A Survey on Diagnostics Methods for Automotive Engines

Javad Mohammadpour*, Matthew Franchek, and Karolos Grigoriadis

Abstract- Faults affecting the automotive engines can potentially lead to increased emissions, increased fuel consumption or engine damage. These negative impacts may be prevented, or at least alleviated, if faults can be detected and isolated in a timely manner. The US Federal and State regulations dictate that automotive engines operate with an On-Board Diagnosis (OBD) system to enable the detection of faults resulting in increased emissions. In this paper, we survey and discuss the different aspects of fault detection and diagnosis in automotive engine systems. The paper aims to describe some of the efforts made in the academia and industry on the fault detection and isolation for a variety of component faults, actuator faults, and sensor faults in automotive engines using various data-driven and model-based methods.

1. INTRODUCTION

The basic concept of automotive On-Board Diagnostic (OBD) systems is to result in malfunction indicator light (MIL) illumination after a fault has been detected on two consecutive driving cycles. Pending fault codes are stored on the first detection and altered to “active” or “confirmed” codes once the MIL comes on. An error is considered to progress to a fault when it leads to produced emission that exceeds a pre-specified threshold.

There have been a number of survey papers on diagnostics of automotive systems including [1-3], with a limited scope of topics and focus on engine subsystems. The present paper is the first attempt to survey the work in the area of fault detection and diagnostics for automotive engines and aftertreatment systems. The aim is to classify the most relevant research articles from an academic perspective. It should be noted that there are many patents issued or pending in this area that will be not surveyed in this paper.

2. MONITORING REQUIREMENTS FOR AUTOMOTIVE ENGINES AND RELEVANT WORK

This section details the major monitors required for automotive engines. For each monitor, we describe the purpose, what needs to be detected (i.e., “malfunction criteria”), and some of the recent work that address each monitor. It is important to note that the OBD regulation only requires the system to be designed and calibrated to detect a single component failure at the required malfunction criteria rather than having to detect every combination of multiple component degradations that can cause emissions to exceed the malfunction threshold (e.g., 1.5 times the standards). In other words, OBD is not required to take into account synergistic effects of multiple component failures. For example, when calibrating an EGR low flow fault that would exceed the threshold, manufacturers would be required to implant only a low flow fault in the EGR system and leave other emission control components/systems (e.g., catalysts) in the nominal condition.

2.1. Fuel system: Manufacturers are required to detect fuel system faults that cause emissions to increase. The faults are involved in the fuel system pressure control (e.g., common rail fuel pressure control or hydraulic pressure control) and the focus is on detecting faults when the feedback system can no longer deliver the desired pressure. Given the critical importance of proper fueling for emission control, monitoring for properly injected fuel quantity and injection timing are also required.

A fuzzy-based pattern recognition method has been applied in [4] for real-time detection of fuel injection system faults in a diesel engine. The fuel system health diagnosis system consists of a piezoelectric pressure sensor to measure fuel injection pressure patterns and a fault diagnosis algorithm to detect abnormal injection pressure patterns and identify the causes contributed to thee abnormal patterns. A multi-net artificial neural network (ANN)-based diagnosis algorithm was proposed in [5-6] to detect a leaking fuel injector nozzle in a diesel engine, where it only used a pressure transducer. Nonlinear estimators using a sufficiently accurate model of powertrain system of an SI direct injection engine are designed by Lee et al. [7] to detect different actuators faults including high-pressure fuel injectors. Recently, Schilling et al. [8-9] developed a system to detect and isolate faults due to aging of the air and fuel path of common-rail direct injected diesel engines using an algorithm based on the information obtained from lambda (air-to-fuel ratio) and NOx emissions sensors. Faults corresponding to quantity of the injected fuel, mass air flow (MAF) and manifold pressure (MAP) sensors are taken into account to explain discrepancies in the expected lambda and/or NOx measurements. The fault detection method of the latter papers is model-based and uses a bank of extended Kalman filters; it is the first work reported to use emission sensors for fault detection purposes in diesel engines.

Payri et al. [10] proposed a diagnosis method for the injection process using the rail pressure measurement. The authors explored and evaluated different data-driven techniques to detect faults in common-rail injection systems. Chandroth et al. proposed to use cylinder pressure and vibration data [11] to detect the presence of a block in fuel injector and poor fuel atomization by training two sets of ANNs and using the features extracted from the cylinder pressure measurements and vibration amplitudes.

2.2. Misfire and Knock monitoring: For the 2010 to 2012 model years, manufacturers are required to detect malfunctions that cause a complete single cylinder misfire (e.g., one cylinder completely dead). For the 2013 and subsequent years, misfire monitoring will be required to be performed continuously (under all loads and speeds) and to look for lower levels of misfire (a cylinder or combination of cylinders that are intermittently misfiring) rather than just monitoring for a complete dead cylinder only at idle.

* Dept. of Mechanical Engineering, Univ. of Houston, TX 77204; email address: jmohammadpour@uh.edu
2.2.1. Misfire detection: Engine misfire detection is an important element of OBD systems since engine misfire can induce an increasing level of exhaust emissions and simultaneously damage the catalytic converter. Many methods have been proposed in the literature to address this problem including algorithms based on variation in engine shaft angular speed (also acceleration and torque), sparkplug voltage [12], oxygen sensor signal [13], knowledge-based expert system, and neural networks.

The first approach for misfire detection is based on the evaluation of the instantaneous angular velocity signal without using an engine model. These methods evaluate the characteristics of the time-domain, angular-domain or frequency-domain engine speed signal. The extracted features are then used to detect misfire through simple threshold check or more complex decision-making algorithms (e.g., [14-18]). Most of these algorithms give satisfactory results at low speeds, but due to the lack of a proper engine model it is difficult to correct the influence of the inertia torque at higher engine speeds.

The second approach is based on the use of model-based techniques, where a dynamic engine model to estimate the indicated torque or in-cylinder pressure is utilized. Rizzoni [19] and Connolly and Rizzoni [20] proposed an algorithm to estimate the effective torque based on the deconvolution in the frequency domain. As the inertia torque depends on the mean angular speed, this term is added to obtain the indicated torque. Rizzoni [21], Kao and Moskwa [22], and Wang and Chu [23] proposed the use of sliding mode observers to estimate the indicated torque, while Kiencke [24] proposed a Kalman filter-based algorithm. To tackle the issue of high complexity of the torque-estimation methods discussed above, researchers have tried to improve the real-time implementability and its use for on-line misfire detection [25-26].

Energy models have been used for misfire detection by Tinaut et al. [27], where they define two dimensionless energy indices for each cylinder with the first index evaluating the change in kinetic energy during the compression stroke and the second evaluating the change in kinetic energy during the expansion stroke. These two indices collectively provide a tool to detect the fault condition of each cylinder.

The methods proposed in [18, 28-29] are aimed to: (i) detect a missing combustion event, and (ii) classify the event into either misfuel event (i.e., missing injection) or misfire (i.e., missing ignition) using a feedback from an appropriate signal. A different approach to misfire detection is proposed in [30-32], where the authors describe a new analysis techniques based on a wavelet approach allowing for the extraction of the frequency components related to a misfire event and its localization in the time domain.

2.2.2. Knock monitoring: Engine knock is caused by spontaneous ignition of a portion of the air-fuel mixture during the combustion cycle. The very fast release of the chemical energy in the mixture results in high local pressure and produces a shock wave. This shock leads to the resonance of the combustion chamber, thus producing the knocking sound. Excessive knock could lead to engine damage, and so knock detection is a very important requirement for the engine control unit.

A number of methods have been proposed to detect the knock phenomenon in spark ignition and internal combustion engines. Samimy and Rizzoni [33] used joint time-frequency signal processing methods to detect knock in internal combustion engines. The idea is based on the use of the relationship between the engine excitation frequency taking into account the combustion chamber geometry and speed of sound in the cylinder charge. The frequency can be estimated using an acoustic model of the combustion chamber given by Draper's equation:

\[ f_{m,n} = \frac{c_0\sqrt{T\eta_{m,n}}}{\pi B} \]

where \( f \) is the resonance frequency, \( \eta_{m,n} \) is nondimensional, and the integers \( m \) and \( n \) refer to the radial and circumferential mode numbers. The parameter \( c_0 \) is the phase velocity constant, \( T \) is the gas temperature, and \( B \) is the cylinder bore diameter. From (2.1), another equation is extracted in [33] to enable the prediction of the existence of frequency shifts in the knock signals suggesting that conventional knock detection methods employing a stationary signal model can be improved by applying a time-varying signal detection method.

In laboratory applications, the in-cylinder pressure is used to represent the knock characteristics. A number of statistical-based methods using in-cylinder pressure signal have been proposed in the literature as a means to determine the knock intensity (KI). The most commonly used KI is the absolute value of the peak magnitude of the filtered in-cylinder pressure defined as \( KI = \max(p_{\text{ic}}) \), where \( p_{\text{ic}} \) is the band-pass filtered pressure signal. The filter cut-off frequencies are selected depending on the engine resonance characteristics given by (2.1) [34]. Borg et al. [34] presented a method to determine the knocking condition of a spark-ignition engine using the discrete wavelet transform as a means of analyzing the engine-block vibration signal and a fuzzy inference scheme to generate an estimate of the knock intensity introduced above. Previous efforts were also made in detecting knock using continuous wavelet transform [30, 35]. In addition to the wavelet transform, Fourier analysis has also been used for knock detection [36-37], where the spectral intensity of the knock resonance frequencies is used as the statistical test for knock determination.

2.3. EGR system monitoring: The exhaust gas recirculation (EGR) system is required to be monitored for three primary failure modes: low flow, high flow, and slow response to achieve the desired flow. EGR is one of the primary oxides of nitrogen (NOx) emission control mechanisms for the majority of engine manufacturers, and it is critical that the desired flow rate is being delivered. Accordingly, most manufacturers utilize feedback control systems to modulate the EGR valve to achieve a desired flow rate. The feedback system usually uses a MAF sensor, and the system compensates for small errors to achieve the desired flow rate. As long as the system can provide the desired flow rate, emissions stay relatively low. However, when the system can no longer achieve the flow it needs or it takes too long to reach the desired recirculation flow, emissions can increase dramatically. For a system that is feedback controlled to an actual flow rate, this emission increase should not occur until the system is close to its control limits, e.g., cannot compensate and deliver the desired flow rate. In addition, the performance of the EGR cooler would also need to be monitored to ensure it has sufficient cooling capacity.

Both data-driven and model-based methods have been proposed to detect faults in the EGR system. Gaussian radial basis function (RBF) neural networks with adaptive classifiers are employed in [38] to detect a stuck EGR valve in SI engines. A neural network-based method using Self Organizing Maps (SOM) was employed in [39] to detect...
malfunctions of the EGR system of a passenger car diesel engine. The SOM outputs a measure of similarity to “typical system behavior patterns”, and for the OBD system, this value can be used as a metric for system anomaly detection. Semi-physical models (identified with local linear neural networks) are used in [40] to detect a leaky or stuck EGR valve in combustion engines, where residuals are generated using signal models and filters. Another diagnosis system for diesel engines proposed in the framework of structured hypothesis tests was developed to detect EGR-valve stuck in closed position in [41], where it was shown that this framework is a useful engineering tool to systematically design model-based diagnosis systems. Another model-based fault detection method to identify and isolate the EGR valve actuator faults was proposed by making use of nonlinear estimators for a model of an SI direct injection engine in [7]. A more recent attempt to address the EGR system diagnostics in diesel engines was made by Mohammadpour et al. [42], where they developed a model-based method (based on the standard orifice flow equation representing the EGR flow back to the intake manifold) to detect low flow and high flow faults in the EGR system. The proposed fault detection scheme used a recursive total-least-squares (RTLS) method to estimate two parameters, whose changes were shown to be indicative of the fault type and its severity.

2.4. Hydrocarbon and NOx catalyst monitoring: Oxidation catalysts located upstream the particulate matter (PM) filter should be monitored to help the regeneration. The requirement also covers monitoring of the other HC converting catalysts such as NOx absorbers and selective catalyst reduction (SCR) catalysts. NOx catalysts, including lean-NOx trap (LNT) catalysts and SCR catalyst systems should be monitored. In general, the catalyst itself would be monitored to make sure it has sufficient NOx conversion to keep emissions below a threshold while additional components such as the SCR injection system components (urea or ammonia) are monitored for proper functioning. For 2010 to 2012 model years, the catalysts would need to be monitored and a fault needs to be detected when emissions exceed the standards by an additional 0.3 g/bhp-hr. For instance, for engines certified to a 0.2 g/bhp-hr standard, a fault would need to be detected when emissions reach 0.5 g/bhp-hr. In 2013 models, however, the threshold will drop down to the standard plus 0.2 g/bhp-hr. For diagnostics purposes, the same NOx sensor for feedback control must be also used for monitoring. For non-feedback SCR systems or passive LNT, manufacturers are allowed to be reaching lower conversion efficiencies. In the following sections, we discuss the efforts recently made on fault diagnosis of engine aftertreatment system to improve the emission control systems efficiency.

2.4.1. DPF system monitoring: As described in detail in [43], the only technology available to meet the diesel particulate filter (DPF) leakage monitoring requirement in 2007 was a pressure sensor combined with a flow measurement. This was generally found to be of limited capability [44] due to little or no separation between healthy and damaged filters and far-reaching implications on the monitor frequency. The main contributing factor in the limited performance of the DPF diagnostics methods based on pressure sensors is the high tolerance caused by the sensor due to noise factors not measured by the engine control system, as well as the driving conditions under which the monitoring occurs. It is suggested by CARB that the model-based methods may result in a more accurate detection rather than a merely pressure sensor-based monitoring.

Most DPF leakage monitors currently in production are based on pressure drop. The Darcy-Forchheimer equation driving the pressure drop across the DPF with a constant soot loading is given by [45]:

\[
\Delta p = a_0 + a_1vF + a_2\rho F^2
\]  

(2.2)

where \( v \), \( F \) and \( \rho \) represent the kinematic viscosity of the exhaust gas, volumetric flow, and gas density, respectively. For a DPF loaded with soot, the above equation is rewritten as

\[
\Delta p = a_0 + a_1R(\text{soot})vF + a_2\rho F^2
\]  

(2.3)

with \( R \) being the (normalized) restriction, which is a function of soot. The monitor detects a leakage in the DPF system when the pressure drop \( \Delta p \) is much lower than what the right hand side of (2.2) predicts. Note that this threshold depends on temperature and exhaust flow. There are many reasons contributing to the discrepancy between the Darcy-Forchheimer model output and reality. These along with an inaccurate approximation of the model coefficients \( a_0, a_1 \) and \( a_2 \) estimated from experimental data, taken with imperfect measurement equipment, lead to a far-from-accurate leakage detection method. The work by Cunningham et al. [46] provides a promising extension to the mean value pressure drop correlation to particulate load through Darcy’s law that is expected to be useful for DPF monitoring and control. van Nieuwstadt and Brahma [43] investigate the ability of the model-based DPF leakage detection methods over the pressure sensor-based DPF leakage monitor, where they presented the noise factors entering the relevant models and a numerical evaluation to assess the capability of the model-based leakage monitor under typical ranges of the noise factors. A recent work by Surve [47] proposed to correlate the pre- and post-DPF temperature and pressure signals to define its transfer function characteristics for the baseline DPF behavior. Assessment of how these characteristics change as a result of a fault in DPF forms the basis of the proposed fault detection algorithm in [47]. The method achieved a fault detection of lightly failed DPF not possible by current algorithms based on mean value pressure drop. In fact, the main contribution of [47] was the extension of dynamic pressure signal analysis from steady-state engine operation (proposed in [46]) to transient operating conditions. Gheorghiu et al. [48] proposed to use a spark discharge soot sensor to detect the presence of a crack in the filter causing the filter to become no longer airtight.

2.4.2. LNT system monitoring: The lean NOx trap (LNT) is one of the promising technologies to control NOx [49]. The LNT can experience performance deterioration and malfunctions that may go undetected. A simplified storage model that can be integrated into the existing control strategy for real-time LNT control and diagnosis was developed in [50] that captures the dynamics of NOx adsorption, reaction rate, and physical mass transfer process. Deactivation of the LNT catalysts is one type of fault that can compromise the NOx conversion efficiency if it is not properly monitored and compensated for. Thermal exposure during high load operating conditions or filter regenerations could also lead to the loss of activity, which is irreversible. Another cause of deactivation is related to the presence of sulfur in the fuel and lubricated oil due to the formation of sulfates on the catalyst surface that reduces the LNT storage
capacity. Therefore, the LNT catalyst degradation must be monitored. There have been very limited contributions to fault detection and isolation of LNT aftertreatment systems [51-54]. Recently, Canova et al. [51] used a time-varying nonlinear ODE model of the LNT system to generate the residuals using the system model, through comparison of the predicted and measured values of selected variables. The fault diagnosis method in [51] was designed to detect and isolate critical faults of the LNT aftertreatment system, including sulfur poisoning, deactivation of the catalyst storage sites due to thermal aging, regeneration controller fault, and faults in the sensors (including outlet NO\textsubscript{x} sensor and temperature sensor). The proposed diagnostic method is based on the parity equation approach and on the analytical redundancy.

2.4.3. SCR system monitoring: Selective catalytic reduction (SCR) is a well-proven NO\textsubscript{x} reduction technology used in power generation for more than 30 years and recently in automotive diesel engines. SCR catalysts are considered as the technology of choice for future heavy-duty applications, while LNTs appear to be more promising for passenger cars and light-duty trucks. This is due to the conversion efficiency, reliability and cost-effectiveness for regenerating the system using the onboard fuel.

There has not been much effort on the development of monitoring methods for SCR systems. It is, however, noted that all the aftertreatment systems share a common type of failure which is to be monitored for: catalyst aging. Ammonia storage capacity is an important parameter directly reflecting the SCR catalyst aging. Estimation of this parameter is then helpful in monitoring the SCR system physical conditions. In [55], a simple yet effective method was proposed to estimate SCR catalyst ammonia coverage ratio and storage capacity based on an extended Kalman filter (EKF). In [56], a sliding mode observer was designed to estimate the SCR catalyst ammonia storage based on the measurements of NO\textsubscript{x}, ammonia and temperature, and an EKF was used to eliminate the NO\textsubscript{x} sensor cross-sensitivity to ammonia.

2.4.4. TWC system monitoring: Modeling and control of three-way catalysts (TWCs) has been a widely discussed research topic [57]. Also, a variety of diagnostics methods based on the detailed thermodynamics-based modeling of the TWC have been developed [58]. Recently, the simplified models capturing TWC dynamics as an oxygen storage/release process have been employed for the catalyst monitoring purposes. Similar to other catalysts discussed so far, during its life, the TWC loses the storage property. The proposed on-line diagnosis methods in the literature based on the oxygen storage model aim to monitor the oxygen storage mechanism in order to detect the difference between a healthy and a deteriorated one. Let $0 < \theta < 1$ be the fraction of oxygen sites occupied in the catalyst, also known as relative oxygen level. The oxygen storage capacity is modeled as a limited integrator as:

$$\theta = \frac{1}{C(MAF)} \times 0.23 \times MAF \times \rho(\lambda_{\text{TWC}}, \theta) \times (1 - \frac{1}{\lambda_{\text{TWC}}}) \quad (2.4)$$

where $MAF$, $C(.)$, $\rho$, and $\lambda$ denote the mass air flow, exhaust catalyst capacity, oxygen exchange between the exhaust gas and the catalyst, and relative air-fuel ratio downstream the catalyst, respectively. Brandt and Grizzle [59] employed the above model to develop a diagnostics algorithm and analyzed that in the context of a hypothesis test based on the oxygen storage capacity of the TWC. The Neyman-Pearson criterion was used in [59] as the basis for the hypothesis test. A slightly modified version of the model (2.4) proposed by Fiengo et al. [60] is used in [60] to present a model-based stochastic approach for TWC fault detection. The method first generates a “residual signal”, which is the difference between the measured quantity of the oxygen storage capacity and the estimated one using the simplified phenomenological model. The diagnostic algorithm then works on the generated signals and implements a stochastic analysis in order to provide a statistical confidence in the TWC condition.

3. DIAGNOSIS OF SENSOR FAULTS AND LEAKS IN AUTOMOTIVE ENGINES

3.1. Detection and diagnosis of engine sensor faults: Sensor systems are critical components in all modern engineering systems. These measuring systems are extensively used not only to obtain system operational information but also to determine control actions. A sensor fault is typically characterized by a change in the sensor parameters or in its operational characteristics. The detection and diagnosis of these undesired changes plays a critical role in the operation of many engineering systems, and automotive systems are no exception to this. The design of sensor fault diagnosis schemes using the hardware redundancy and analytical redundancy approaches have been addressed in the literature [61]. In the hardware redundancy approach, redundant sensor systems are incorporated into the control system to improve the reliability of sensor measurement and enable sensor fault detection; however, cost and space make this approach unattractive. In contrast, the analytical redundancy-based fault diagnosis architectures use system physics-based models and information processing methods to achieve the necessary redundancy.

In the automotive systems literature, data-driven and model-based approaches have been proposed to diagnose different sensor faults. An early work using analytical redundancy method in [62] described the applicability of the model-based detection filters to diagnose a variety of sensor failures in automotive engines. UEGO sensor fault detection was studies in [63],[64-66].

In addition to UEGO sensors, throttle position (TP) [64, 66-70], MAF sensor [41, 64-65, 67], engine speed (ES) [64, 66, 71], and MAP sensors [41, 65-67, 69, 72-75] are among those extensively studied and experimented sensors to be monitored. Model-based and data-driven methods are also proposed to diagnose the ambient pressure sensors (both intake and exhaust) [65, 75-76]. The sensor faults are treated as either additive (bias term) or multiplicative uncertainties in model-based approaches.

3.2. Detection of leaks in automotive engines

As described in detail in [77], the detection of a leak particularly an air leak in the intake manifold can be difficult since under a range of operating conditions, the turbocharger wastegate inherently counteracts the fault and maintain the manifold boost pressure at a pre-specified level. Depending on the location of the leak within the intake manifold and the control method applied, the EGR process may be also affected leading to an increase in NO\textsubscript{x} emissions [77].

Efforts have been made to detect and diagnose different types of leaks in automotive engine systems. The use of nonlinear model-based adaptive observers to detect intake leaks in diesel engines is proposed by Ceccarelli et al. [78-79], where the authors design observers with fixed gains and
variable gains. Vinsonneau et al. [73] also designed a nonlinear observer to detect the manifold leakage in SI engines in real-time. The idea in [73] was to model the leak effect on the air flow as

$$m_{ai} = (j_1(\theta, N) + f_1(\theta, p_m)) \Psi(p_{aim})$$

where the term $f_1$ represents the flow perturbation due to the leak. The model was obtained following the same lines as described next. Structured hypothesis tests (SHT) are employed to detect a manifold leak (a leak in the intake manifold) and a boost leak (a leak between air mass flow and throttle right after the intercooler) in SI engines [69, 80].

The aforementioned model-based leak detection methods use flow equation through a restriction, i.e., the model for flow past the throttle. It is noted that due to the pressure difference direction, the air flow through a boost leak is in the direction out of the air tube. The manifold leak can be also described in a different way noting that in this case, the leak air flow is in the direction into the intake manifold. The authors in [41, 69, 80] used the above model along with SHT to diagnose the leaks, whereas a simple parameter identification based on the above leak models is employed in [81-82] to detect the faults by tracking the changes in the estimated parameters of interest. The manifold leak model was added to a nonlinear state-space representation of a diesel engine’s air path system model, and an extended Kalman filter was designed to estimate the flow corresponding to the leak and to detect possible intake manifold leakage.

REFERENCES:

8. Schilling, A., Model-based detection and isolation of faults in the air and fuel paths of common-rail DI diesel engines equipped with a lambda and a nitrogen oxides sensor. 2008, ETH Zurich: Zurich, Switzerland.
13. Plapp, G. and M. Klenk, Methods of on-board misfire detection, in SAE International Congress & Exposition. 1990: Detroit, MI.


44. van Nieuwstadt, M.J. and D.F. Trudell. Diagnostics for Diesel Particulate Filters. in SAE World Congress & Exhibition. 2004. Detroit, MI.


47. Surve, P.R., Diesel Particulate Filter Diagnostics Using Correlation and Spectral Analysis, in ECE, 2008. Purdue.


