2002.

S. R. Bull, "Renewable Energy Today and Tomorrow," *Proceedings of the IEEE*, Vol. 89, No. 8, pp. 1216 – 1226, August 2001.

L. Carmichael, "Nonintrusive Appliance Load Monitoring System," *EPRI Journal*, Electric Power Research Institute, pp. 45 – 47, September 1990.

H. Chen, W. C. Kong and C. C. Ko, "Comparison of Pitch Prediction and Adaptation Algorithms in Forward and Backward Adaptive CELP Systems," *IEE Proceedings on Communications, Speech and Audio Vision*, Vol. 140, No. 4, pp. 240 – 245, 1993.

Y.-F. Chen and T.-D. Chiueh, "Baseband Transceiver Design of a 128-Kbps Power-Line Modem for Household Applications," *IEEE Transactions on Power Delivery*, Vol. 17, No. 2, pp. 338 – 344, April 2002.

Y.-S. Choi and D.-H. Youn, "Fast Harmonic Estimation Method for Harmonic Speech Coders," *Electronics Letters*, Vol. 38, No. 7, pp. 346 – 347, 2002.

A. I. Cole and A. Albicki, "Data Extraction for Effective Non-Intrusive Identification of Residential Power Loads," *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, Vol. 2, pp. 812 – 815, May 1998.

A. I. Cole and A. Albicki, "Algorithm for Non-Intrusive Identification of Residential Appliances," *Proceedings of the IEEE International Symposium on Circuits and Systems* (ISCAS), Vol. 3, pp. 338 – 341, 1998.

B. Cook, "Introduction to Fuel Cells and Hydrogen Technology," *IEE Engineering Science and Education Journal*, pp. 205 – 216, December 2002.

T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, Second Edition, MIT Press, Cambridge, Massachusetts, 2001.

M. L. Crow, "Power System Deregulation," *IEEE Potentials*, pp. 8 – 9, January 2002.

P. K. Dash, R. K. Jena, G. Panda and A. Routray, "An Extended Complex Kalman Filter for Frequency Measurement of Distorted Signals," *IEEE Transactions on Instrumentation and Measurement*, Vol. 49, No. 4, pp. 746 – 753, August 2000.

J. R. Deller, J. H. L. Hansen and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Piscataway, NJ, 2000.

E. L. Draper, "Assessment of Deregulation and Competition," *IEEE Power Engineering Review*, pp. 17 – 18, July 1998.

S. Drenker and A. Kader, "Nonintrusive Monitoring of Electric Loads," *IEEE Computer Applications in Power*, Vol. 12, No. 4, pp. 47 – 51, October 1999.

J. Driesen, T. Van Craenenbroeck and D. Van Dommelen, "The Registration of Harmonic Power by Analog and Digital Power Meters," *IEEE Transactions on Instrumentation and Measurement*, Vol. 47, No. 1, pp. 195 – 198, February 1998.

M. El-Habrouk and M. K. Darwish, "Design and Implementation of a Modified Fourier Analysis Harmonic Current Computation Technique for Power Active Filters using DSPs," *IEE Proceedings on Electric Power Applications*, Vol. 148, No. 1, pp. 21 – 28, January 2001.

A. Emadi, M. Ehsani and J. M. Miller, "Advanced Silicon Rich Automotive Electrical Power Systems," *Proceedings of the Eighteenth Digital Avionics Systems Conference*, Vol. 2, pp. 8.B.1-1 – 8.B.1-8, 1999.

A. E. Emanuel, "Powers in Nonsinusoidal Situations: A Review of Definitions and Physical Meaning," *IEEE Transactions on Power Delivery*, Vol. 5, No. 3, pp. 1377 – 1389, July 1990.

A. E. Emanuel, J. A. Orr, D. Cyganski and E. M. Gulachenski, "A Survey of Harmonic Voltages and Currents at Distribution Substations," *IEEE Transactions on Power Delivery*, Vol. 6, No. 4, pp. 1883 – 1890, October 1991.

A. E. Emanuel, J. A. Orr, D. Cyganski and E. M. Gulachenski, "A Survey of Harmonic Voltages and Currents at the Customer's Bus," *IEEE Transactions on Power Delivery*, Vol. 8, No. 1, pp. 411 – 421, January 1993.

A. E. Emanuel, "Harmonics in the Early Years of Electrical Engineering: A Brief Review of Events, People and Documents," *Proceedings of the Ninth International Conference on Harmonics and Quality of Power*, Vol. 1, pp. 1 – 7, 2000.

L. Erup, F. M. Gardner and R. A. Harris, "Interpolation in Digital Modems- Part II: Implementation and Performance," *IEEE Transactions on Communications*, Vol. 41, No. 6, pp. 998 – 1008, June 1993.

M. Farooque and H. C. Maru, "Fuel Cells- The Clean and Efficient Power Generators," *Proceedings of the IEEE*, Vol. 89, No. 12, pp. 1819 – 1829, December 2001.

G. F. Franklin, J. D. Powell and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Third Edition, Addison Wesley, Reading, MA, 1994.

K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, NY, 1972.

F. M. Gardner, "Interpolation in Digital Modems- Part I: Fundamentals," *IEEE Transactions on Communications*, Vol. 41, No. 3, pp. 501 – 507, March 1993.

J. E. Gibson, *Nonlinear Control System*, McGraw-Hill, NY, 1963.

H. S. Gill, V. Filipovic-Gledja, H. P. Mukherjee, T. S. Sidhu and L. Cui, "Software Quality Tools for the Deregulated Market," *IEEE Computer Applications in Power*, pp. 33 – 38, October 2001.

E. H. Graf and M. S. Lee, "All-Air Systems," in *HVAC Systems and Components Handbook*, N. R. Grimm and R. C. Rosaler, Eds, McGraw-Hill, NY, 1998.

D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 36, No. 8, pp. 1223 – 1235, 1988.

E. W. Gunther, "Interharmonics- Recommended Updates to IEEE 519," *IEEE Power Engineering Society Summer Meeting*, Vol. 2, pp. 950 – 954, 2002.

P. G. Harris, "Impacts of Deregulation on the Electric Power Industry," *IEEE Power Engineering Review*, pp. 4 – 6, October 2000.

G. W. Hart, "Nonintrusive Appliance Load Monitoring," *Proceedings of the IEEE*, Vol. 80, No. 12, pp. 1870 – 1891, December 1992.

IEEE Industrial Applications Society/Power Engineering Society, "IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems," *IEEE Standard 519-1992*, NY, 1993.

IEEE Working Group on Nonsinusoidal Situations, "A Survey of North American Electric Utility Concerns Regarding Nonsinusoidal Waveforms," *IEEE Transactions on Power Delivery*, Vol. 11, No. 1, pp. 73 – 78, January 1996.

IEEE Working Group on Nonsinusoidal Situations: Effects on Meter Performance and Definitions of Power, "Practical Definitions for Powers in Systems with Nonsinusoidal Waveforms and Unbalanced Loads: A Discussion," *IEEE Transactions on Power Delivery*, Vol. 11, No. 1, pp. 79 – 101, January 1996.

M. Ilic, P. Skantze and P. Visudhiphan, "Electricity Troubles in California: Who's Next?" *IEEE Spectrum*, pp. 11 –13, February 2001.

E. Jacobsen and R. Lyons, "The Sliding DFT," *IEEE Signal Processing Magazine*, Vol. 20, No. 2, pp. 74 – 80, March 2003.

W. C. Karl, S. B. Leeb, L, A. Jones, J. L. Kirtley and G. C. Verghese, "Applications of Rank-Based Filters in Power Electronics", *IEEE Transactions on Power Electronics*, Vol. 7, No. 3, pp. 437 – 443, July 1992.

D. C. Karnopp, D. L. Margolis and R. C. Rosenberg, *System Dynamics: A Unified Approach*, Second Edition, Wiley-InterScience, NY, 1990.

J. G. Kassakian, H.-C. Wolf, J. M. Miller and C. J. Hurton, "Automotive Electrical Systems circa 2005," *IEEE Spectrum*, pp. 22 – 27, August 1996.

J. G. Kassakian, "Automotive Electrical Systems- The Power Electronics Market of the Future," *Fifteenth Annual IEEE Applied Power Electronics Conference and Exposition*, Vol. 1, pp. 3 – 9, 2000.

S. M. Kay and S. L. Marple, "Spectrum Analysis- A Modern Perspective," *Proceedings of the IEEE*, Vol. 69, No. 11, pp. 1380–1419, 1981.

S. M. Kay, "Spectral Estimation," in *Advanced Topics in Signal Processing*, J. S. Lim and A. V. Oppenheim, Eds, Prentice Hall, Englewood Cliffs, NJ, 1988.

S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall, Englewood Cliffs, NJ, 1993.

S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Prentice Hall, Upper Saddle River, NJ, 1998.

G. Kendall, "Power Outages during Market Deregulation," *IEEE Control Systems Magazine*, pp. 33 – 39, December 2001.

B. W. Kennedy, *Power Quality Primer*, McGraw-Hill, New York, 2000.

I. A. Khan, "Automotive Electrical Systems: Architecture and Components," *Proceedings of the Eighteenth Digital Avionics Systems Conference*, Vol. 2, pp. 8.C.5-1 – 8.C.5-10, 1999.

P. Koponen, R. Seesvuori and R. Bostman, "Adding Power Quality Monitoring to a Smart kWh Meter," *IEE Power Engineering Journal*, pp. 159 – 163, August 1996.

T. I. Laakso, V. Valimaki, M. Karjalainen and U. K. Laine, "Splitting the Unit Delay," IEEE Signal Processing Magazine, Vol. 13, No. 1, pp. 30 – 60, January 1996.

C. W. Lander, *Power Electronics*, Second Edition, McGraw-Hill, London, 1987.

C. Laughman, K. D. Lee, R. Cox, S. Shaw, S. B. Leeb, L. K. Norford and P. Armstrong, "Advanced Nonintrusive Monitoring of Electric Loads," accepted by *IEEE Power and Energy Magazine* (to appear on March/April 2003).

W. J. Lee and C. H. Lin, "Utility Deregulation and Its impact on Industrial Power Systems," *IEEE Industry Applications Magazine*, pp. 40 – 46, May 1998.

S. B. Leeb, *A Conjoint Pattern Recognition Approach to NonIntrusive Load Monitoring*, Ph. D. Thesis, MIT, Department of Electrical Engineering and Computer Science, 1993.

S. B. Leeb, S. R. Shaw and J. L. Kirtley, "Transient Event Detection in Spectral Envelope Estimates for Nonintrusive Load Monitoring," *IEEE Transactions on Power Delivery*, Vol. 10, No. 3, pp. 1200 – 1210, July 1995.

S. B. Leeb, B. C. Lesieutre and S. R. Shaw, "Determination of Load Composition using Spectral Envelope Estimates," *Proceedings of the 27th Annual North American Power Symposium*, IEEE Power Engineering Society, Bozeman, Montana, October 2 – 3, 1995.

C. Li, W. Xu and T. Tayjasanant, "Interharmonics: Basic Concepts and Techniques for Their Detection and Measurement," to appear on *Journal of Electric Power Systems Research*, (Paper No.: 1859), 2003.

J. S. Lim and A. V. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 26, No. 3, pp. 197 – 210, 1978.

J. S. Lim, A. V. Oppenheim and L. D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 26, No. 4, pp. 354–358, 1978.

Y.-J. Lin, H. A. Latchman, M. Lee and S. Katar, "A Power Line Communication Network Infrastructure for the Smart Home," *IEEE Wireless Communications*, pp. 104 – 111, December 2002.

D. Luo, *Detection and Diagnosis of Faults and Energy Monitoring of HVAC Systems with Least-Intrusive Power Analysis*, Ph. D. Thesis, MIT, Department of Architecture, 2001.

D. Luo, L. K. Norford, S. R. Shaw, and S. B. Leeb, "Monitoring HVAC Equipment from a Centralized Location- Methods and Field Test Results", *ASHRAE Transactions* Vol. 108, No. 1, 2002.

Z. Luo, Z. Xu, Y. Zheng and X. Lu, "DFT and DSP-based Electric Energy Measurement Algorithm of Harmonic Source Load," *Proceedings on the International Conference on Power System Technology (PowerCon)*, Vol. 4, pp. 2487 – 2490, 2002.

S. L. Marple, "A New Autoregressive Spectrum Analysis Algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, pp. 441 – 454, 1980.

S. L. Marple, *Digital Spectral Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1987.

P. Martin, "Comparison of Pitch Detection by Cepstrum and Spectral Comb Analysis," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Paris, France, pp. 180 – 183, May 1982.

R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Detection based on a Sinusoidal Speech Model," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Vol. 1, pp. 249 – 252, April 1990.

R. H. McClanahan, "Electric Deregulation," *IEEE Industry Applications Magazine*, pp. 11 – 18, April 2002.

J. H. McClellan, "Parametric Signal Modeling," in *Advanced Topics in Signal Processing*, J. S. Lim and A. V. Oppenheim, Eds, Prentice Hall, Englewood Cliffs, NJ, 1988.

W. McNamara, *The California Energy Crisis: Lessons for a Deregulating Industry*, Penn Well, Tulsa, Oklahoma, 2002.

J. M. Miller, "Multiple Voltage Electrical Power Distribution System for Automotive Applications," *Proceedings of the 31st Intersociety Energy Conversion Engineering Conference* (IECEC), Vol. 3, pp. 1930 – 1937, August 1996.

J. M. Moreno-Eguilaz, J. Peracaula and A. Esquivel, "Neural Network based Approach for the Computation of Harmonic Power in a Real-time Microprocessor-based Vector Control for an Induction Motor Drive," *Proceedings of the IEEE International Symposium on Industrial Electronics*, Vol. 1, pp. 277 – 281, 2000.

D. P. Morgan, E. B. George, L. T. Lee and S. M. Kay, "Cochannel Speaker Separation by Harmonic Enhancement and Suppression," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 5, pp. 407–424, 1997.

R. Naik, N. Mohan, M. Rogers and A. Bulawka, "A Novel Grid Interface, Optimized for Utility-scale Applications of Photovoltaic, Wind-electric and Fuel-cell Systems," *IEEE Transactions on Power Delivery*, Vol. 10, No. 4, pp. 1920 – 1926, October 1995.

J. A. Naylor and S. F. Boll, "Techniques for Suppression of an Interfering Talker in Co-channel Speech," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Dallas, TX, pp. 205 – 208, April 1987.

J. Naylor and J. Porter, "An Effective Speech Separation System which Requires No *A Priori* Information," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Toronto, Ontario, Canada, pp. 937 – 940, May 1991.

J. Neubert, "Powering up," *IEE Review*, pp. 21 – 25, September 2000.

A. Nijenhuis and H. S. Wilf, *Combinatorial Algorithms*, Academic Press, NY, 1975.

Nils J. Nilsson, *Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, 1998.

L. K. Norford and S. B. Leeb, "Non-intrusive Electrical Load Monitoring in Commercial Buildings based on Steady-state and Transient Load-detection Algorithms," *Energy and Buildings*, Vol. 24, No. 1, pp. 51 – 64, 1996.

J. A. Oliver, R. Lawrence and B. B. Banerjee, "Power Quality: How to Specify Power-Quality-Tolerant Process Equipment," *IEEE Industry Applications Magazine*, pp. 21 – 30, September 2002.

A. V. Oppenheim, A. S. Willsky and S. H. Nawab, *Signals and Systems*, second edition, Prentice-Hall, Englewood Cliffs, NJ, 1998.

A. V. Oppenheim, R. W. Schafer and J. R. Buck, *Discrete-time Signal Processing*, second edition, Prentice-Hall, Upper Saddle River, NJ, 1999.

S. Osowski, "Neural Network for Estimation of Harmonic Components in a Power System," *IEE Proceedings-C*, Vol. 139, No. 2, pp. 129 – 135, March 1992.

S. Osowski, "SVD Technique for Estimation of Harmonic Components in a Power System: A Statistical Approach," *IEE Proceedings on Generation, Transmission and Distribution*, Vol. 141, No. 5, pp. 473 – 479, September 1994.

R. Palomera-Arias and L. K. Norford, *Applications of Non-intrusive Electrical Load Monitoring to the Detection and Diagnosis of Faults Associated with the On-off Switching of HVAC Equipment*, Technical Report to Honeywell, Inc., MIT Department of Architecture, August 9, 2001.

A. Papoulis, S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, Fourth Edition, McGraw-Hill, NY, 2002.

T. W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *Journal of the Acoustical Society of America*, Vol. 60, No. 4, pp. 911 – 918, 1976.

E. Parzen, "Some Recent Advances in Time Series Modeling," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 723 – 730, 1974.

N. Pavlidou, A. J. Han Vinck, J. Yazdani and B. Honary, "Power Line Communications: State of the Art and Future Trends," *IEEE Communications Magazine*, pp. 34 – 40, April 2003.

L. Philipson and H. L. Willis, *Understanding Electric Utilities and De-Regulation*, Marcel Dekker, NY, 1998.

V. F. Pisarenko, "The Retrieval of Harmonics from a Covariance Function," *Geophysical Journal of Royal Astronomical Society*, Vol. 33, pp. 347 – 366, 1973.

A. Price, S. Bartley, S. Male and G. Cooley, "A Novel Approach to Utility Scale Energy Storage," *IEE Power Engineering Journal*, pp. 122 – 129, June 1999.

J. G. Proakis, C. M. Rader, F. Ling, C. L. Nikias, M. Moonen and I. K. Proudler, *Algorithms for Statistical Signal Processing*, Prentice Hall, Upper Saddle River, NJ. 2002.

T. F. Quatieri and R. G. Danisewicz, "An Approach to Co-channel Talker Interference Suppression using a Sinusoidal Model for Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, No. 1, pp. 56 – 69, 1990.

S. Rechka, E. Ngandui, J. Xu and P. Sicard, "A Comparative Study of Harmonic Detection Algorithms for Active Filters and Hybrid Active Filters," *The 33rd Annual IEEE Power Electronics Specialists Conference (PSEC)*, Vol. 1, pp. 357 – 363, 2002.

E. M. Reingold, J. Nievergelt and N. Deo, *Combinatorial Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

J. Rissanen, "Modeling by Shortest Data Description", *Automatica*, Vol. 14, pp. 465-471, 1978.

J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *Annals of Statistics*, Vol. 11, No. 2, pp. 416 – 431, 1983.

R. Roy and T. Kailath, "ESPRIT- Estimation of Signal Parameters via Rotational Invariance Techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 7, pp. 984 – 995, 1989.

H. Saari, P. Koponen, E. Tahvanainen and T. Lindholm, "Remote Reading and Data Management System for kWh Meters with Power Quality Monitoring," *IEE Power Engineering Journal*, pp. 164 – 168, August 1996.

S. R. Shaw, C. B. Abler, R. F. Lepard, D. Luo, S. B. Leeb and L. K. Norford, "Instrumentation for High Performance Nonintrusive Electrical Load Monitoring," *ASME Journal of Solar Energy Engineering*, Vol. 120, No. 3, August 1998.

S. R. Shaw and S. B. Leeb, "Identification of Induction Motor Parameters from Transient Stator Current Measurements," *IEEE Transactions on Industrial Electronics*, Vol. 46, No. 1, pp. 139 – 149, February 1999.

S. R. Shaw, *System Identification Techniques and Modeling for Nonintrusive Load Diagnostics*, Ph. D. Thesis, MIT, Department of Electrical Engineering and Computer Science, 2000.

S. R. Shaw, D. Luo, L. K. Norford and S. B. Leeb, "Detection and Diagnosis of HVAC Faults via Electrical Load Monitoring," *International Journal of HVAC&R Research*, Vol. 8, No. 1, pp. 13 – 40, 2002.

P. Stoica and A. Nehorai, "MUSIC, Maximum Likelihood and Cramer-Rao Bound," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 5, pp. 720 – 741, 1989.

J. Stones and A. Collinson, "Power Quality," *IEE Power Engineering Journal*, pp. 56 – 64, April 2001.

R. J. Stubbs and Q. Summerfield, "Evaluation of Two Voice-separation Algorithms using Normal-hearing and Hearing-impaired Listeners," *Journal of the Acoustical Society of America*, Vol. 84, No. 4, pp. 1236 – 1249, 1988.

F. Sultanem, "Using Appliance Signatures for Monitoring Residential Loads at Meter Panel Level," *IEEE Transactions on Power Delivery*, Vol. 6, No. 4, pp. 1380 – 1385, October 1991.

J. L. Sweeney, *The California Electricity Crisis*, Hoover Institution Press, Stanford, California, 2002.

J. G. Truxell, *Automatic Feedback Control System Synthesis*, McGraw-Hill, NY, 1955.

T. J. Ulrych and R. W. Clayton, "Time Series Modeling and Maximum Entropy," *Physics of the Earth and Planetary Interiors*, Vol. 12, pp. 188 – 200, 1976.

T. L. Vincent and W. J. Grantham, *Nonlinear and Optimal Control Systems*, John Wiley & Sons, NY, 1997.

R. P. Walker, "Why in the World Would an Electric Utility be Interested in Renewable Energy?" *IEEE Power Engineering Review*, pp. 9 – 12, May 1998.

J. V. Wijayakulasooriya, G. A. Putrus and P. D. Minns, "Electric Power Quality Disturbance Classification using Self-adapting Artificial Neural Networks," *IEE*

Proceedings on Generation, Transmission and Distribution, Vol. 149, No. 1, pp. 98 – 101, January 2002.

J. D. Wise, J. R. Caprio and T. W. Parks, "Maximum Likelihood Pitch Estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 24, No. 5, pp. 418–423, 1976.

H. Wold, *A Study in the Analysis of Stationary Time Series*, Second Edition, Almqvist & Wiksell, Stockholm, Sweden, 1954.

R. H. Wolk, "Fuel Cells for Homes and Hospitals," *IEEE Spectrum*, pp. 45 – 52, May 1999.

B. F. Wollenberg, "The Price of Change," *IEEE Potentials*, pp. 14 – 16, January 1998.

W. Xu, H. W. Dommel, M. B. Hughes, G. W. K. Chang, L. Tang, "Modeling of Adjustable Speed Drives for Power System Harmonic Analysis", *IEEE Transactions on Power Delivery*, Vol. 14, No. 2, pp. 595 – 601, April 1999.

W. Xu, "Status and Future Directions of Power System Harmonic Analysis," to appear on IEEE Power Engineering Society General Meeting, Toronto, Canada, July 13 – 17, 2003.

R. Yacamini, "Power System Harmonics, Part 1: Harmonic Sources," *IEE Power Engineering Journal*, pp. 193 – 198, August 1994.

R. Yacamini, "Power System Harmonics, Part 4: Interharmonics," *IEE Power Engineering Journal*, pp. 185 – 193, August 1996.

G. Zaccour, Ed., *Deregulation of Electric Utilities*, Kluwer, Norwell, MA, 1998.