Telemetry Parameter Synthesis System to Support Machine Learning Tuning and Validation

Guang Chao Wang
Oracle Physical Sciences Research Center
OracleLabs, Oracle Corp.
San Diego, CA
guang.wang@oracle.com

Kenny Gross
Oracle Physical Sciences Research Center
OracleLabs, Oracle Corp.
San Diego, CA
kenny.gross@oracle.com

Abstract—

Advanced machine learning (ML) prognostics are leading to increasing Return-on-Investment (ROI) for dense-sensor Internet-of-Things (IoT) applications across multiple industries including Utilities, Oil-and-Gas, Manufacturing, Transportation, and for business-critical assets in enterprise and cloud data centers. For all of these IoT prognostic applications, a nontrivial challenge for data scientists is acquiring enough time series data from executing assets with which to evaluate, tune, optimize, and validate important prognostic functional requirements that include false-alarm and missed-alarm probabilities (FAPs, MAPs), time-to-detect (TTD) metrics for early-warning of incipient issues in monitored components and systems, and overhead compute cost (CC) for real-time stream ML prognostics. In this paper we present a new data synthesis methodology called the Telemetry Parameter Synthesis System (TPSS) that can take any limited chunk of real sensor telemetry from monitored assets, decompose the sensor signals into deterministic and stochastic components, and then generate millions of hours of high-fidelity synthesized telemetry signals that possess exactly the same serial correlation structure and statistical idiosyncrasies (resolution, variance, skewness, kurtosis, auto-correlation content, and spikiness) as the real telemetry signals from the IoT monitored critical assets. The synthesized signals bring significant value-add for ML data science researchers for evaluation and tuning of candidate ML algorithmics and for offline validation of important prognostic functional requirements including sensitivity, false alarm avoidance, and overhead compute cost. The TPSS has become an indispensable tool in Oracle's ongoing development of innovative diagnostic/prognostic algorithms for dense-sensor predictive maintenance applications in multiple industries.

Keywords—Fourier Decomposition, Signal Synthesis and Projection, Signal Spike Detection.

I. INTRODUCTION

Machine learning (ML) researchers and data scientists are proliferating throughout the world thanks to the expansion of Internet-of-Things (IoT) dense-sensor applications across many industrial segments, including manufacturing, transportation, oil&gas, and utilities. A common problem lamented by ML researchers is to acquire sufficient real data that can be used for evaluation, tuning, training, optimization, and validation of candidate ML innovations.

Time series databases make up a large and growing portion of the tech community's data base business, thanks to the expansion of Internet-of-Things dense-sensor applications across many industrial segments. For example, a modern oil refinery these days has 1M sensors recording time-series signals 24x7x365. A typical large commercial airplane has 75,000 sensors these days, and a medium-size enterprise or cloud data center can have 1 million sensors. One significant challenge for industrial use cases such as these, when the number of sensors and the sampling rates for those sensors both are climbing every year, the challenge becomes acquiring and retaining sufficient volumes of data to be able to validate ML prognostic specifications. For example, to validate that false-alarm probabilities (FAPs) are being met, when the specifications on FAPs are very small (example: 1 in 10^5 over 10K Hrs of operation), it would require over 5 years of signals to establish with a 99% confidence factor that the desired FAP is being met. We will show in this paper how validation objectives can be met with much shorter telemetry archives (and storage media) using a novel approach to high-fidelity signal synthesis.

One significant challenge for large-scale signal synthesis methodology has been that conventional approaches cannot accommodate time series signals containing spikes in the signals. The best technique in the classical industry literature for handling signals that contain spikes is a well-known spike detection algorithm developed by Goring & Nikora [1], a technique that suffers from several limitations that prevent using it as a basis for high-fidelity reconstruction of synthesized signals because it uses a guiding metric for detection of changes in signal quality, changes that are deemed "abnormal" with respect to the variance of the "base" signature (i.e. the base signal just before and just after spikes). Especially it is not uncommon in many areas of big data analytics for two or more moderate or wide spikes to superimpose and "fool" conventional spike detection algorithms into counting superimposed spikes as one very wide spike. Another area of applications for which conventional spike detection algorithms can severely underperform is for use cases where the "base signals" are noisy and the height-to-width ratios for the spikes becomes smaller (i.e. within 2 Standard Deviations of the noise for the base signal). For these types of challenging use cases, classical "state of the art" spike detection algorithms may have poor performance.

In this paper we propose a novel algorithmic infrastructure for processing a database of time series signals to solve the above challenges: the Telemetry Parameter Synthesis System

(TPSS). It advances a prior innovation of spectral decomposition and reconstruction of telemetry signals [2] to the next level. Specifically, TPSS allows any database of time series signals to be processed and decomposed into their deterministic and stochastic components, and then generates synthesized signals that possess exactly the same deterministic structure (including serial correlation for individual signals, cross correlation for multivariate signals, and periodicities for any number of seasonality components) and stochastic distributions for any amount of noise on the signals, including variance, skewness, and kurtosis. It features projecting synthesized signals into future time window of interest (optimal time series forecasting) without discontinuities at the end of the learning window. Moreover, our systematic and parametric spike detection, despiking/respiking algorithm demonstrates outstanding feasibility, practicability, and fidelity to handle spike distributions that arise in real signals.

The proposed technique is capable of producing an extremely high-fidelity synthetic database of signals possessing the following characteristics:

(1) The synthetic signals are statistically indistinguishable from the original time series database. The synthesized database of time series have exactly the same serial correlation structure, cross correlation structure, and stochastic content (matching means, variance, skewness, and kurtosis, and kolmolgorov-smirnov statistic).

(2) To be able to filter out the principal serially-correlated, deterministic components from telemetry variables so that the remaining stochastic signal (i.e., residual function) can be analyzed with signal validation tools that are designed for signals drawn from independent random distributions.

(3) High sensitivity for detection of large as well as small spikes contained in signals through a comprehensive systematic parametric iterative procedure. The integrated methodology creates high-fidelity spike detection, characterization, "despiking" analysis, and then synthesizes all conceivable spike distributions in the synthetic signals, including positive and negative amplitudes, shapes, widths, and temporal distributions.

From a ML Research perspective, all of the goals and objectives of data scientists conducting ML R&D for prognostics use cases, will find the new TPSS capabilities essential for:

- Assessing false alarm probabilities for new ML algorithms (FAPs)
- Assessing missed alarm probabilities for new ML algorithms (MAPs)
- Assessing sensitivity and "time-to-detection" metrics for discovery of subtle
  anomalies creeping into time-series processes
- Assessing overall compute cost for various new (and old) ML algorithms

II. METHODOLOGY

A. TPSS- Spectral Decomposition and Reconstruction

TPSS performs Fourier decomposition and reconstruction of telemetry signals based on the number of modes (sinusoidal periodicities) detected in the spectral domain of the telemetry signals. The reconstructed signals possess exactly the same statistical noise idiosyncrasies as the original telemetry characteristics.

To proceed, the time series signal is first analyzed through Fast Fourier Transform (FFT) analyses, and the highest peak in the resulting power spectral density (PSD) is captured. With the frequency and amplitude corresponding to the PSD peak, a Fourier composite signal in the time domain is reproduced. Then the synthesized Fourier composite is optimally synchronized with and subtracted from the original signal. The resulting residual is sent to a next round of spectral decomposition and reconstruction. This process is iterated until no more prominent modes are left in the residual, confirmed by both the kolmolgorov-smirnov test and the common white noise test [3]. Finally, the stochastic content of the residual is analyzed, synthesized, and superimposed onto the Fourier composites constructed during the previous iterations.

(c)

Figure 1a-2c illustrate the performance of TPSS through a use case where a telemetry signal comprising an envelope of three modes is presented. In this illustrative example, each mode discovered by TPSS is iteratively recreated, synced and subtracted from the original signal while at the same time the Fourier composite is constructed and superimposed (Figure 2a). After the final residual becomes normal and white noise, or the number of detected modes has reached a pre-specified value, the stochastic content of the residual is synthesized and superimposed onto the prior Fourier composite (Fig. 2b). Note that the histograms in the two figures are showing two different entities. It can be concluded that the top histogram (Figure 2a) of original signal is improved by using TPSS because the resulting residual exhibits a more Gaussian structure, per the bottom histogram in Figure 2b. This nature of signal pre-whitening suggests guaranteed optimal performance of the sequential probability ratio test, SPRT [4-5], for advanced prognostics and proactive anomaly detection. Finally, the synthesized signals can be projected into the future (or even "backcasting" into the past) given the length of a time window of interest (Fig. 2c). Observe there are no discontinuities between the synthetic signal and projected signal, because the Fourier composite (i.e. envelope of superimposed sinusoids) are repetitive and continuous.

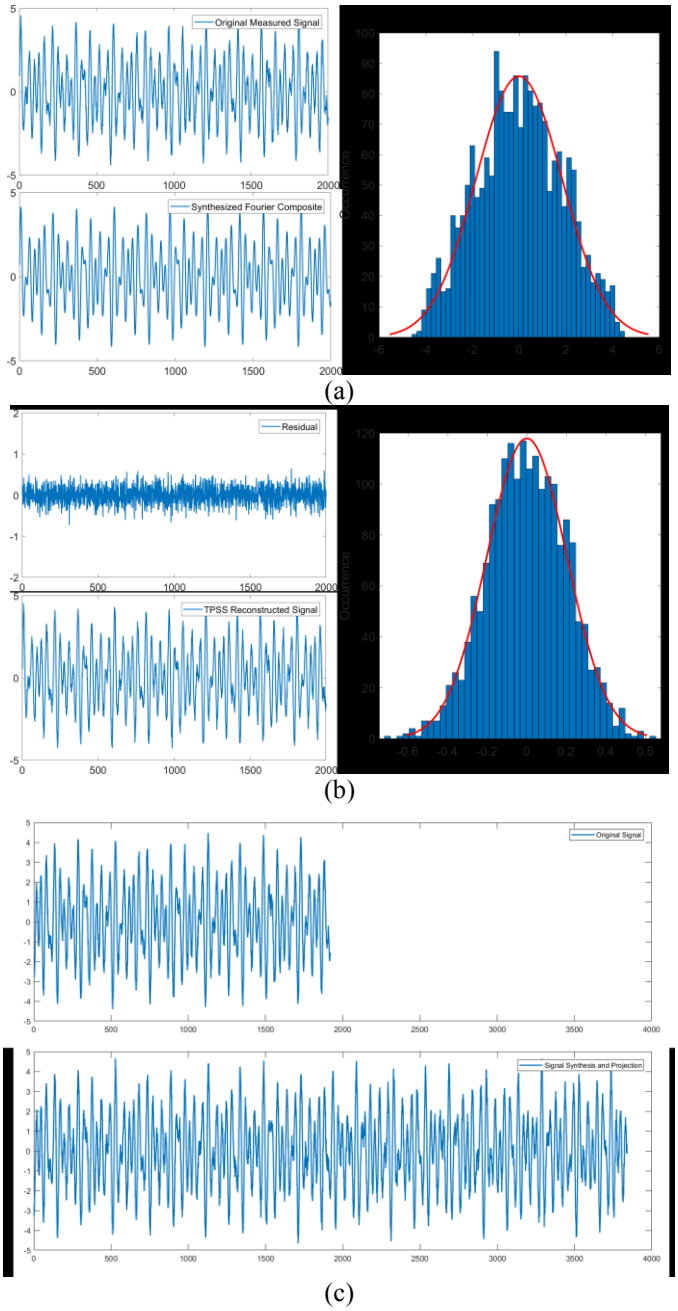

Figure 1: Use case of synthesis of telemetry signals by using TPSS.

The new TPSS reported herein is being used for empirical testing and validation of any ML-based prognostic and diagnostic surveillance techniques, and for evaluation of control system efficiency, when only limited archives of normal or faulted signals are available. One of the most valuable features of TPSS is the ability to analyze a relatively short segment of actual asset signals, then generate millions of hours of synthesized signals to be used in verification and validation (V&V) for advanced ML techniques for assurance that target false-alarm probability functional specifications are being achieved.

B. TPSS-Spike Detection Algorithm

To ensure a high-fidelity signal synthesis, we need to remove spikes in the signal upstream. The improved spike detection algorithm (red box in Fig. 1) is further introduced in this section. The prototype of the most popular classical spike detection technique (Wavelet Thresholding method) originates from an article which proposed a "universal threshold" for detecting and removing spike noise from a signal [6]. Later, Goring and Nikora [1] further implemented a universal threshold based Phase-Space Thresholding method to improve the spike detection for their acoustic signals. The universal threshold is a function of the number of observations (i.e. sampling rate) and the standard deviation (STD) for the signal. However, when spikes become wide, the STD becomes larger and exhibits a different nature so the thresholds derived for prior "needle spikes" no longer hold true, and conventional spike detection methods break down as the widths of the spikes increases, and can severely miss-characterize signals when occurrences of two or more spikes can overlap. Thus the classical state of the art spike detection methodology does not work well for "longer period fluctuations", which commonly manifest as wide spikes in the telemetry signals.

To address the above challenges, an enhanced spike detection approach is required to characterize and mimic the patterns of stochastic and dynamic spikes in the original signals. We advance the traditional spike-detection algorithm to next level with a novel and enhanced technique that 1) possesses increased sensitivity for detection of large as well as small spikes, and 2) is able to quantitatively evaluate itself and reports the spike detection efficiency using a new parametric Monte Carlo simulation based approach. We demonstrate the improved performance with use cases for which we compare "ground truth" spikes with "detected" spikes.

We propose an adjustable (and hence optimizable) parameter called the damping factor (DF) to suppress the universal threshold in [1, 6] and enhance the sensitivity for spike detection so that small, large, and overlapped spikes can be detected. While DF is just a scale factor ranging from 0 to 1, our real innovative work is to develop a comprehensive automated optimization framework which allows future users to automatically identify near-optimal thresholds instead of manually trying a scale factor for their specific signals. We present our innovative improvement to state-of-art spike-detection algorithms in two aspects:

1) SimSpike: a systematic parametric recursive iteration technique for evaluating the overall detection efficiency in terms of True Detections (Ts) and False Detections (Fs) vs. DF.
2) Performing near-optimal despiking for any given use case for any given end customer with any batch of spiky signals without knowing the true spikes.

First, to understand the relation between quantification of DF and the resulting spike detection performance, we introduce a systematic parametric characterization analysis (denoted as

"SimSpike") of *True* spike detections (Ts) and *false* spike detections (Fs), wherein the detection efficiency becomes a multidimensional function of multiple variables. A monte-carlo nested loop structure is devised to perform a comprehensive parametric investigation on the correlation between a DF value and resulting Ts/Fs ratio performance with any given set of spike characteristics (i.e. height, width, base-signal noise level, and Inter-arrival times (IATs) of the spikes, refer to next section for details).

Figs 4-7 together illustrate how damp factors and Ts/Fs performance correlate to varying spike widths, amplitudes, and signal noise level. In each of the following four figures, there are two surface plots: the top one (as suggested in the z-axis) represents the Ts value while the bottom one represents the Fs value. The x-axis indicates the STD of the noise added onto the base signal. Duration offset (y-axis) suggests the number of observation points by which the base spikes are expanded. For example, duration of 1 indicates all base spikes are widen by 1 extra observation point. The color represents the optimized damp factor that yields the best Ts/Fs performance. The spike height is specified in the figure title. It is defined as the number of observation points by which the peaks of the base heights are lifted.

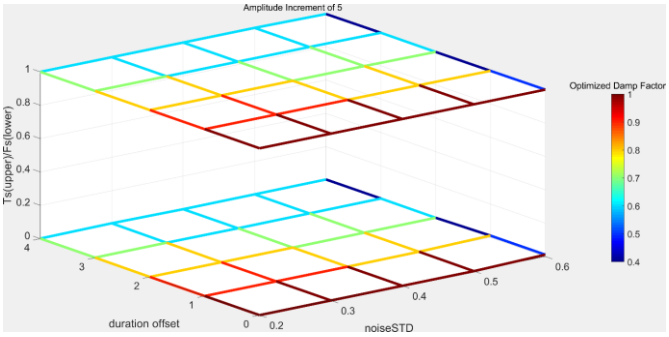

Figure 4: parametric nested loop analysis with spike height increment = 5 observation points.

Figure presents the first case where the height of base spikes is lifted by 5 observation points, suggesting prominent spikes in the signal. It can be concluded that if all spikes are substantially taller than base signal, it is easy to capture them all (100% Ts and 0% Fs). However, to ensure a perfect job, a nice DF value needs to be optimally analyzed from case to case. For example, on the near corner of the figure, when spikes have moderate widths, and signal noise level is low, the optimal DF is 1, indicating that suppressing the phase-space threshold is actually not necessary and the state-of-art spike detection algorithm [1] would work adequately for this "easy" use case. When the noise level becomes high, and/or all spikes become wider, a smaller DF needs to be applied to make the spike detection algorithm sensitive enough to capture them all. As a result, the color gradually turns from dark to light and consequently the optimized DF is lowered to 0.4 at the far corner of the surface when the signal is noisy and the spikes are wide.

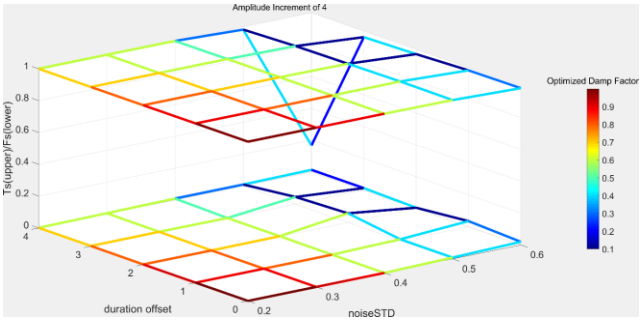

Figure 5: parametric nested loop analysis with spike height increment = 4 observation points.

Fig. 5 presents a comparison where the spikes become less prominent. At the near corner, the Ts and Fs can still reach 100% and 0% respectively with a less than 1 DF value, while the far corner area of the surface of Ts starts to drop even the damp factor is lowered to 0.1, and accordingly the far corner of the surface of Fs starts to hump. This phenomena illustrates that if spikes are not prominent and/or narrow enough, the signal noise and wide spikes can affect the performance of the conventional spike detection method, which is expected as discussed in the introduction section.

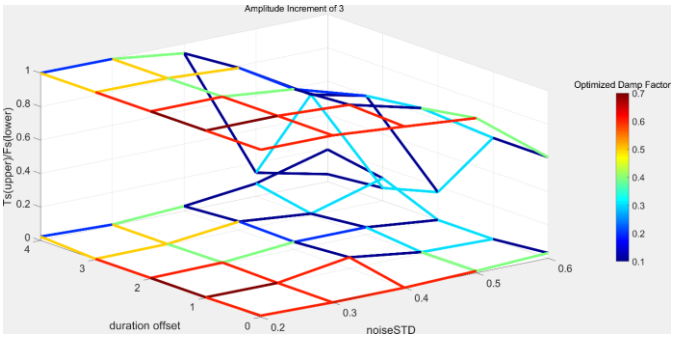

Figure 6: parametric nested loop analysis with spike height increment = 3 observation points.

Fig. 6 further reduces the spike height, causing more areas of the Ts to collapse and more areas of the Fs to further hump. Finally, Figure 7 presents an extreme case where the spike height is short enough to hide behind the signal noise, causing suboptimal spike detection.

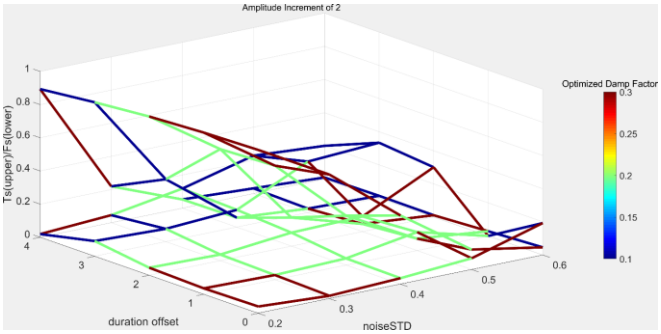

Figure 7: parametric nested loop analysis with spike height increment = 2 observation points.

In real world IoT applications, the customer's dataset will have no "ground truth" signals that define exactly how many spikes are in each signal. Consequently it is impossible with classical prior-art techniques to fully validate a spike-detection algorithm in terms of Ts and Fs, and the detection efficiency ratio Ts/Fs. Moreover, since for prior-art techniques we could not know the true Ts and Fs for an algorithm, it's not possible to select an appropriate DF ahead of time. This is because if we simply "turn the knob" on DF, and witness a different number of spikes getting identified each time we change DF, we still would have no way (with conventional techniques) of counting Ts and Fs for the new number of detected spikes with each new setting of DF. Thus to make our spike detection enhancement operational for real data, we have introduced an optimization framework of DF autonomous identification, which supplies us a near-optimal DF even though the original dataset of signals does not have any "ground truth" labeling of spikes. Figure 8 presents the idea of damping factor (DF) multidimensional optimization in detail.

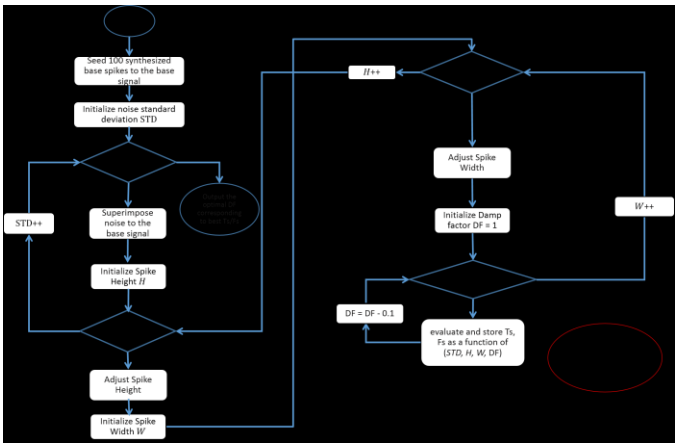

Figure 8: Flowchart of the damping factor (DF) multidimensional optimization for any given use case without a-priori knowledge the true spikes.

We start the first iteration by conducting a run for the original measured spiky signals, with a "nice" initial guess of DF value (based on our experience, a good starting value is 1.0), and the spikes with that DF value are extracted and removed (Fig. 9).

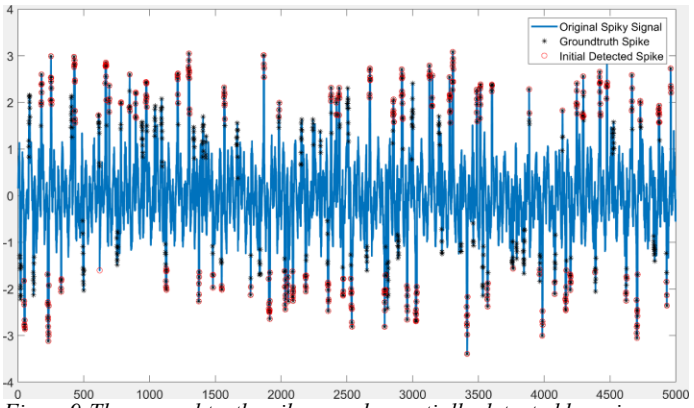

Figure9:The ground truth spikes can be partially detected by using an initial guess of DF value in standard spike detection algorithm.

Next we take those extracted spikes and expand their characteristic metrics by ±10 % of the range of the measured spike Heights, Widths, and IATs. We generate "ground truth" spikes by sampling from the new expanded distributions, and seed (i.e., the respiking process, which will be elaborated in the next section) to the prior despiked signal (Figure 20).

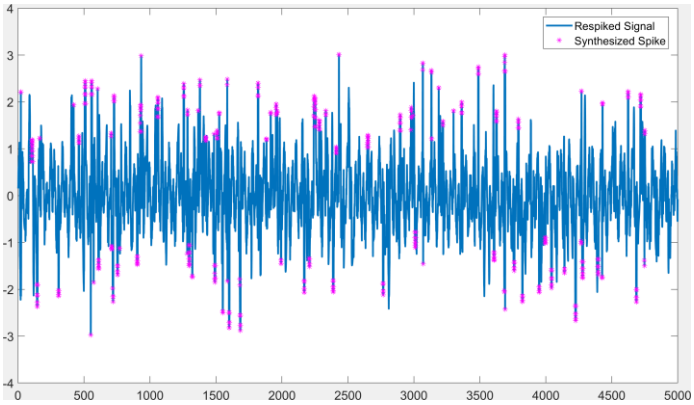

Figure 2: The distributions of the measured spike characteristics are expanded and the resulting synthesized spikes are seeded into the prior despiked signal.

Since we are simulating these signals and spike characteristics, we now know in all these simulation replications the "ground truth", hence we can truly optimize DF as a function of spike height width, and IAT through the procedures presented in . Finally, after the best possible DF value is determined, we now perform spike detection process again to the original spiky signal and capture more real spikes than the initial run (Fig. 11).

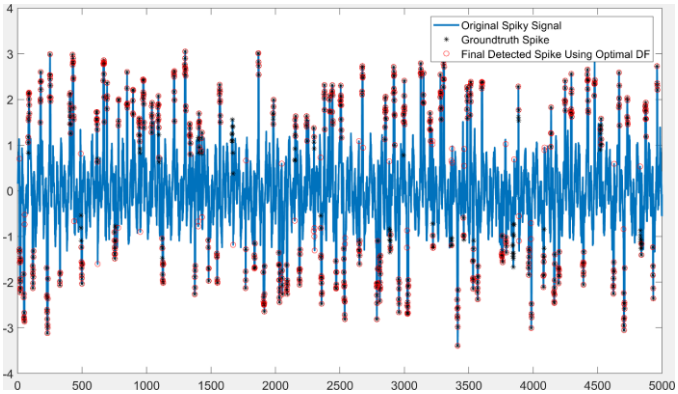

Figure 11: The final optimal DF value is reapplied to detect the spikes in the original measured signal, resulting in an improved spike detection performance.

By the above procedures we reach the goal of optimal despiking and spike-characterization for high-fidelity synthesis of time series signals containing spikes. This capability will give ML researchers significantly enhanced capabilities for high-fidelity synthesis of IoT signals.

C. TPSS-Signal Respiking Algorithm

The last feature of the proposed TPSS enhancement includes a high-fidelity respiking algorithm which characterizes the temporal distribution of the spikes regarding IATs, widths of both positive and negative spikes (WoP, WoN), and positive/negative amplitudes (AoP, AoN) of spikes, then simulate a set of spikes in a manner that possesses a nearly identical distributions of IATs, widths and heights. Then the simulated spikes are seeded into the synthesized signals for higher-fidelity synthesis of time series signals with any types of spikes.

To demonstrate the capability of the respiking algorithm, we present a side by side comparison on histogram of amplitude, width, and IAT of both original and simulated spikes. In Figure 123, the spikes previously captured during the spike detection process are first characterized. To compute the widths of the spikes identified for removal, we use the "full width at half maximum" metric, FWHM, which is defined as the width of the spike at exactly one-half of its peak amplitude. To determine the amplitude of the spikes, we replace the detected spiky points by interpolating the nearest non-spike neighbors and then subtract spiky points from the interpolated baseline. Then, after the despiking process, we generate empirical distributions of spike characteristics for producing simulated spikes with high fidelity. Finally, the simulated spikes are seeded onto the synthesized signals in a manner that matches closely the spike distributions of amplitudes, widths, and temporal sequences (via IATs) of the raw measured signals (Figure 1343).

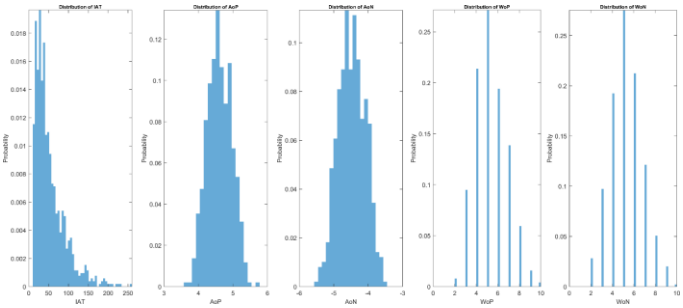

Figure 123: Distribution of the spike characteristics (IAT, amplitude, width) that were characterized from the original signals by the systematic technique introduced in Section II.A and II.C.

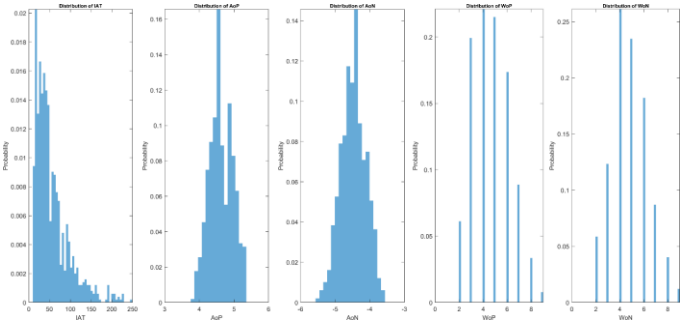

Figure 134: Distribution of the simulated spikes that will be seeded into the synthesized signal.

III. CONCLUSION

In this paper, we demonstrate our ongoing development and enhancement for the prior telemetry parameter simulation system (TPSS). It has the capability to process and synthesize any stochastic telemetry variable from all types of sensors for ML researchers to use in development, evaluation, tuning, and optimization of new ML related research for prognostics. TPSS employs Fourier-based decomposition and reconstruction methodology with the capability to efficiently decompose any signals into their deterministic and stochastic components, then reconstructs new, simulated signals that possess exactly the same statistical noise idiosyncrasies as the original telemetry variables. From a ML research perspective, all of the goals and objectives of ML innovators will obtain identical conclusions whether the candidate ML algorithms are applied to the synthesized telemetry database or to the original telemetry database.

In addition, the enhanced despiking and respiking feature further make the technique even more robust to real signals across dense-sensor applications in IoT industries. Lastly, the ability to generate long signal streams by only analyzing a short segment of actual asset signals is an extremely valuable capability for use for prognostics Proof-of-Concept Demos, which otherwise can be become prohibitively costly in empirical validation studies for big-data IoT applications in dense-sensor industries such as manufacturing, transportation, oil&gas, utilities, healthcare [7] and of course IT datacenters [8].

REFERENCES

[1] Goring, D. G., & Nikora, V. I. (2002). Despiking acoustic Doppler velocimeter data. *Journal of Hydraulic Engineering*, *128*(1), 117-126.

[2] Gross, K. C., & Schuster, E. (2005). Spectral Decomposition and Reconstruction of Telemetry Signals from Enterprise Computing Systems. In *CDES* (pp. 240-246).

[3] Mahan, M. Y., Chorn, C. R., & Georgopoulos, A. P. (2015). White Noise Test: detecting autocorrelation and nonstationarities in long time series after ARIMA modeling. In *Proceedings 14th Python in Science Conference (Scipy 2015), Austin, TX*.

[4] Gross, K. C., & Lu, W. (2002, June). Early Detection of Signal and Process Anomalies in Enterprise Computing Systems. In *ICMLA* (pp. 204-210).

[5] Masoumi, T., & Gross, K. C. (2016, December). SimSPRT-II: Monte Carlo simulation of sequential probability ratio test algorithms for optimal prognostic performance. In *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on* (pp. 496-501). IEEE.

[6] Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, *81*(3), 425-455.

[7] K. C. Gross and D. Gawlick, "Combining Advanced Machine Learning with Situation Awareness for Big Data Health Informatics Applications," ***4th IEEE Intn'l Conf. on Health Informatics and Medical Systems (HIMS'18)***, Las Vegas, NV (July 30- Aug 2, 2018).

[8] K. Whisnant, K. C. Gross and N. Lingurovska, "MSET Proactive Fault Monitoring in Enterprise Servers," *Proc. 2005 IEEE Intn'l Multiconference in Computer Science & Computer Eng.*, Las Vegas, NV (June 2005).