Diagnosis and Prognosis of Automotive Systems: motivations, history and some results

Giorgio Rizzoni [*,**], **Simona Onori** [**], **Matteo Rubagotti** [***]

[*] *The Ohio State University, Department of Mechanical Engineering Department, and Center for Automotive Research, email: rizzoni.1@osu.edu*
[**] *The Ohio State University, Center for Automotive Research, email: onori.1@osu.edu*
[***] *University of Pavia, Department of Computer Engineering and Systems Science Italy, email: matteo.rubagotti@unipv.it*

Abstract: This paper presents an overview of diagnostic needs and methodologies in the automotive field. The field of automotive engineering has seen an explosion in the presence of electronic components and systems on-board vehicles since the 1970s. This growth was initially motivated by the introduction of emissions regulations that led to the widespread application of electronic engine controls. A secondary but important consequence of these developments was the adoption of on-board diagnostics regulations aimed at insuring that emission control systems would operate as intended for a prescribed period of time (or vehicle mileage). In addition, the presence of micro-controllers on-board the vehicle has led to a proliferation of other functions related to safety and customer convenience, and implemented through electronic systems and related software, thus creating the need for more sophisticated on-board diagnostics. Today, a significant percentage of the software code in an automobile is devoted to diagnostic functions. This paper presents an overview of diagnostic needs and requirements in the automotive industry, illustrates some of the challenges that are associated with satisfying these requirements, and proposes some future directions, in particular with respect to prognostics.

Keywords: Automotive applications, Fault-tolerant control, Linear and robust methods, Nonlinear methods

1. INTRODUCTION

Why diagnosis in the automotive field? The field of automotive engineering has seen an explosion in the presence of electronic components and systems on-board vehicles since the 1970s. This growth was initially motivated by the introduction of emissions regulations that led to the widespread application of electronic engine controls. A secondary but important consequence of these developments was the adoption of on-board diagnostics regulations aimed at insiring that emission control systems remained functional for a prescribed period of time (or vehicle mileage). In addition, the presence of microcontrollers on-board the vehicle led to a proliferation of functions implemented through electronic systems and related software, related to safety and customer convenience, creating the need for more sophisticated on-board diagnostics. Today, a significant percentage of the software code in an automobile is devoted to diagnostic functions. Of course, the use of diagnostic methods in automobiles is as old as the automobile itself. Repair technicians have adopted a variety of diagnostic techniques based on the human senses, from sight to sound to smell, for over a century. However, over the past quarter century this field has evolved from being in the domain of expert technicians, aided by test and measurement equipment, to the development of software that is embedded in the microcontrollers that manage functions ranging from engine performance and emissions, to braking, traction control, stability, and myriad customer convenience functions. The aim of this paper is to present an overview of the state of the art in on-board diagnostics in today's automobiles, and to suggest some future directions.

1.1 Emission regulation

The original motivation for the introduction of real-time on-board diagnostics in automotive vehicles originates from the California Air Resources Board (CARB) requirements introduced in the early 1990's to guarantee the integrity of the engine exhaust emissions control systems. The idea behind the original on-board diagnostics regulations [obd05], [EPA] was to guarantee that the exhaust emissions control system would be functional for a period of time associated with warranty or with regulated requirements. OBD regulations mandate that any fault in the emission control system affecting software algorithms, sensors, actuators or other hardware that could lead to

⋆ This work was supported in part by someone.

an increase of tailpipe emissions such that the vehicle would no longer meet the emissions regulations, should be detected in real-time and codified according to a set of on-board diagnostic codes that are described in the OBD legislation. These regulations first came into effect in 1988 and were further expanded in 1994 through OBD-II regulations, and affect every single component or subsystem that could increase engine exhaust emissions above a pre-specified threshold. With the growth in complexity in exhaust emissions regulations, and the attendant increase in complexity in the hardware and software required to meet such regulations, the task of meting OBD regulations has become quite challenging. In particular, the OBD challenge for Diesel engines and associated exhaust aftertreatment systems is notable, requiring the detection of faults that lead to very small changes in regulated exhaust gas emissions, of the order of tens of ppm. In a later section, this paper reviews three distinct approaches to solving OBD problems related to exhaust emissions controls.

1.2 Safety

A second motivation for the introduction of on-board diagnostic algorithms has been the introduction of safety systems on-board vehicles. In recent years, increasing attention to safety has led to the introduction of anti-lock braking systems, traction control systems, electronic stability control systems, and passive and active restraints. Many safety functions are also the subject of increasingly stringent regulations. The introduction of active systems that can affect the safety of a vehicle, such as braking, traction and stability control, and the introduction of by-wire systems to implement these functions, has generated different needs in diagnostics. In this context, diagnosis is a precursor to fault-tolerant control: if a safety-critical component is malfunctioning because of a fault or failure in a sensor, actuator or other component, or a malfunction in one of the software algorithms, then it is necessary to identify such safety-critical failures very quickly so as to be able to take corrective actions and ensure the safety and reliability of the vehicle.

1.3 Customer satisfaction

The third area that has seen a growth in diagnostcs is related to customer satisfaction. Even in subsystems where diagnostic requirements are not legislated, nor are they mandated by the presence safety-critical functions, there may be some significant advantages in having diagnostic algorithms on-board the vehicle for the purpose of guaranteeing customer satisfaction and overall quality. The use of diagnostic algorithms to reduce false positives that may lead to significant warranty costs for manufacturers has been a subject of interest to automotive manufacturers, and also to subsystem suppliers, who may incorporate diagnostics directly into sensors or actuators. Accurate diagnosis can reduce the incidence of faulty components or the incidence of components being replaced when they are in fact still good (i.e., false positives). This is an issue of particular concern in a vehicle that has significant electronics content because it is often too easy, in the face of a perceived malfunction, to replace expensive components, such as for example an electronic control unit, rather than pinpoint the specific cause. Studies conducted in the industry have shown that the percentage of false positives, for example as they pertain to replacement of the engine control unit, is as high as 80% of the cases ([MPY+06]).

2. PROBLEMS AND CHALLENGES

In the face of the different requirements outlined in the preceding section, there is growing interest on the part of the automotive industry in the ability to systematically design diagnostic algorithms. Further, automakers have also shown a desire to extend warranty periods to provide consumers with a worry-free experience. As a consequence, in addition to on-board diagnosis of different functions, the prognosis of various functions and subsystems in the vehicle has also become important. Manufacturers would like to be able to predict when maintenance or replacement may be needed for specific components, for example the 12V battery, or components in subsystems related to the emissions control system. So, prognosis is beginning to take on a role in automotive electronic systems that was not on the horizon even just five or ten years ago.

The implementation of diagnostic and prognostic algorithms of this type in automotive systems presents a number of challenges due to the scale of the implementation. Such algorithms must be adaptable to millions of vehicles and must be robust enough to be valid over a broad range of different realizations of the same vehicle platform, with choice of different engines, transmissions, and accessories. Further, vehicles that might be architecturally identical, will unavoidably require different software calibrations in different markets. Thus, the design and implementation of OBD algorithms is not a 'one size fits all' kind of design approach.

The second issue is related to the fact that automotive systems tend to be complex and highly nonlinear. For example, engine and exhaust emissions processes are characterized by complex thermochemical behavior (combustion processes, exhaust emissions formation), that is strongly affected by chemical reaction kinetics, fluid motion and heat transfer. Further, the presence of sensors and actuators, such as fuel injectors, or systems that could be pneumatically or hydraulically actuated increases the overall complexity of an engine emission control system. Therefore, it is difficult to imagine that simple, linear algorithms could be very effective unless a substantial amount of thinking and a deep understanding of the physics of the processes goes into their design.

Another important aspect is the speed of execution, in the face of limited computational capabilities (both CPU speed and memory). On-board computers used in automotive applications have relatively low power relative to the number of functions that they perform, because cost is a significant constraint in the automotive industry. So, one of the main challenges is to develop effective diagnosis algorithms that can be implemented in fixed-point arithmetic microcontrollers with limited amount of memory and limited CPU speed. Some algorithms may require truly real-time implementation. For instance, in safety-critical diagnosis algorithms (e.g.: vehicle stability control, or brake-by wire or steer-by wire applications), one

is obviously concerned with the implementation of these algorithms in real-time so that any fault that is detected can be compensated for in a fault-tolerant control scheme or by entering a limp-home mode as safely and as quickly as possible. On the other hand, other types of algorithms, such as those that may be used to diagnose malfunctions in the emission control systems, may not have such stringent real-time requirements, in the sense that on-board diagnostics regulations typically require that the diagnosis be carried out within what is called a 'one trip'.

Finally, diagnosis must be as transparent as possible to the user, while the designer must be very cognizant of the relative weight of false alarms vs. missed detections: such weighting will vary depending on the application, with missed detections being especially costly (to the user) in safety-critical applications, while false alarms can be very costly (to the manufacturer and consumer alike) when non-safety-critical applications that have warranty implications are considered.

In short, the subject of system diagnosis in the most complex consumer device in existence today - the automobile - is one that presents numerous technical challenges that range form the theory of estimation and detection, to real-time software implementation issues. We hope that the reader will come to appreciate the true complexity of the problem after reading the overview presented in the following sections.

3. APPROACHES TO FAULT DIAGNOSIS

Generally speaking, one can characterize approaches to fault diagnosis as

- data-based: based on data and employing system identification techniques to identify models;
- physics model-based, that is, based on a model that has some predictive quality based on physical first principles.

In reality, no approach is completely physics-model-based or completely data-based, because every approach that uses data typically has some kind of an underlying model, and every physics-model-based approach requires a certain amount of empirical calibration, so that in truth every approach we use is really a combination of data-based and physics-model-based techniques. It is our opinion that, whenever possible, models based on the physics of the process should be used. What constitutes a model is subject to interpretation, of course. A model is, in general, intended to be a collection of differential and/or algebraic equations, but can take many forms. The key idea is that one should taking into account a physical understanding of the system at the very onset. It is also important to understand that when one talks about model-based approaches in diagnostics, these may require models that have greater fidelity than the models that are required to develop the control algorithms. This is one of the fundamental challenges in diagnostics: in order to be able to detect minor differences between the normal operation of a system and its faulty operation, one needs to have reasonably sophisticated models, or a very solid understanding of the physical principles that underlie the processes.

The ultimate accuracy of diagnostic algorithms is dependent on accuracy of the model we use to predict the behavior of the system. Whether these models are data-based or physics-based, whether we obtain them through system identification or whether they are grey-box models in which we parameterize a physically-based model with empirical coefficients, it is always true that any diagnostic algorithm is fundamentally limited in its robustness by modeling capability. Modeling errors will unavoidably arises due to the use of a simplified model to represent a more complex phenomenon, resulting in unmodeled or inaccurately modeled dynamics, for example, whenever we approximate a distributed-parameter process using a lumped-parameter model. A second and equally important

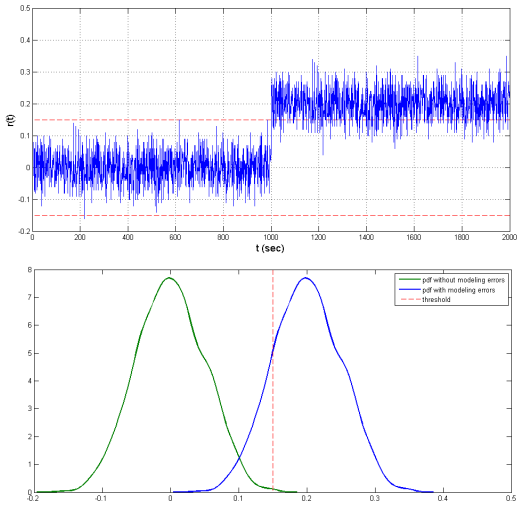

Fig. 1. Effect of the modeling errors on residual (above), and on the probability density function (below)

source of model uncertainty in automotive systems is due to the fact there are unavoidable plant differences even in nominally identical realizations of the same system or subsystem due to production variability across tens or hundreds of thousands of vehicle units. This fact is clearly recognized by the automotive industry, and is the reason for the development of sophisticated calibration procedures. Work has been done in characterizing this different types of model uncertainty in fault diagnosis, but this remains one of the great challenges in the diagnosis of automotive systems.

4. APPLICATIONS - EMISSION CONTROL

4.1 On-Board Diagnostics (OBD)

To combat the smog problem in the LA basin, the State of California started requiring emission control systems in 1966 model-year passenger vehicles. The federal government extended these controls nationwide in 1968. In 1970 the Environmental Protection Agency (EPA) was established. This started a series of graduated emission standards and requirements for maintenance of vehicles. To meet these standards, manufacturers turned to electronically controlled fuel feed and ignition systems. In 1988, the Society of Automotive Engineers (SAE) set a standard connector plug and set of diagnostic test signals.

The EPA adapted the standards from the SAE on-board diagnostic programs [EPA]. OBD-II is developed by SAE and adopted by the EPA and CARB (California Air Resources Board).

EPA: OBD-I and OBD-II The first-generation OBD-I requirements (1988) were relatively simple as compared to today's requirements:

- Emission-related inputs to the ECU were required to be monitored for opens and shorts;
- The components requiring performance monitoring included the ECU, fuel metering system, ignition and EGR system (if present).

Since 1994, OBD-II regulations have been imposed in the U.S.A. on gasoline, Diesel and alternative fuel vehicles. A distinction is made between passenger cars and light-duty trucks on one side, and heavy-duty vehicles on the other.

The European Union (EU) has developed a set of regulations for onboard diagnostics for emissions controls that is similar to those defined for the U.S.A. It is expected that common standards will be used to define the EU OBD requirements such that engines developed for the U.S. and Europe will comply for OBD in both domains. The European On-Board Diagnostics (EOBD) requirements are similar to the EPA OBD requirements, with the exception that higher malfunction thresholds are tolerated. In practice, most manufacturers have simply utilized US OBD-II software/strategy to meet the EOBD requirements.

Any component that directly or indirectly affects emissions, for example, coolant or intake manifold temperature sensors must be monitored. If a malfunction is detected that would cause emissions to exceed the regulated standard by more than 50%, a Malfunction Indicator Light (MIL) must be lit on the dash-panel to warn the operator that repair is required.

In addition, the OBD-II regulations require monitoring of comprehensive components, such as any electronic powertrain component/system which either provides input to, or receives commands from the on-board computer for: malfunctions which can affect emissions during any reasonable in-use driving condition, or electronic powertrain components/systems used as part of the diagnostic strategy for any other monitored system or component. This stringent requirement is intended to ensure that anything that can affect emissions, even to very small degree, is monitored.

In the next three sections we illustrate three diagnostic applications motivated by OBD requirements, chosen to place emphasis on three distinct diagnostic methodologies:

- Model-based fault diagnosis of a NO_x aftertreatment system. In ([PCO$^+$09], [PSYJ06]) an open-loop approach based on parity equations is used;
- fault-tolerant powertrain control achieved by integrating control and diagnostics, in which [KRU98], [KRU01a] use nonlinear (sliding-mode) observers;
- and, finally a signal-processing-based engine misfire detection [KRSW95] algorithm.

4.2 NO_x after treatment system

A fundamental application of FDI in the automotive field consists in monitoring the after-treatment systems in Diesel engines. In [PCS] and [PCO$^+$09] a scheme based on a parity equation approach is proposed and analyzed. One of the most important factors for permitting the use of Diesel engines according to the OBD regulations is the control of the engine emissions, in particular particulate matter (PM) and nitrogen oxides (NO_x). A methodology for NO_x emission reduction called Lean NOx Trap (LNT) [BCSB98] (see Fig. 2), is taken into account here. This strategy has the advantage of not requiring reductant supply aboard, is effective within a broad range of temperatures ($250 - 450°C$), and has a high NO_x conversion efficiency (more than 90%).

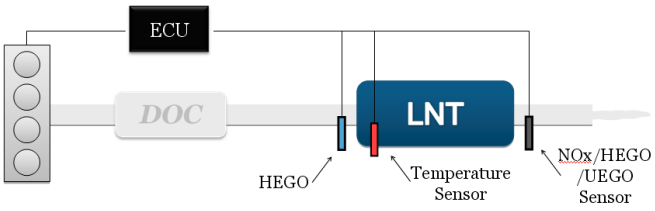

Fig. 2. Schematic of typical LNT aftertreatment system

The main challenge for NO_x after treatment systems implementation on the vehicle is that a robust control system must be implemented, in order to ensure that the after treatment device operates with high conversion efficiency, regardless of variability in the operating conditions, and with restrictions on the available sensors. Moreover, dedicated algorithms are required to monitor the aftertreatment system for faults, in compliance with the recent OBD-II regulations. In [PCS] and [PCO$^+$09], the described strategy permits to detect faults regarding sensor malfunctioning, LNT sulfur poisoning and thermal deactivation in the LNT scheme. The model is obtained with a grey-box approach, where the equations regarding the conservation of mass and energy are used, together with simplified stoichiometric reactions for mass balances. It is assumed that the mixture is at thermal and chemical equilibrium (quasi-steady conditions), and that NOx are only present as NO. A high level scheme of the model can be seen in Fig. 3.

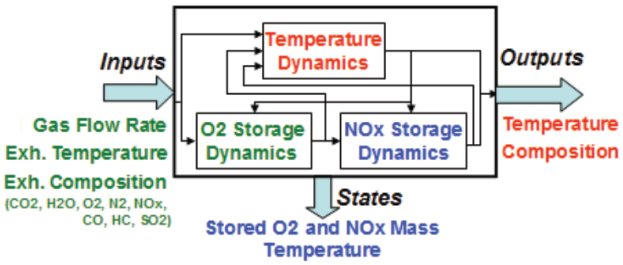

Fig. 3. Schematic of the model used for the NO_x after treatment diagnostics

The fault diagnosis method used to detect the LNT aftertreatment system faults with steady-state input conditions, is based on parity equations with constant thresholds. Fig. 4 represents the whole diagnostics scheme: the

blocks in the upper part, from left to right, represent the Diesel engine with air-to-fuel ratio (AFR) control, the Diesel Oxidation Catalyst (DOC) and the LNT. The block in the left-lower part represents the model of the engine that is run in the diagnostics scheme, in order to generate the output in case of non-faulty conditions: the difference between the outputs of this system and those generated by the 'truth model' implemented in the upper part of the scheme are the residuals used for fault detection and isolation. Please note that the implementation of the parity equations is achieved in this case through a computational model, not an analytical one, and that the computational model, however simplified, is still quite complex.

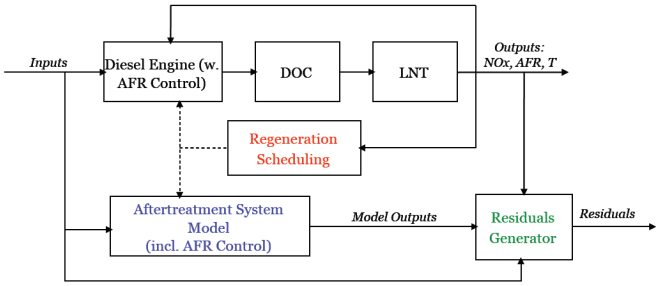

Fig. 4. Schematic of the FDI system in the NO_x after treatment application

The method described above yields promising results; for instance, Fig. 5 shows the results obtained in detecting a fault in the temperature sensor in [PCO+09], These results were obtained in simulation, using a validated model of the process. It should be remarked that today, the ability to run models of such complexity in real time in an onborad processor is probably still beyond the capabilities of production microcontroller hardware. Nonetheless, this approach shows what can be accomplished if a relatively high-fidelity model that can be computed in real time is available.

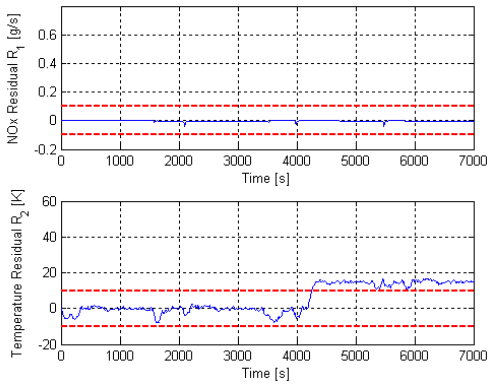

Fig. 5. Residuals for temperature sensor fault diagnosis in the NOx diagnostics scheme

4.3 Fault tolerant powertrain control

The second application that pertains to the exhaust emissions regulations area is the one described in ([KRU01a]) and [KRU98], where the integration of control and diagnostics with the objective of achieving fault tolerant control is considered. The references cited describe a strategy that incorporates detection and isolation, and guarantees desired system stability and performance in the presence of certain component malfunctions.

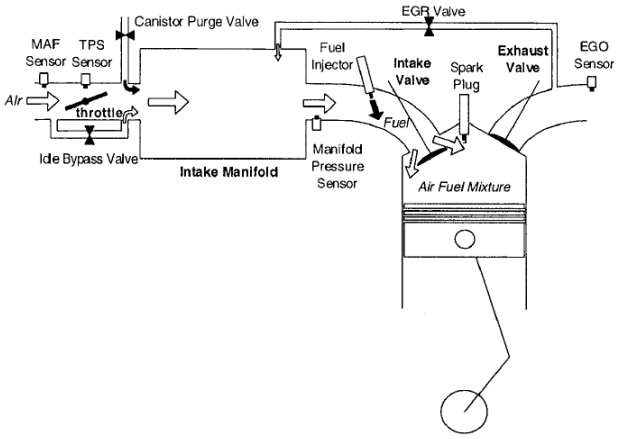

Fig. 6. Air-fuel related subsystem of an IC engine.

The system under consideration considers the breathing and fueling dynamics of a port fuel-injected 4.6L spark ignition (SI) engine, shown in Figure 6. This type of model is generally referred to as a "mean value" model. The elements of the system include flow through the throttle, filling and emptying of the intake manifold, fuel injection and fuel film dynamics in the manifold runners, exhaust gas recirculation (EGR), system delays including the induction-to-power stroke delay inherent in the four-stroke engine cycle and transport delay in the exhaust manifold, and sensor dynamics.

The objective of the work is two-fold: first, it is required to identify a fault in the proportional oxygen sensor (UEGO), and, second, the air-fuel-ratio (AFR) controller must be capable of reacting to a fault maintaining good control performance, that is, keeping the AFR as close as possible to the reference value. To solve both these problems, the control strategy is designed as follows: a feed-forward term that uses cylinder air charge estimation and fuel film estimation provides the base value for the fuel to be injected; feedback comes in the form of a proportional term (using the UEGO sensor) and of and Integral Sliding Mode term. The latter consists of an additive discontinuous control term that compensates for possible actuator faults. A sliding mode observer is used to track the value of the output, in order to avoid that a fault on the sensor could lead to decreased performance of the control strategy. Fig. 7 shows a schematic of the fault tolerant control strategy: the observer structure takes into account the possible different faults.

The strategy was implemented first in simulation, then on an engine; Fig. 8 shows how the proposed control methodology reacts to an input fault (in simulation). After a short time, the fault is identified, while the Integral Sliding Mode strategy permits to react very quickly to it keeping the AFR around the reference value.

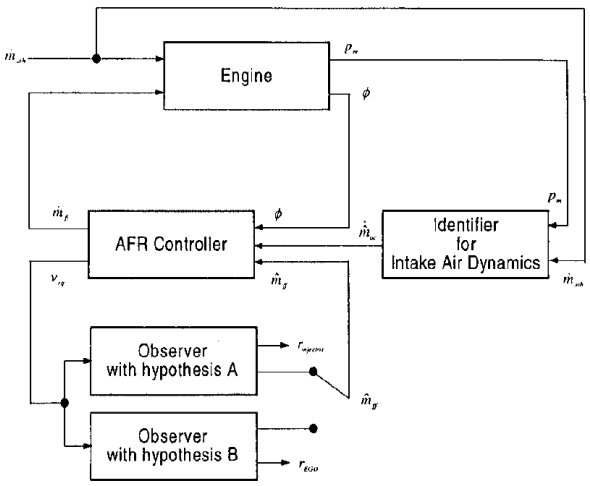

Fig. 7. Schematic of control and diagnostics for engine air and fuel management system.

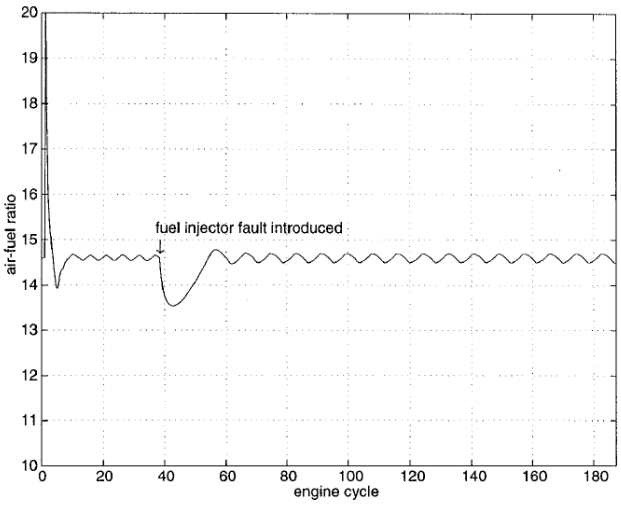

Fig. 8. Air-fuel ratio response with fuel injector fault in the fault tolerant control scheme

4.4 Engine misfire detection

In this section we summarize a family of approaches that use signal processing and statistical signal classification to detect and isolate individual occurrences of engine misfire. Misfire is defined as as a lack of combustion in the cylinder due to absence of spark, poor fuel metering, poor compression, or any other cause. This does not include lack of combustion events in non-active cylinders due to default fuel shut-off or cylinder deactivation strategies [obd05]. According to the OBD-II regulations, even a single instance of misfire in the engine has to be recorded, while repeated instances have to be reported for maintenance. Automotive engines, as with any rotating machinery, present signature torsional vibrations, and any changes in this signature can be used to detect faults in the engine. Angular velocity measurement have been used to detect changes in torsional vibration characteristics of engines [Riz89],[GW90],[CR94],[PGG+95],[KRSW95].

The diagnosis of abnormalities in the torsional vibration of the engine crankshaft is more readily apparent when the variables of interest are viewed with respect to the angle of rotation of the shaft. These variables are periodic when measured in the domain of angle of rotation, and it does not matter if angle of rotation is a non-linear function of time [CR94]. Among the various methods used to analyze the angular velocity data, there are: spectral analysis using discrete fourier transform, principal components analysis, discrete wavelet transform and change detection [KRSW95].

When torsional vibrations are measured in the angular domain, they are quasi-periodic with period equal to one engine cycle (two revolutions for a four stroke engine). Discrete Fourier Transform (DFT) based approach are especially appropriate because of the periodic nature of the signal. If we define the window to be 4π radians, then the engine firing frequency will be N times per period, where N is the number of cylinders in the engine. Assuming that each cylinder produces identical torque, the spectral content of the signal will be at the firing frequency of the engine, and at higher harmonics. Any spectral content at lower frequencies will indicate a significant non-uniformity in engine torque production, and is an excellent indicator of misfiring conditions. Fig. 9 shows the case of the amplitude spectrum of the angular velocity signal for the normal and misfire case for a 12-cylinder engine [PGG+95].

The fact that a misfire is readily visible in the order of rotation spectrum, as shown in Fig. 9, does not make the isolation task automatic. Isolation of the misfiring cylinder can actually be a complex process that can require sophisticated pattern recognition methods. In the especially challenging case, that of a 12-cylinder engine, [Riz89] shows that using a combination of Principal Components Analysis (PCA) and statistical clustering it is possible to correctly detect and isolate single misfires even in a V-12 high-performance engine in which combustion events have significant overlap with one another, rendering the isolation of the misfiring cylinder quite difficult.

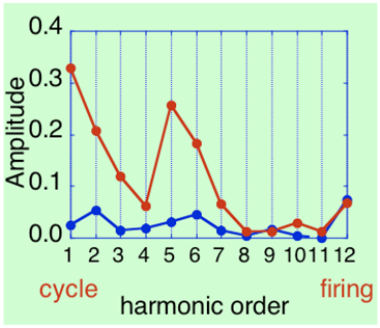

Fig. 9. Amplitude spectrum of the angular velocity signal for the normal and misfire case for a 12-cylinder engine

The PCA method is a well-known multi-variate statistical technique that permits analysis of complex data by providing the ability to compress large data sets into a lower dimensional representation. The data compression characteristics of PCA also make it suitable for the ex-

traction of salient features from the data, and facilitate the solution of pattern recognition ad signature analysis problems. In this example, PCA is used in conjunction with clustering techniques to isolate the location of the misfiring cylinder. The real and imaginary parts of the first twelve frequency components of the engine angular velocity signal (excluding the DC component), form the initial data vector for PCA. As explained earlier, these frequency components are computed on a cycle-by-cycle basis. Let a sample data matrix be

$$X = x_{ij} = x_j \tag{1}$$

where $i = 1, \ldots, I$ is the engine cycle index and $j = 1, ..., 24$, is the index of the real and imaginary parts of the twelve frequency components. Note that $I >> 24$, and that x_i is the row vector corresponding to the ith engine cycle. The matrix X is typically generated by acquiring data at various engine speeds and loads, under nominally constant speed and load conditions. The x_i vectors can then be represented in 24-dimensional space.

Once a normalized data matrix exists, it becomes possible to define a new orthonormal basis, $u_1, u_2, ..., u_J$. The u_1 axis is determined in such a way that the square distances between vectors z_i and the u_1 axis are minimized. This is analogous to maximizing the square projections of the vectors z_i on u_1. Summing up these square projections, the variance of the random vectors z_i is obtained. The u_2 axis is found following the same procedure with the supplementary constraint of its orthogonality with u_1. The u_3 axis is defined in the same way adding the orthogonality to u_1 and u_2 and so on.

The projection of the i-th cycle vectors onto the new basis is calculated by means of the transformation matrix U: $F = f_{ij} = f_i = ZU$ where F is a $(I \times 24)$ matrix, called the factor matrix. The $f_j = f_{ij}$ columns (the Principal Components) contain the projections of all the I points onto the u_j axis. The next step is to order the J columns f_i, according to the decreasing value of the variance on the u_j axis. Taking into account only the most significant Principal Components, the computing time is reduced, without a substantial loss of information. The detection and isolation of a misfire is then based on the calculation of the distance between the coordinates of the z^* vector and the coordinates of the pre-computed centers of gravity of each of the clusters corresponding to each of the twelve possible misfire conditions. Figure 10 depicts such clustering in three-dimensional space simply for visualization purposes showing clusters of misfiring and normal data.

5. APPLICATION - SAFETY

5.1 Introduction

The first results in electronically controlled vehicle chassis systems can be found in [WI95], where a diagnostic system for the lateral vehicle motion is designed using a discrete parity space approach. In [KR95a] and [KR95b] a solution to the problem of diagnosing faults in a vehicle steering system is proposed using nonlinear observers constructed by sliding mode design techniques. In [ISS00], after a description of a drive-by-wire system and of the possible sensor, actuator and component faults, an application

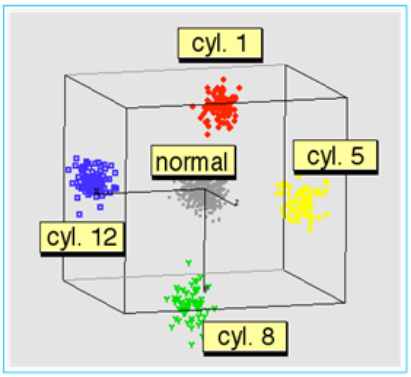

Fig. 10. Example of clustering in three-dimensional space

of a simple fault detection and isolation scheme for a electromechanical brake pedal in a brake-by-wire system is presented. Three residuals are obtained by using a parity equation method and by comparing the signal from one sensor with the reconstructed value from another sensor obtained by an analytical pedal model. In [Ise00], the fault detection and isolation approach is applied to vehicle suspensions and to a simplified bicycle model for the steering system. Some experimental results are presented for the vehicle suspension related to the tire stiffness estimation. For the bicycle model a fault detection scheme based on the parity space approach and neural networks is used to classify the different faults by training it with special patterns. For an overview on fault-tolerant drive-by wire system one can refer to [ISS02], while recent results in FDI for lateral and vertical vehicle dynamics are presented in [FBSI07]. Fault diagnosis for engine and powertrain systems, and vehicle dynamics and control, has been studied for instance in [KRU01b], [DAR99], [PSR03].

During recent years, the applications of FDI strategies for improving safety in the vehicle have received increasing attention.

In the following, a hierarchical approach proposed in [Pis02] in drive-by-wire system is presented which permits the use of simple models for fault diagnosis.

5.2 Hierarchical FID in a drive-by-wire system

The general structure for the Hierarchical FDI is based on a hierarchical decomposition of a system, as represented in Fig. 12. The basic idea is to view the system as an interconnection of lower dimensional subsystems determined according to a certain partitioning methodology. For the decomposed system it is possible to create a FDI scheme comprised of many 'Residual Generation Units', and each of them outputs a residual that is sent to a 'Residual Evaluation Unit' that performs the residual evaluation for the selected subsystem. An example of hierarchical decomposition scheme for a vehicle chassis system is depicted in Fig. 11.

The methodology combines enhanced version of model-based fault diagnosis with qualitative techniques and models. The principle is based on the fact that system failures occur in two stages: failure sources where the fault originated, and failure propagation of these faults to other units. So, a process for failure analysis must try to locate

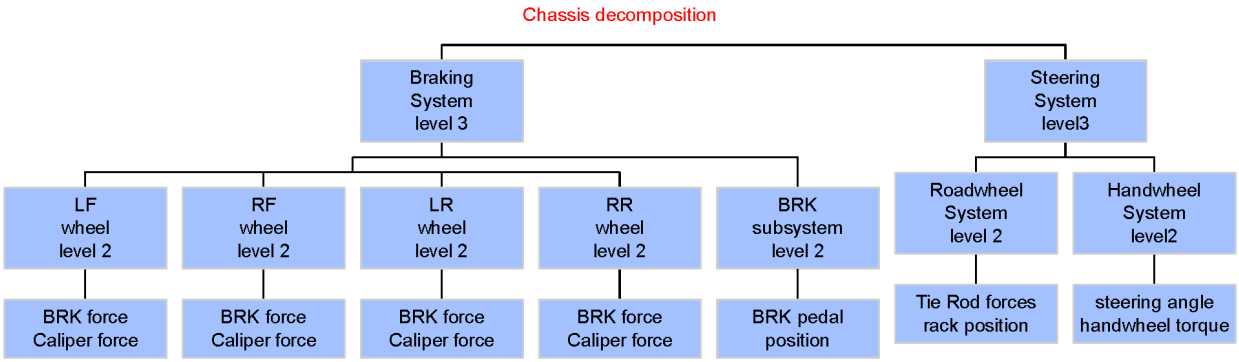

Fig. 11. Generalized structure decomposition for hierarchical FDI (upper part), vehicle X-by-wire system (lower part)

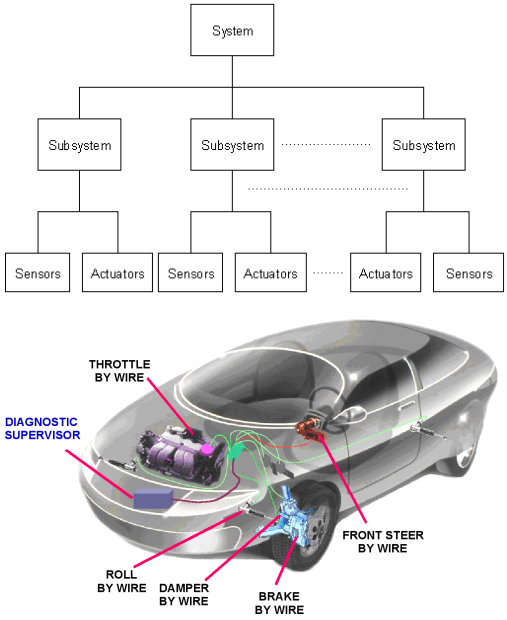

Fig. 12. Generalized structure decomposition for hierarchical FDI (upper part), vehicle X-by-wire system (lower part)

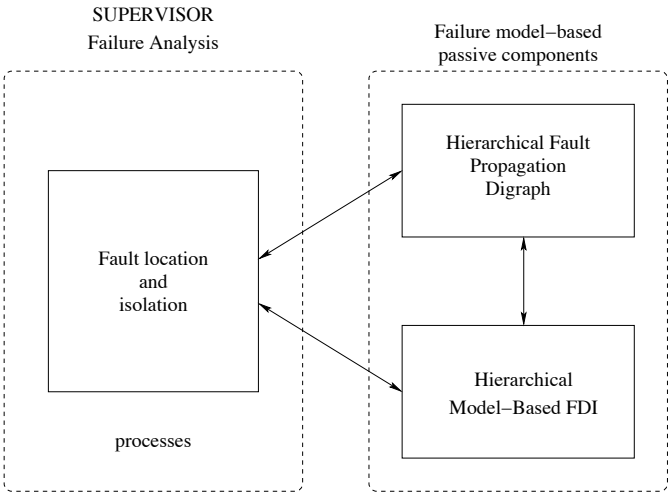

Fig. 13. A framework for hierarchical FDI

the failure sources, and isolate the faults cause of failure, according to the scheme in Fig. 13. The framework is structurally divided into two components. The model-based passive component represents the knowledge about the fault behavior of the system under analysis, and consists of two different models. The failure analysis instead is an active constituent comprising processes that use the model-based knowledge for diagnosis. One model for the passive component is the Hierarchical Model-Based FDI previously described, while the second model is constituted by a Hierarchical Fault Propagation Digraph.

The fault propagation digraph contains a hierarchical representation of available knowledge about the characteristics of fault propagation within the system. Each level in the hierarchy contains one or more structures that together represent a view of the system under a particular granularity. The granularity of view increases with levels. Thus as one traverses down the levels of the model, the resolution of view increases. The Fault Location and Isolation Process (FLIP) operates on individual structures of the hierarchical fault propagation model.

FLIP starts at the top level of the model constituted by a single structure containing a single element that represents the system under analysis. Once the FLIP returns the failure sources in all structures at one level, the process migrates to the next level. The FLIP interacts also with the hierarchical model-based FDI model by activating the available residual generator units to reduce the possible fault locations and, at the end, to perform the final isolation. When a FLIP exits from the lowest level, the failure source set will contain the set of faults that are the source of failure.

A scheme representing the application of such methodology to a chassis system is depicted in Fig. 14.

For the same application, i.e. the drive-by-wire system, adaptive thresholds are used for fault detection in sensors when a steer-by-wire strategy is used, as described in [PSYJ06], where it is also shown how modeling errors can be drastically reduced by this technique.

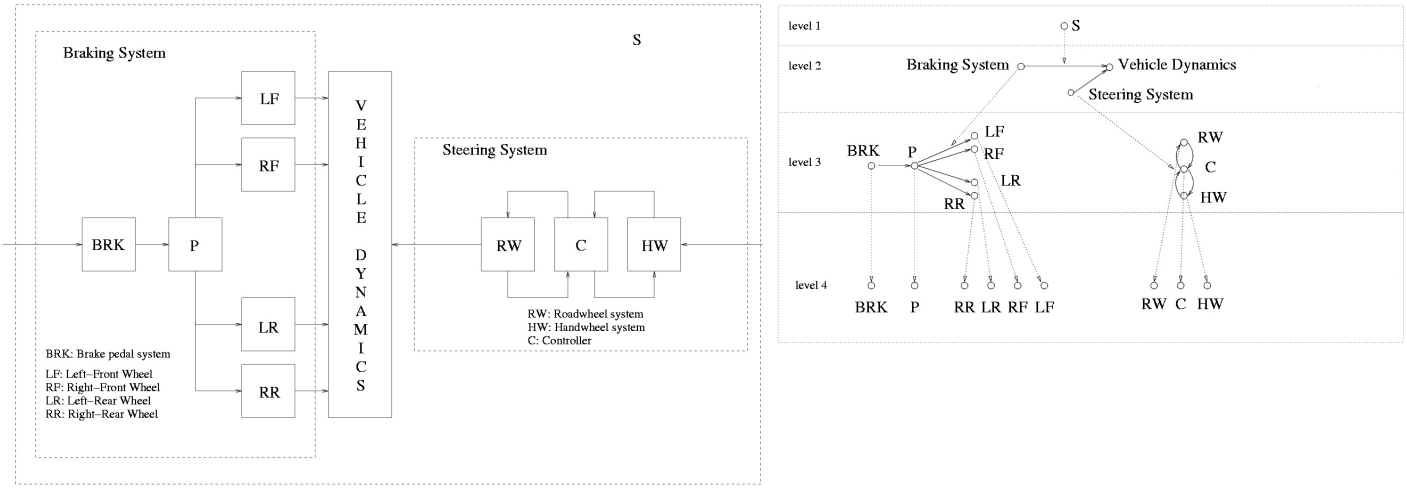

Fig. 14. An example of hierarchical propagation model for a chassis system

6. APPLICATION - CONSUMER SATISFACTION

6.1 Introduction

The demand for electric power systems in automobiles has increased substantially in the last decade due to the addition of many electrical systems and electronic devices needed to comply with regulations and meet customer needs. As a consequence of this dramatic growth in power requirements, and in order to maintain optimal vehicle performance, attention must be given to deriving more effective fault diagnosis algorithms for electrical systems. Vehicle diagnosis has become an important component in a vehicle operation both for safety and consumer satisfaction reasons. As vehicles have become equipped with more and more complex electronic systems and sub-systems, it has also become more challenging to identify the defective part of the system, whether it is the ECU, sensor, actuator, wiring harness, etc. ([TH02]). This is mainly due to the following reasons. First of all, because different models of vehicles generations have versatile architecture and features, they can exhibit distinct fault behavior and it can be sometimes difficult to uniquely associate a malfunction feature to a specific component. Moreover, the automotive industry has experienced an increased of both "no defect found" problems, not reproducible, and "intermittent defects", that exhibits a minimal degree of repeatability but are hard to locate. That leads to an intrinsic difficulty of accurately identifying faults on vehicles. This challenging problem has led both the academic and the industry world to put a lot of efforts in it [KN05], [AM02], [BI03].

The next section focuses on the description of development of fault diagnosis systems for the automotive Electric Power Generation and Storage (EPGS) system. It will be extremely helpful to give an early warning to the user when the EPGS system leaves its safe operating area due to whatever reasons. Meanwhile, such a capability will also improve resource management via condition-based maintenance, and minimize the operational costs for automotive dealers. For such a system two different fault diagnosis approaches have been proposed:

- model-based approach ([**?**], [LCS+08b]), which exploits a analytical model of the system for designing a parity-equation based fault diagnosis scheme, and
- a hierarchical approach, presented in [LCS+08a], that allows to go from more general information of the system malfunctions to a detailed knowledge of the fault, hence achieving fault isolation.

6.2 Electric Power Generation and Storage system (EPGS)

The EPGS system, shown in Fig. 15, is composed of a belt, an alternator with rectifier, a voltage regulator, a battery and several electrical loads. When the engine is

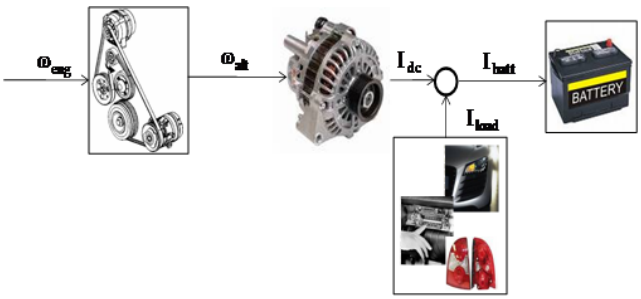

Fig. 15. EPGS system

running, the alternator, driven by the engine through a belt, supplies power to the loads and charges a 12-V lead-acid battery. The battery provides the high power needed by the engine starter motor, and supplies power when the engine is not running or when the demand for electrical power exceeds the output power of the alternator [Bos03]. The diagnostic problem focuses on the detection and isolation of a specific set of alternator faults, including belt slipping, rectifier fault and voltage regulator fault.

The faults aimed to be diagnosed are:

- Open diode fault: a diode of the passive rectifier is open.
- Regulator electronic circuit fault: the electronic circuit of the regulator can break.
- Belt slip fault: the belt can break or can have a significant slip.

The first three faults are different failure modes of the alternator, whereas the last two are faults related to the belt. The two approaches presented have been both tested in simulation and implemented on the EPGS experimental test bench, (Fig. 16) developed at the Center for Automotive Research, OSU.

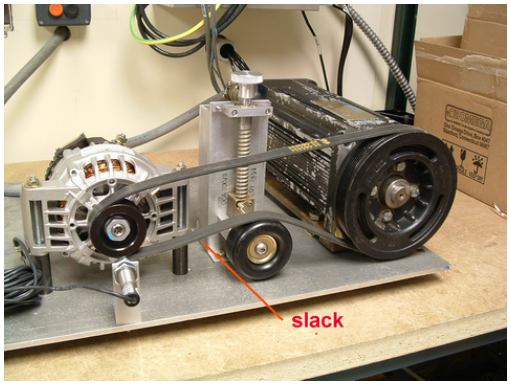

Fig. 16. EPGS test bench

6.3 Model-based fault diagnosis for the EPGS system

Model-based fault detection and isolation (FDI) is based on the ability to construct residual generators based on models of the system (for example, through the design of state observers or parity equations). Unfortunately, due to the highly nonlinear behavior of the components of the system, the complexity of the EPGS system is significant. For the alternator system, the combination of the nonlinear dynamics of the three-phase generator with the switched, state-dependent behavior of the diode bridge rectifier, make the design of a model-based fault diagnosis system very challenging. Linearization is, for example, virtually impossible in the presence of the hard nonlinearities present in the rectifier. Meanwhile, a direct non-linear parity equation or observer design for such a complex non-linear switch system will also be extremely difficult.

In order to obtain a robust diagnostic algorithm, and in light of its implementation in the vehicle, the approach in [LCS+08b] utilizes an equivalent alternator model based on its input-output relationship ([SRP07]). This allows for the identification of an equivalent DC generator model of the alternator by the replacement of the AC synchronous generator and diode bridge rectifier with an equivalent DC generator, as shown in the scheme of Fig.17. The designed fault diagnosis algorithm uses a parity equation approach based on the equivalent model and compare the behavior of the alternator with the behavior of the equivalent model to produce the residuals that contain the information of the faults, according to the scheme in Fig. 18.

The components of the EPGS system have a high nonlinear behavior which can be hard to model. For the alternator system, the combination of the nonlinear dynamics of the three-phase generator with the switched, state-dependent behavior of the diode bridge rectifier, make the design of a model-based fault diagnosis system very challenging. Linearization is, for example, virtually impossible in the presence of the hard nonlinearities present in the

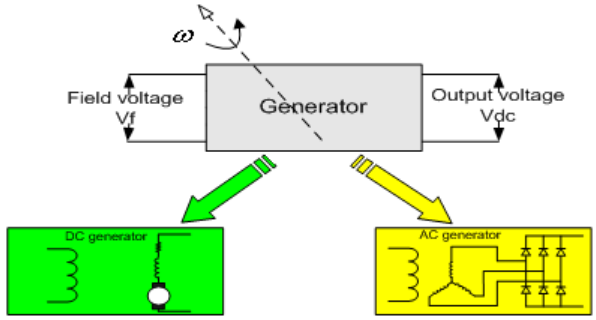

Fig. 17. Input-output perspective of an alternator

rectifier. Meanwhile, a direct non-linear parity equation or observer design for such a complex non-linear switch system will also be extremely difficult. It is important

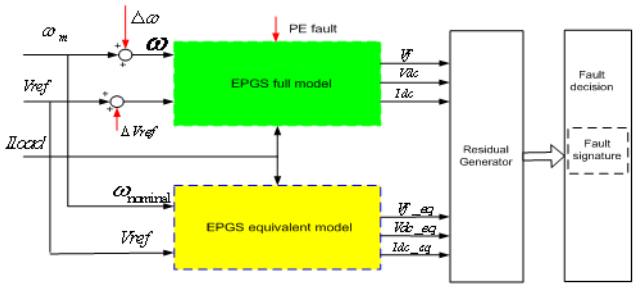

Fig. 18. Model-based fault diagnosis scheme

to notice that the equivalent model is used as an open loop estimator for the full model without fault. Another important part of FDI design is the residual processing and threshold calibration. In fact, because of model inaccuracy, disturbance or measurement noise, conditions for perfectly robust residual generation cannot be met in practice. The threshold calibration has been conducted by a statistical approach, with the aim of reducing the probability of false detection and miss detection. In Fig. 19, for example, it is shown the behavior of the residual $r_2 = V_f - V_{f_{eq}}$, in the case of no fault, belt fault and diode fault condition.

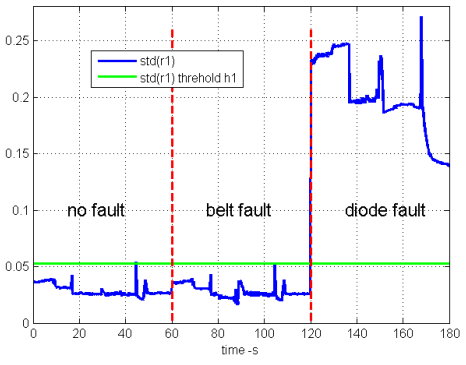

Fig. 19. Experimental results

6.4 Hierarchical fault diagnosis for the EPGS system

The main drawback of the model-based FDI solution presented in the previous section is that it requires the measurements of both the battery current and alternator

current, whereas only the alternator current is available on-board. A different approach of fault diagnosis is then pursued in [LCS+08a]. In this work, the frequency content of the measured signals is deeply analyzed and exploited to design the hierarchical FDI algorithm. Specifically, from an accurate analysis of the signals, one can extract crucial information about the symptoms of the faults occurring at the system. For instance, it is known that, if an open diode fault occurs, the ripple amplitude increases.

Hierarchical diagnostic strategy, as implemented in [LCS+08a] refers to a top-down methodology that uses a-priori knowledge of the system signal behavior in order to detect possible system malfunctions. This allows to go from more general information of the system malfunctions to a detailed knowledge of the fault, hence achieving fault isolation. The main idea is that, starting from a high level analysis of the signals, the occurrence of a possible fault is detected and isolated by analyzing the frequency content of the signals. An important advantage of such a hierarchical approach is that the computation load is greatly reduced when compared to a model-based FDI algorithm. Figure 20 depicts the diagnostic logic which runs when the alternator is in operation. The diagnosis logic has a hierarchical structure, composed of three levels The first level analyzes both the battery current signal I_{batt}, by checking amplitude and frequency of its ripple, and the V_{dc} signal by monitoring its mean value. If an anomaly is detected, with respect to the nominal condition, the algorithm activates the lower levels to isolate the fault. This algorithm has been also experimentally validated on

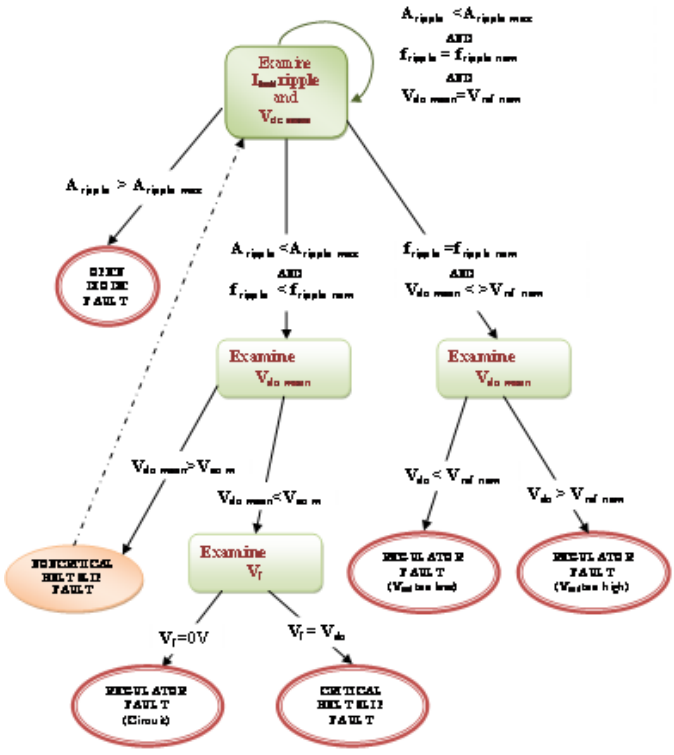

Fig. 20. Top-down hierarchical scheme

the EPGS test bench (Fig. 16) and it turns out to be suitable for on-board implementation.

7. FUTURE CHALLENGES

7.1 Hybrid Vehicles

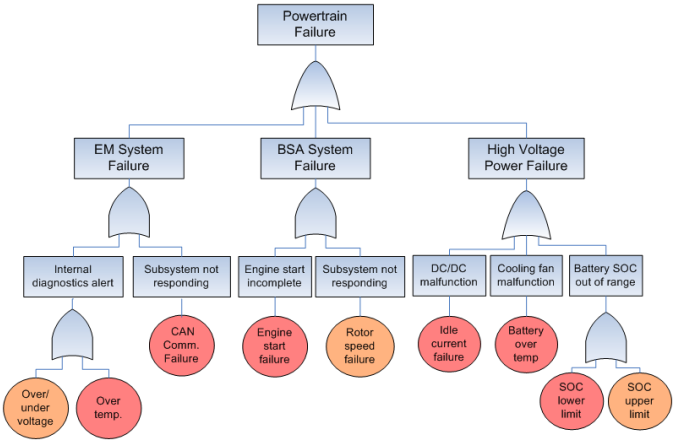

Fig. 21. Fault-tree diagram of the electric powertrain

7.2 Prognosis

During the last ten years, many researchers from different engineering communities (reliability, fault detection and diagnosis, mechanical fatigue analysis, identification and control of linear and nonlinear dynamical systems) started focusing on understanding how a complex electromechanical system can age, and to predict how much time remains before the life of the system comes to the end, that is, to estimate the Remaining Useful Life (RUL). This emerging engineering discipline is usually referred to as Prognostics and Health Management (PHM). Health monitoring and prognostics of complex systems is a basic requirement for condition-based maintenance in many application domains where safety, reliability, and availability of the systems are considered mission critical. Unlike health monitoring technology, prognostics technology is still in its infancy although some research work on developing the technology has been done over the recent years. Most of the relevant work in the field of prognostics comes from the structural engineering community, where failure due to structural fatigue can lead to catastrophic consequences (e.g., in aeronautical and marine applications).

Like in diagnostics, prognosis methods can be divided in data-driven and model-based ([?]). With respect to diagnosis, in a data-driven approach for prognosis a more intensive data collection process is needed in order to characterize the damage accumulation and progression.

The main challenge is to analyze a multidimensional and noisy data stream from many of sources (use conditions, environmental conditions, and so on) in a population of similar components. It is important to say that the management of uncertainty is fundamental in RUL estimation. When the component is new and the accumulated damage is negligible, the uncertainty margins on the exact time of failure are very large. These margin of course become narrower as the component ages. Data-driven applications span a large number of techniques, from probabilistic ones [?] to neural networks [?]. Model-based approaches are useful to obtain more precise results, but of course their

design requires a deep knowledge of the system. First, it is necessary to identify and experimentally validate damage variables not always an easy and generalizable task, as it usually involves very lengthy experiments under controlled conditions, which do not necessarily reflect actual aging in real life. Second, once a damage variable is identified, there remains the challenge of reliably extracting features or estimating parameters from experimental data that closely correlate with the damage variable. Third, damage evolution is invariably a nonlinear phenomenon, making the modeling of it more difficult, and is also dependent on initial conditions (e.g. structural or material defect distribution). The applied methodologies comprehend, for instance, observer-based methods [LBP+03], [OKG+08], and mechanics-of-failure related strategies [CC04].

As for the automotive field, at the moment a lot of efforts are being put in the estimation of the RUL of batteries. A fundamental role is played by batteries, especially in electric and hybrid-electric vehicle applications, where estimating their life can eventually be crucial. Estimation of the state-of-charge (SOC) is indeed an important variable to tarck for battery prognosis. In fact, the SOC value can be useful to monitor the variation in the capacity of the battery, and consequently the battery calendar life.

A very good example of model-based SOC estimation can be found in [Ple04a], [Ple04b], and [Ple04c], where an Extended Kalman Filter is exploited.

A new prognostic methodology developed in ([SOGR09], [CSGR06], [SCGR05]) is presented in this section.

The model-based prognosis method in ([SOGR09]), introduces an analytical aging model to describe the system degradation for predicting the remaining life. The analytical aging model is based on experimental results collected on three different kinds of batteries: lead-acid (PbA), nickel-metal hydride (NiMH) and lithium-ion (Li-Ion) and on a suitable curve fitting. If the life of the battery is defined in terms of residual capacity, S, then the progression of the aging process can be expressed through the evolution of the normalized *damage measure* ξ, defined as,

$$\xi = \frac{S_0 - S}{S_0 - S_f} \tag{2}$$

where S_0 is the capacity of a new battery, and S_f the capacity of a battery at the end of its life, which is a predefined value. In fact, the end of life is usually defined as the moment in which the capacity becomes 80% of its original value. By definition, the damage measure ξ takes values between 0 (new battery) and 1 (end-of-life battery). Based on the Palmgren-Miner comulative mechanical fatigue model, a prognosis algorithm is formulated to predict the evolution of the damage, as shown, for example, in Fig. 22.

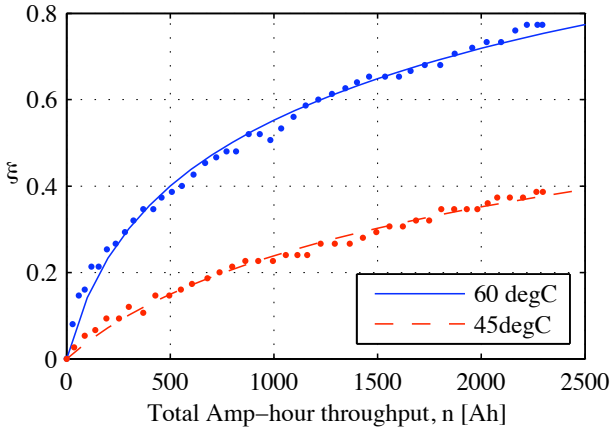

Fig. 22. Damage progression for Li-Ion batteries

8. CONCLUSIONS

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Marcello Canova, Dr. Yong-Wha Kim, Prof. P. Pisu, Delphi, GM, NSF.

REFERENCES

[AM02] S. Amberkar and B. Murray. Diagnostic strategies for advanced automotive systems. 2002.

[BCSB98] M.S. Brogan, AD Clark, MJ Spencer, and R.J. Brisley. Recent progress in nox trap technology. In *Proc. International Congress & Exposition*, Detroit, MI, February 1998.

[BI03] M. Borner and R. Isermann. Supervision, fault detection, and sensor fault tolerance of passenger cars. pages 327–334, 2003.

[Bos03] Bosch. Automotive electric and electronic systems. 2003.

[CC04] D. Chelidze and J.P. Cusumano. A dynamical systems approach to failure prognosis. *Journal of Vibration and Acoustics*, 126:2, 2004.

[CR94] F.T. Connolly and G. Rizzoni. Real time estimation of engine torque for the detection of engine misfires. *Journal of Dynamic Systems, Measurement, and Control*, 116:675, 1994.

[CSGR06] Z. Chehab, L. Serrao, Y. Guezennec, and G. Rizzoni. Aging characterization of nickel – metal hydride batteries using electrochemical impedance spectroscopy. *Proceedings of the 2006 ASME International Mechanical Engineering Congress and Exposition*, 2006.

[DAR99] L. Dinca, T. Aldemir, and G. Rizzoni. A model-based probabilistic approach for fault detection andidentification with application to the diagnosis of automotive engines. *IEEE Transactions on Automatic Control*, 44(11):2200–2205, 1999.

[EPA] Code of federal regulations (cfr [epa]), title 40, part 86, section 86.094-17 - control of air pollution from new motor vehicles and new motor vehicle engines; regulations requiring on-board diagnostic systems on 1994 and later model year light duty vehicles and light duty trucks (obd).

[FBSI07] D. Fischer, M. Borner, J. Schmitt, and R. Isermann. Fault detection for lateral and vertical vehicle dynamics. *Control Engineering Practice*, 15(3):315–324, 2007.

[GW90] Rizzoni G. and Ribbens W.B. Onboard diagnosis of engine misfire. In *SAE Transactions, Section 6*, volume 99, 1990.

[Ise00] R. Isermann. Mechatronic systems: concepts and applications. *Trans. of the Institute of Measurement and Control*, 22(1):29–45, 2000.

[ISS00] R. Isermann, R. Schwartz, and S. Stolzl. Fault-tolerant drive-by-wire systems - concepts and realizations. In *Proc. 4th IFAC Symposium on Fault Detection Supervision and Safety for Technical Processes*, Budapest, Hungary, June 2000.

[ISS02] R. Isermann, R. Schwarz, and S. Stolzl. Fault-tolerant drive-by-wire systems. *IEEE Control Systems Magazine*, 22(5):64–81, 2002.

[KN05] Line K. and Clements N. A systematic approach for developing prognostic algorithms on large complex systems. 2005.

[KR95a] V. Krishnaswami and G. Rizzoni. Model based health monitoring of vehicle steering system usingsliding mode observers. In *Proc. American Control Conference*, 1995.

[KR95b] V. Krishnaswami and G. Rizzoni. Sliding mode observer design with emphasis on vehicle dynamics applications. In *Proc. ASME International Mechanical Engineering Congress and Exposition*, San Francisco, CA, November 1995.

[KRSW95] YW Kim, G. Rizzoni, B. Samimy, and YY Wang. Analysis and processing of shaft angular velocity signals inrotating machinery for diagnostic applications. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 5, 1995.

[KRU98] YW Kim, G. Rizzoni, and V Utkin. Automotive engine diagnosis and control via nonlinear estimation. *Control Systems Magazine, IEEE*, 18:84–89, 1998.

[KRU01a] YW Kim, G. Rizzoni, and V Utkin. Developing a fault tolerant power-train control system by integrating design of control and diagnostics. *International Journal of Robust and Nonlinear Control*, 11:1095 – 1114, 2001.

[KRU01b] Y.W. Kim, G. Rizzoni, and V.I. Utkin. Developing a fault tolerant power-train control system by integrating design of control and diagnostics. *International Journal of Robust and Nonlinear Control*, 11(11):1095–1114, 2001.

[LBP+03] J. Luo, A. Bixby, K. Pattipati, L. Qiao, M. Kawamoto, and S. Chigusa. An interacting multiple model approach to model-based prognostics. In *IEEE International Conference on Systems, Man and Cybernetics, 2003*, volume 1, 2003.

[LCS+08a] Bologna L., Guerini C., Onori S., Rizzoni G., Salman M. A., and Zhang X. Hierarchical diagnosis and prognosis strategy for electrical power generation and storage system. volume 27, Ann Arbor, Michigan, USA, October 20-22 2008.

[LCS+08b] Weiwu L., Suozzo C., Onori S., Rizzoni G., Salman M. A., and Zhang X. Experimental calibration and validation of fault diagnosis and prognosis algorithms for automotive electric power generation and storage system. volume 27, Ann Arbor, Michigan, USA, October 20-22 2008.

[MPY+06] Salman M., Popp P., Zhang Y., Zhang X., and Chin Y. K. Vehicle diagnosis and prognosis: Concepts, trends, and applications to batteries. 2006.

[obd05] Malfunction and Diagnostic System Requirements for 2004 and Subsequent Model-Year Passenger Cars, Light-Duty Trucks, and Medium-Duty Vehicles and Engines, November 2005. Title 13, California Code Regulations, Section 1968.2.

[OKG+08] M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, and G. Vachtsevanos. Advances in uncertainty representation and management for particle filtering applied to prognostics. In *Proc. International Conference on Prognostics and Health Management*, Denver, CO, October 2008.

[PCO+09] A. Pezzini, M. Canova, S. Onori, G. Rizzoni, and A. Soliman. A Methodology for Fault Diagnosis of Diesel NOx Aftertreatment Systems. In *Proc. Safeprocess*, 2009.

[PCS] P. Pisu, M. Canova, and A. Soliman. Model-Based Fault Diagnosis of a NOx Aftertreatment System.

[PGG+95] Azzoni P., Cantoni G., Minelli G., Moro D., Rizzoni G., Ceccarani M., and Mazzetti S. Measurement of engine misfire in the lamborghini 533 v-12 engine using crankshaft speed fluctuations. In Journal of Engines, editor, *SAE Technical Paper 950837*, volume 99, 1995.

[Pis02] P. Pisu. *Hierarchical Model-based Fault Diagnosis with Application to Vehicle Systems*. PhD Thesis, Ohio State University, 2002.

[Ple04a] G.L. Plett. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 1. Background. *Journal of Power Sources*, 134(2):252–261, 2004.

[Ple04b] G.L. Plett. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 2. Modeling and identification. *Journal of Power Sources*, 134(2):262–276, 2004.

[Ple04c] G.L. Plett. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs-Part 3. State and parameter estimation. *Journal of Power Sources*, 134(2):277–292, 2004.

[PSR03] P. Pisu, A. Soliman, and G. Rizzoni. Vehicle chassis monitoring system. *Control Engineering Practice*, 11(3):345–354, 2003.

[PSYJ06] P. Pisu, A. Serrani, S. You, and L. Jalics. Adaptive threshold based diagnostics for steer-by-wire systems. *Journal of Dynamic Systems, Measurement, and Control*, 128:428, 2006.

[Riz89] G. Rizzoni. Diagnosis of individual cylinder misfires by signature analysis of crankshaft speed fluctuations. *SAE Transactions, Section 3*, 1989.

[SCGR05] L. Serrao, Z. Chehab, Y. Guezennec, and G. Rizzoni. An aging model of Ni-MH batteries for hybrid electric vehicles. *Proceedings of*

the 2005 IEEE Vehicle Power and Propulsion Conference (VPP05), pages 78–85, 2005.

[SOGR09] L. Serrao, S. Onori, Y. Guezennec, and G. Rizzoni. A novel model-based algorithm for battery prognosis. *accepted to the 2009 Safeprocess*, 2009.

[SRP07] A. Scacchioli, G. Rizzoni, and P. Pisu. Hierarchical model based fault diagnosis for an electrical power generation storage automotive system. New York, USA, July 11-13 2007.

[TH02] Ogawa T. and Morozumi H. Diagnostics trends for automotive electronic systems. 2002.

[WI95] M. Wurtenberger and R. Isermann. Supervision of lateral vehicle motion using a discrete parity space approach. In *Proc. 1st Workshop on Advances in Automotive Control*, Ascona, Switzerland, March 1995.