EVALUATION OF A NOVEL DYNAMIC THRESHOLDING AND TREND ALERT GENERATION TECHNOLOGY ON A HUMS-EQUIPPED FLEET[*†]

Steven M. Krick
RMCI Inc., U.S. Army RDECOM
Aviation Engineering Directorate
Redstone Arsenal, AL 35898

Daniel R. Wade
U.S. Army RDECOM
Aviation Engineering Directorate
Redstone Arsenal, AL 35898

Kenneth Pipe
Universal Pipe Enterprises
Southampton, United Kingdom

Abstract
The US Army Condition Based Maintenance Program has utilized static thresholding techniques to monitor the condition of gearboxes, bearings, and shafts on nearly 2000 aircraft over the last decade. In some cases, vibration based diagnostics thresholds do not produce the targeted detection or accuracy rates prescribed by Aeronautical Design Specification 79C. The US Army Aviation Engineering Directorate has evaluated a dynamic thresholding technique produced by HUMAWARE. This paper discusses the selection of a ground truth data set for this evaluation and presents the results of the new technique when applied to a small fleet of Apache rotorcraft.

INTRODUCTION

Throughout the history of the U.S. Army Condition Based Maintenance (CBM) program, static thresholds have been used to define caution and alert exceedance levels for vibration-based diagnostics. Experience has shown that these levels can be inadequate for differentiating between faulted and healthy populations.[1] Separability, one of the four key attributes of vibration diagnostics defined by ADS-79C, must be achieved for successful implementation of CBM.[2] Good separability clearly distinguishes faulted populations from the fleet norm. The ADS-79C requirements for separability state that the false positive (FP) rate should be less than 10 percent and the false negative (FN) rate should be 10^{-6}, depending on the criticality of the failure.[2] The Army evaluated CFAR Autotrend to assess the performance of its dynamic thresholding and trend analysis (DTTA) techniques as an augmentation to the current static thresholding technique.

DYNAMIC THRESHOLD SETTING

CFAR

The CFAR technique functioned by splitting the threshold setting process into two stages. Each stage of the processing treated sensitivity to a defect and false positive rate separately without compromising the performance of either. A further strength of the CFAR technique is that it made no assumptions about the statistical or frequency domain characteristics of the data. The first stage of the processing established a primary threshold set at a level that produced a measurable exceedance rate for the data stream (Figure 1). The primary threshold level was controlled by a feedback loop to maintain the exceedance rate at a constant value, hence the name Constant False Alarm Rate (CFAR). The feedback loop set and maintained a threshold which maximized the probability of detection. The second stage of the processing addressed the problem of false alerts and subjected the data that exceeded the primary threshold to a binomial (binary) integration process, also known as M out of N processing (MooN). The MooN processing filtered random noise components out of the alerts to ensure that the primary threshold was only exceeded by an underlying trend or other deterministic component of the signal. Trending, or a similar underlying pattern caused by a defect, would correlate the data and trigger the secondary processing, whereas random noise or absence of a defect would not. The secondary processing had the benefit of remaining on once it was triggered by an alert, unlike static threshold techniques where the noise in the signal can cause the alert to switch on and off during the transition.

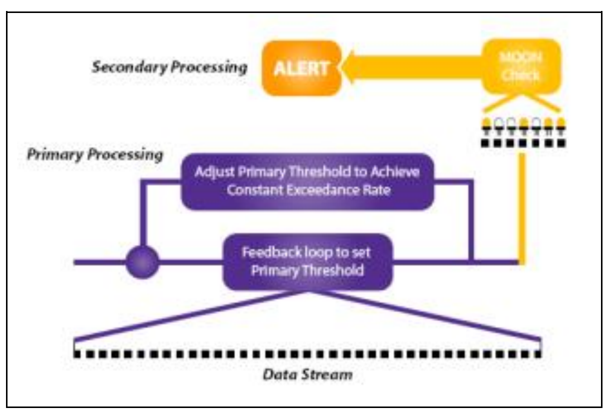

Figure 1 – CFAR Processing

[*] *Under 17 U.S.C. 105, the United States Government does not hold any copyright in works produced by U.S. Government employees within the scope of their employment.*

[†] *Presented at the American Helicopter Society 68th Annual Forum, Forth Worth, TX, 1-3 May, 2012.*

The amount of binomial integration was determined by the target FP rate. Just as the primary threshold exceedance rate was controlled to maintain a constant rate, the secondary processing output was also controlled to remain constant. This control allowed both the sensitivity to defects and the False Alert Rate (FAR) to be managed successfully.

Autotrend

The Autotrend technique functioned by using the Box-Car methodology (Figure 2). The box trend was calculated using the latest data in the data stream while the car trend was calculated using the remaining data. The turning point occurred where the two trends separated. For every new point in the data stream, the trend rate between the box and the car was calculated and compared using null hypothesis testing and a separation factor to determine the significance of the difference between them.

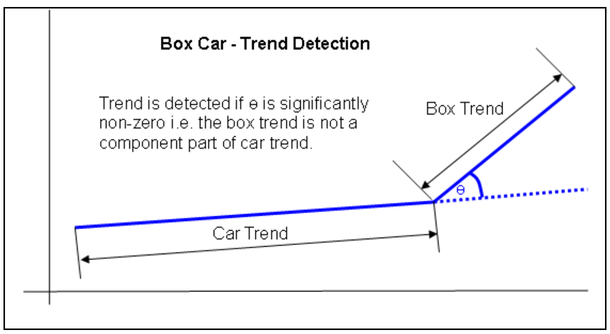

Figure 2 – Autotrend Processing

INITIAL SETUP

Evaluation Data

The U.S. Army assembled a sample dataset for this evaluation that included a total of 29 AH-64 aircraft. The aircraft were further subdivided into 5 64A and 24 64D model aircraft. Approximately half of the dataset contained a homogenous mixture of static thresholding true positive (TP), true negative (TN), and FP conditions as determined by detailed component teardown analyses. The other half of the dataset contained a sample of random and presumably healthy aircraft designed to boost the number of TNs in the dataset. Prior to beginning the evaluation, the Army was unaware that CFAR Autotrend required a minimum amount of data prior to component failure in order to identify anomalies. Several of the cases intended to demonstrate the ability of DTTA to distinguish between healthy and faulted data did not have the required number of data points prior to component removal; consequently, those cases were removed from the study.

Evaluation Methodology

The U.S. Army used an evaluation methodology that corresponded to that described in ADS-79C for Detection Algorithm Development (DAD) in order to evaluate the performance of DTTA.[2] The methodology dictated that, to be successful, DTTA must:

- Achieve earlier detection than static thresholding without degrading the current diagnostic accuracy, identifiability, and detectability.

- Correctly identify features that were not originally detected by static thresholding (verified by referencing the maintenance actions reported in the DA Form 2410 and component tear down analyses).

Parameter Optimization

The CFAR Autotrend algorithms used control parameters to optimize the detection performance of DTTA. The parameters were defined in terms of identifiable features in the data streams and did not require specialized mathematical knowledge. For the CFAR algorithm, the principal parameter is the Primary Threshold Exceedance (PTE) rate. The PTE rate controlled the primary threshold level and, hence, the sensitivity of the alerts to defects. This sensitivity could be reduced, if required, by setting the Conservatism Factor. The FAR

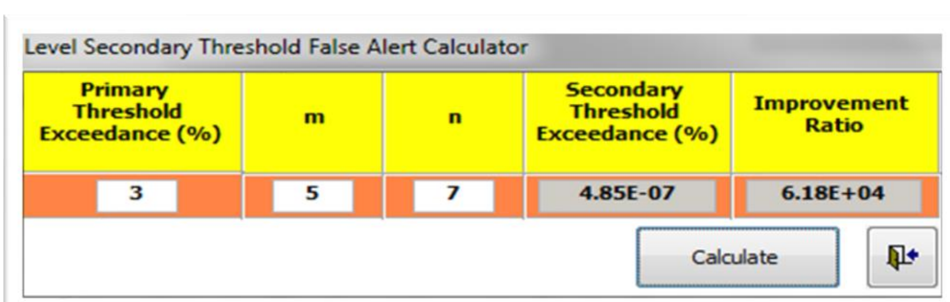

Figure 3 – False Alert Calculator

was controlled by setting the *M* and *N* parameters for MooN processing. A special calculator was used to predict the resulting FAR (Figure 3). The calculator determined whether or not a target FAR would be achieved. Parameters could also be set to determine whether the level change was large enough in amplitude to be considered a discontinuity or step change. The Autotrend algorithm was primarily controlled by Box and Car lengths. The Box and Car lengths determined if and when a trend would be detected. The parameters ensured that the FAR for Autotrend could be made similar to that of CFAR. The detailed parameter descriptions and the procedures for optimization are contained in References 3 and 5, respectively.

To establish a basic parameter set for the evaluation, a small sample of the data was analyzed by Humaware using the CFAR Autotrend analysis software. This parameter set was then applied to the evaluation dataset. With the base parameter set in place, CFAR Autotrend immediately began analyzing the evaluation data. Figure 4 demonstrates the ability of CFAR's primary threshold (magenta) to track rising (top) or falling (bottom) data streams while avoiding confusion from random peaks in the data. Once an alert was triggered (yellow), it remained triggered despite some periodic drop-outs in the data.

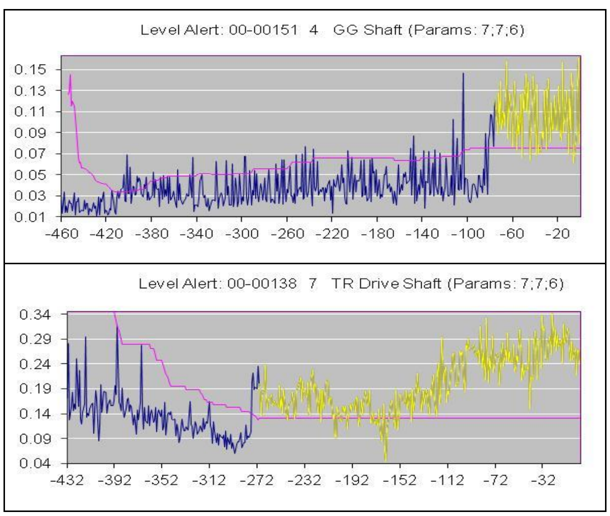

Figure 4 – Primary Threshold Adapting to Data Stream

The large number of threshold value changes in the charts demonstrates the requirement for the dynamic threshold setting functionality. Figure 5 demonstrates Autotrend's ability to accurately identify trends and their turning points. The figure also demonstrates that the algorithm was not confused by trivial events in the data, such as level changes or shallow trends that were not diagnostically useful.

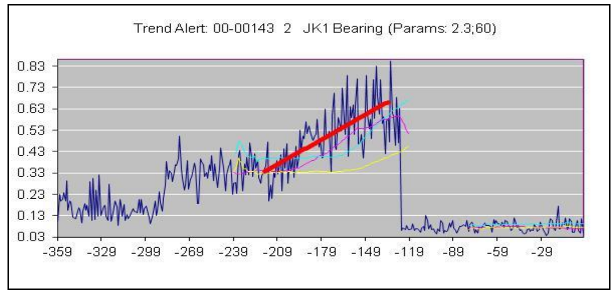

Figure 5 – Primary Threshold Adapting to Trend

The control parameters for the CFAR Autotrend were reviewed to optimize the results of the 4 ADS-79C metrics: TP, FP, TN, and FN. This was of particular significance for the Autotrend algorithm, which initially produced an uncharacteristically large number of multiple alerts as shown in the upper chart in Figure 6. The inflated number of true alerts indicated that the Box Length (the number of points that constitute a trend by the diagnostics process) was too short. The Box Length parameter was increased to reduce the number of trends that produced alerts as shown in lower chart in Figure 6.

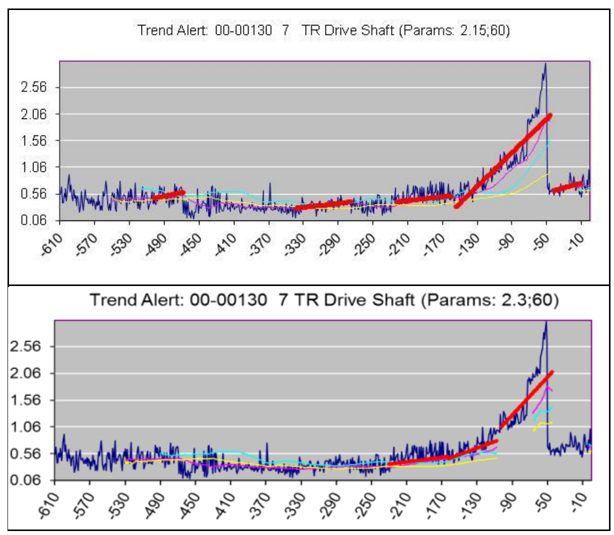

Figure 5 – Multi-Trend Detections

A number of trends occurred as a result of level changes which were not alerted by the CFAR algorithm as shown in the upper chart in Figure 7. This problem occurred because the primary threshold in the CFAR algorithm was too high. To correct this, the PTE rate was increased to heighten sensitivity to the level changes. The secondary processing MooN was maintained to keep the FAR in excess of 1 in 100,000. The correction is shown in the lower chart in Figure 7.

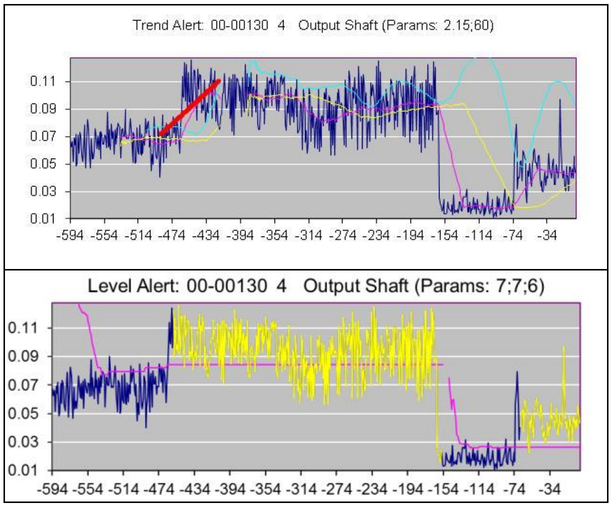

Figure 6 – Trends that Should Be Step Level Alerts

The discontinuity, or Step Conservatism Factor, was adjusted to remove ambiguities that existed between step and level alerts. The resulting changes to the parameter sets were implemented and the analysis was reinitiated. These modifications to the control parameters resulted in a positive increase in performance as shown in Table 1. The unverified alerts were compared to DA Form 2410 data to determine which alerts were true and which were false. The 2410 data drastically reduced the number of unverified alerts, although 11 alerts could not be resolved. The resulting 33 FP alerts were randomly distributed across aircraft, components, sensors, and condition indicator (CI) type and therefore yielded an irreducible residual FP rate.

Metric	Old	New
TP	480	367
FP	123	33
Unverified	165	11
Total	768	411

Table 1 – Alerts Before & After Parameter Optimization

RESULTS

Preliminary Evaluation

The preliminary evaluation data set consisted of 29 examples. Of these examples, 20 either produced alerts upon the component's installation or provided insufficient data to be trended. DTTA identified relative changes in data, but since installed defects did not produce a relative change in the data, static thresholding was still required to detect these defects. For the purpose of this evaluation, such cases were excluded from the analysis. The remaining 9 cases produced the metrics as shown in Table 4.

Metric	Static Threshold	DTTA	Ground Truth
TP	7	8	8
FP	1	–	–
TN	–	1	1
FN	1	–	–

Table 2 – Metrics Summary with Ground Truth

Current operational procedures affirm that a single CI is acceptable to produce a fault indication. Of the DTTA alerts, only one was a single CI alert. The remaining alerts were multi-CI, which may improve the accuracy, identifiability, detectability, and separability of the diagnostic processing as mandated in the evaluation methodology. The DTTA techniques identified two defects that were not detected by the static thresholds, which is evidence of a greater sensitivity to defects. The alerts for CFAR occurred, on average (with standard deviations), 11 (16) data points or 24 (30) days earlier than the static thresholding alerts; and the trend detections occurred 35 (16) data points or 34 (5) days in advance of the static thresholding alerts. The large standard deviation in the "days" interval was due to the random occurrence of aircraft downtime in the evaluation data set. The variance would be greatly improved if the metrics had been converted to flight hours; however, such a conversion of metrics was outside the scope of this evaluation.

Additional Findings

In addition to accurately corresponding to the ground truth results, DTTA identified an additional 28 events that indicated a possible component defect, fulfilling the second mandate in the evaluation methodology. All but 4 of these detections were confirmed to be legitimate by the DA Form 2410 data. The performance analysis produced the following Key Performance Indicators (KPIs) as shown in Table 5.

The majority of the static thresholding FNs discovered by DTTA occurred because the static thresholds were set artificially high. This was done during the early learning stages of CBM to ensure an acceptable FP rate. Regardless, since the data did come from a fielded system and since the defects were not alerted to the maintainer, it is reasonable to classify these defects as FNs for the static thresholding technique.

Metric	Static Threshold	DTTA
TP	2	24
FP	–	–
TN	–	–
FN	22	–

Table 3 – Metrics Summary without Ground Truth

Combining the two sets of results (Tables 4 and 5) with the uncertainty caused by the 11 unverified detections produces a best and worst case scenario as shown in Table 6.

Metric	Static Threshold		DTTA	
	Best	Worst	Best	Worst
TP	9	9	36	32
FP	1	1	–	4
TN	4	–	1	1
FN	23	27	–	–

Table 4 – Combined Metrics Summary

The FN rate for DTTA is idealistic. The total population of defects was discovered by DTTA techniques, not selected from an independent ground truth set of defects. This analysis was sufficient to compare the static threshold and DTTA techniques, but it did not provide the absolute performance measurement for the DTTA FN rate. Overall, there were 411 DTTA alerts. This abnormally large number was due to an artificially high number of events embedded in the data set and, as mentioned earlier, most alerts were multi-CI. In all cases, the alerts were analyzed as true or false, yet there remained 11 alerts that could not be resolved. This represents a level of uncertainty in the analysis as can be seen in Table 7.

Metric	Best Case	Worst Case
TP	378	367
FP	33	44
Percent of FP	9	12

Table 5 – Total DTTA Alerts Metrics Summary

The FP rate was within the target of 10 percent in the best case, but it was slightly outside the target in the worst case. Given that most of the uncertain CFAR Autotrend detections have been resolved as true, the best case scenario may be the more likely of the two. The alerts for CFAR occurred, on average, 17 points or 27 days earlier than the static thresholding alerts, and the alerts for Autotrend occurred 18 points or 15 days in advance of the static thresholding alerts. As in the previous case, the inconsistent distribution of time versus points is attributed to the large intervals of non-flying times that were randomly distributed throughout the data set. The number of DTTA alerts relating to a single maintenance event is shown in Table 8.

Number of Alerts	1	2–3	4–10	10+
Number of Events	10	12	13	2

Table 6 – Number of Alerts to Events Detected

Seventy-three percent of maintenance events resulted in more than one alert being generated. These results correlate to the number of CIs used on each component. If there was a rich set of CIs, then there was a rich set of alerts associated with an event.

Independent Evaluation

In addition to the primary evaluation, AMRDEC's Aviation Engineering Directorate (AED) Aeromechanics Division performed an independent evaluation of the features and performance of the DTTA techniques. This evaluation examined the functionality and simplicity of setting the various CFAR Autotrend parameters for the purpose of optimizing the results. The evaluation also extended the test set of data to include multiple new data sets containing known component failures. The DTTA techniques accurately detected 100 percent of the known failures, including one failure that the static thresholds were unable to detect (Table 9).

Metric	Static Threshold	DTTA	Ground Truth
TP	9	10	10
FP	–	–	–
TN	–	–	–
FN	1	–	–

Table 7 – Independent Evaluation Metrics Summary

CONCLUSIONS

The principal findings from the evaluation program are that the DTTA techniques:

- Exceeded the static thresholding detection accuracy.
- Provided significantly earlier detection than the static thresholds.
- Detected maintenance events (material faults) that were not identified by static thresholds.

A power distribution for the DTTA based on the valid test cases was 100 percent. Test cases included a number of faults present upon installation. CFAR Autotrend is designed to identify a relative change in the signal level and is not a suitable technology for detecting installed faults. The results clearly demonstrate that the ADS-79 False Alert target of 10 percent can be met. The known false alert in the test cases was not alerted by CFAR Autotrend. The automatic production of thresholds resulted in a large number of defect discoveries in the data set. This was due to the static thresholds being set arbitrarily high. Since the static thresholds are part of a technology in service, it is reasonable to say that these discoveries add to the FN rate. Due to the design of the evaluation—all faults were discovered by one or the other of the techniques—the DTTA techniques did not identify any FNs.. To determine the true FN rate, a test case would need to be constructed from the DA Form 2410 records without reference to the defect detection source, and the performance of the DTTA techniques would need to be evaluated.

RECOMMENDATIONS

The assessment finding is that DTTA techniques are ready for the next stage of development. The following work items represent the issues that need to be addressed in order to develop an in-service DTTA capability:

- Verify all relevant CI data sets for HUMS-equipped fleets. This is required to ensure that the performance reported in the assessment is valid for the entire fleet and that the parameters are correctly set and allocated to the CI data types.

- Verify the levels of conservatism to be applied to the setup parameters to minimize the No Fault Found (NFF) rate, yet not compromise the TP and FN rates. The assessment shows that the processing can be too sensitive to defects, raising the prospect of increasing the NFF rate and costing useful life. The parameters can be lowered to reduce sensitivity.

- Assess other data sets, such as engine Health Indication Test (HIT) check trending. There are other data sets being recorded in the HUMS, engine performance, and usage that could benefit from the use of the technology. This should be evaluated before fielding.

- Verify integration with existing static thresholds.

- Revise the training and doctrine for the use of trend-based alerts by personnel utilizing the system.

- Validate the software functionality and the operational interfaces for the control and processing of the CI data.

- Develop a maintenance action reset function.

ACKNOWLEDGEMENTS

The authors would like to thank Humaware and Dr John Berry of the U.S. Army International Technology Center-Atlantic (USAITC-A) office for their support of this project.

REFERENCES

1. Wade, Daniel; Antolick, Lance; Branning, Jeremy; and Dempsey, Paula, "Evaluation of Gear Condition Indicator Performance on Rotorcraft Fleet," American Helicopter Society Annual Forum 66, Phoenix, AZ, 2010.

2. ADS-79C, http://www.redstone.army.mil/amrdec/sepd/tdmd/StandardAero.htm.

3. Pipe, Kenneth, "Improved Methods of Alert Generation in HUMS," American Helicopter Society Annual Forum 58, Montreal, Canada, 2002.

4. Dempsey, Paula J., "Signal Detection Theory Applied to Helicopter Transmission Diagnostic Thresholds," Glenn Research Center, Cleveland, OH, NASA/TM—2008-215262.

5. Pipe, Kenneth, "Parameter Setting Quick Reference Guide," Version 1.5, March 2010.