Fast Oscillation Detection and Labeling via Coarse Grained Time Series Data for ML Applications

Xin Xu, Chetan Mishra, Chen Wang,
Kevin D. Jones
Engineering Analytics & Modeling
Dominion Energy, Richmond, VA
chetan.mishra@dominionenergy.com

Luigi Vanfretti
School of Engineering
Rensselaer Polytechnic Instiute
Troy, NY
vanfrl@rpi.edu

Sean Murphy
PingThings, Inc.
Sacramento, CA
sean@pingthings.io

***Abstract*— Oscillation detectors use an RMS signal's energy in pre-specified frequency bands to detect the presence of different kinds of oscillations. While originally intended for situational awareness applications, they can be used for labeling the periods were they appear in large historical data sets. Labeled data sets are a requirement in machine learning applications. One such application is that of deriving associations between system operating conditions and the appearance of certain oscillations. However, there are two main challenges when realizing such application in practice. Firstly, a reliable detector requires properly set thresholds for the energy values. Secondly, scaling the algorithm to label multi-year historical data archives containing hundreds of terabytes of signals requires fast data access. The PingThings PredictiveGrid time series data platform deployed at Dominion Energy stores statistical averages of synchrophasor data at increasingly coarse resolutions as well as the original measurements. This work explores the use of coarse as opposed to full resolution PMU data for fast oscillation detection. Furthermore, K-means clustering is used to automatically determine energy thresholds based on the energy distribution of historical data in relevant frequency bands. Results using synchrophasor data from a STATCOM in the Dominion system expose a local mode.**

***Index Terms*—Machine learning, oscillation detection, synchrophasors, spectral analysis**

I. INTRODUCTION

The structure and dynamics of power grids all over the world are drastically changing due to the retirement of conventional generators and the proliferation of renewable generation. This has resulted in unprecedented challenges [1], which often require the help of fast-acting, power electronics-based assets such as STATCOMs, HVDC, etc. System operators, utilities, and owners usually do not have access to transparent models of these devices from the vendor due to proprietary technology. Therefore, measurement data in the form of synchrophasors plays a key role in helping uncover the dynamics of power electronic devices, as well as identifying dynamic performance issues that cannot be simulated with the models available to utilities.

Power system operations largely take place in ambient conditions and their dynamic response is normally linear and characterized by oscillations [2]. Therefore, in the frequency domain, the signal's energy content is concentrated in largely separable frequency bands. Moreover, power system data is dense in local dynamics i.e. most spectral peaks can be traced back to a specific device or a group of neighboring devices using techniques such as those in [3].

In day-to-day operations, the grid undergoes many changes, e.g. power system components such as generators and lines going in and out of service, internal switching in the components' control modes, etc. Naturally, these result in changes to the dynamic behavior of the system and, consequently, in the spectral content of various measurements. The ability to identify specific operational changes that produce a given oscillation can greatly help with locating its origin. This is particularly helpful in regions with poor sensor coverage, where the mode shape estimated from the available signals alone cannot be computed with sufficient space resolution to locate the source. For example, correlating an oscillation observed over a large area in the Dominion system to daylight hours [4] helped identify a solar PV farm as the source of a sustained oscillation as shown in Fig. 1. This can also often help explain the underlying phenomenon behind certain oscillations. Furthermore, this type of information can serve as a guide for making operating decisions that avoid undesirable dynamic responses or deterioriated performance.

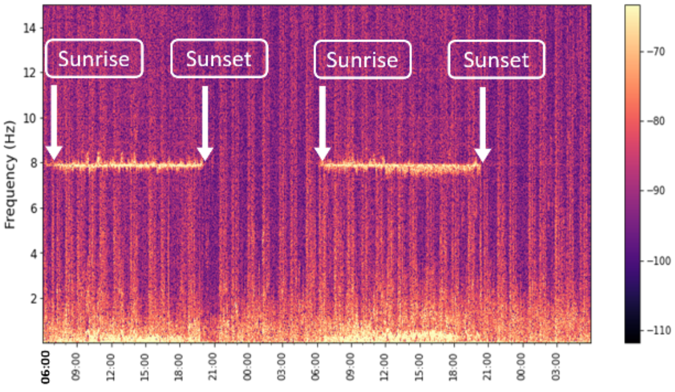

Figure 1. 8 Hz Solar PV Oscillation.

Deriving such associations is an extremely hard problem in practical systems due to the large number of variables at play, suggesting that machine learning (ML) approaches may be applicable. The idea of using ML to perform dynamic security assessments by mapping the operating conditions to stability dates back to the 1980's [5]. However, the problem is more challenging when only measurement data is available. The first and foremost requirement is to label periods where specific oscillations occur and, more importantly, to label a sufficient number of cases to create a rich training dataset.

In this regard, oscillation detectors have been developed over the past three decades. One of the earliest successful approaches in [6] monitors the envelope of the signal and compares it against a threshold for a specified amount of time before triggering an alarm. The RMS energy detector [7], which is an extension of [6], is one of the most widely accepted approaches in the industry [8] and therefore is exploited in this work. It operates by comparing the signal's RMS energy in a specific frequency band (containing the oscillation of interest) to a threshold. [9] extends this approach to detect subsynchronous oscillations. One drawback with existing approaches is that the threshold is set manually using baselining studies and experience. In this regard, [10] proposes a statistics-based threshold meant for online applications, where the thresholds are obtained using a short data window and succesivily updated. In the present work, we explore the use of clustering for arriving at a threshold that can be used for large-scale historical data sets.

The main bottleneck when labeling years' worth of historical synchrophasor data for the presence of oscillations is the speed and memory required to access the data. The consequent computations, e.g. spectrum calculation, only comprise a small portion of the overall processing time. Therefore, decreasing the time resolution of the data being accessed can substantially accelerate the process. Note that this will impact the frequency range of the dynamics that can be detected; in other words, decreasing the time resolution means that such approach is only applicable to slower dynamics. The PingThings PredictiveGrid platform stores precomputed, multi-resolution statistical summaries of the original time series data. Therefore, in this work, we investigate the spectral content of these summaries to justify using temporal aggregations instead of the original raw data for accelerated oscillation detection and labeling.

This paper is organized as follows. Section II provides background on the data platform and spectral analysis techniques used. Section III gives a brief overview of the application of K-means for oscillation detection on coarse data spectral results. The proposed approach is demonstrated on sychrophasor data from a STATCOM in the Dominion system and the result is shown in Section IV. Conclusions and future works are discussed in Section V.

II. Time Series Data Platform at Dominion Energy

A. *PingThings and the PredictiveGrid*

Dominion Energy uses the PingThings PredictiveGrid time series data platform for time series data ingestion, storage, visualization, and analysis on its transmission system, including for synchrophasor and continuous point on wave data. The PingThings platform is a scalable distributed computing system deployed in the cloud. It was architected to provide fast access not just to real time sensor data but also to all historical data up to at least 50 petabytes. Benchmarks indicate that the platform can read and write greater than ten million measurements per second per node with linear performance scaling characteristics. The platform is currently handling over 100 terabytes of data for Dominion from hundreds of PMUs with over one hundred thousand, 30 Hz data streams.

The PingThings platform uses a unique, purpose built database, the Berkley Tree Database (BTrDB) [11], to store time series measurements. The platform provides high-performance, temporally hierarchical access to stored data by using a novel data structure that guarantees consistent versioning, fast change-set identification, and multi-resolution statistical summaries of the raw high-resolution measurements, as explained below.

The data is stored in a time partitioned, tree data structure. The nodes at the bottom of the tree represent the original or raw measurements from a PMU or other sensor. Consequently, the depth of the tree is determined by the reporting rate f_s of the sensor, with up to 1 GHz per stream supported. Each node in the tree above the bottom captures a statistical summary (mean, min, max, count, and variance) of the child nodes below and an integer for data versioning. Nodes on the same level of the tree correspond to the same time resolution with the time spans represented increasing in size as we go higher in the tree. This multi-resolution storage can be exploited for efficient analytics, as illustrated in this paper.

B. *Spectral Content of Different Tree Levels*

The statistical summaries provided by the platform at different nodes or tree levels will be used for oscillation detection. The key tool for detection is the computation of the power spectral density (PSD). Thus, we illustrate the implications of computing the PSD on statistical summaries at exponentially increasing window lengths N of powers of 2, i.e. $N = 2^n \; \forall n \geq 0$. Let the data stored at n^{th} level be denoted by the ordered set $\{y_n(0), y_n(1) \dots\}$, where $n = 0$ denotes the original full-resolution data. Let the full resolution sampling rate be denoted by f_s. Consequently, the sampling rate for n^{th} level is given by $\frac{f_s}{2^n}$. In the Z domain,

$$Y_n(z) = \sum_i y_n(i) z^{-i}. \tag{1}$$

Now, the data stored at $(n+1)^{th}$ level is the data stored at n^{th} level undergoing a 2 sample moving average and downsampled by half. In the Z domain,

$$\begin{aligned} Y_{n+1}(z) &= \sum_i \frac{\left(y_n(i) + y_n(i+1)\right)}{2} z^{-i} \\ &= \frac{1}{4}\left(\left(1 + z^{1/2}\right) Y_n\left(z^{1/2}\right) + \left(1 - z^{1/2}\right) Y_n\left(-z^{1/2}\right)\right) \end{aligned} \tag{2}$$

The frequency domain estimate can be obtained for $(n+1)^{th}$ level by substituting $z = e^{(j\omega 2^{n+1})/f_s}$.

The data stored at higher levels in the tree are coarse grained and larger time spans can be queried faster than the data at low levels. Moreover, the coarse grained data can still preserve the spectral peaks up to their corresponding Nyquist frequencies, which enables fast data query for specral analysis when the oscillation frequency of interests is slow enough for the specific coarseness of the data. As an example, a piece of voltage magnitude data representing a 20 minutes time span is selected of which the highest sampling rate is 60 Hz. The spectral content of the data stored at different tree levels (or sampling rate) is shown in Figure *2*, which shows the spectral peaks are preserved up to their Nyquist frequencies. The resulting computational efforts (measured in seconds) at different sampling rate is visualized in Figure 3 which shows the time saving due to coarse grained data. Not only this has an impact on time, but also on resources. When using cloud based technologies, cost is based on resource consumption. A computation that completes more quickly is less expensive. Hence, in the rest of the paper, we will use coarse grained data for a faster, more resource efficient analysis.

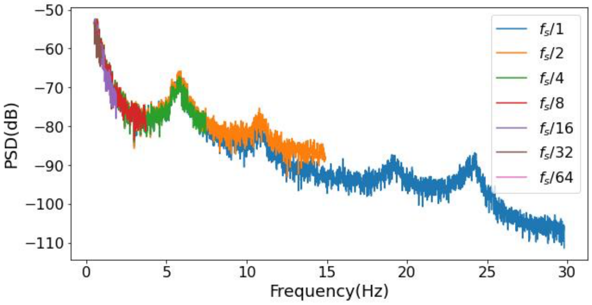

Figure 2. Spectral content at different sampling rates. In this example, the highest sampling rate is 60 Hz.

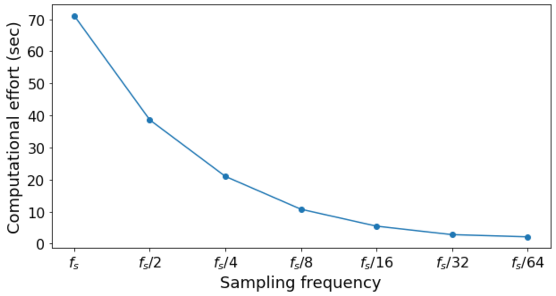

Figure 3. Computational effort at different sampling rates.

Note that these moving averages cannot be used for analyzing phase angle measurements, which requires full resolution data. Since the power system is never precisely at 60 Hz, the absolute phase angles change continuously and, therefore, often surpass the $\pm\pi$ radians limit. This condition is accompanied by a wrapping operation, introducing a $\pm 2\pi$ [12] offset in the data that significantly distorts the spectral content. One method to overcome this issue would be to store relative phase angles instead of the absolute angles. This could heavily reduce the likelihood of phase angle wrapping since, in a stable system, phase angle differences should not go beyond $\pm\pi$. Currently, the PingThings platform natively stores the absolute angles so we chose to focus our investigation on phasor magnitude data.

III. Oscillation Detection using K-Means Clustering

A. *Power Spectral Density*

Power Spectral Density (PSD) describes how the signal's power is distributed over frequency. Let $y(n)$ denote a stationary process with autocorrelation function,

$$r(k) = E_n(y^*(n)y(n+k)) \tag{3}$$

where, $E_n(\quad)$ is the expectation operator over n. PSD value at frequency f, denoted by $S(f)$ can be defined as,

$$\begin{aligned} S(f) &= \sum_{k=-\infty}^{\infty} r(k)e^{-j2\pi fk} \\ &= \lim_{N\to\infty} E\left(\frac{1}{N}\left|\sum_{n=0}^{N-1} y(n)e^{\frac{-j2\pi fn}{N}}\right|^2\right). \end{aligned} \tag{4}$$

In practice, given a sampled data window of finite length N, to obtain a robust (low variance) spectrum estimate, Welch's method [13] is popularly used. It divides the data window into smaller blocks, estimates the spectrum for each using a windowing function, and finally averages across all the blocks. In the present work, to process 20 minute windows, each block is of 2 minutes duration, processed using a Hanning window function with a 50% overlap between successive blocks.

B. *Methodology*

Practically, the RMS energy at a specific frequency f_s can be obtained by considering the maximum PSD within a narrow band around that frequency $[\,f_s - \varepsilon, f_s + \varepsilon\,]$, where ε is a small value like 0.02 Hz. Then, the detection of the presence of an oscillation mode is performed by observing the RMS energy, i.e. the mode is excited if the RMS energy increases and meets a certain criterion such as an explicitly defined threshold. The threshold can be determined by assuming the spectral density follows certain distributions. Alternatively, the approach proposed herein for oscillation detection is to use clustering algorithms to collect the computed RMS energy in different groups. The rationale behind this idea is that it is common for the RMS energy to have a steep change when the oscillation mode is present. Thus, clustering algorithms can be used to capture those steep changes and the identified clusters can be used to detect/predict the presence of the oscillation mode. In addition, when the oscillation appears it may reach different energy levels, the clustering algorithms can easily capture them by identifying more clusters, hence providing important insights into the oscillation mode of interests.

The clustering algorithm used in this paper is the K-means clustering algorithm. The procedure of this algorithm is introduced in the next subsection, together with other details to reliably identify the clusters.

C. K-means Clustering

K-means clustering, one of the most popular algorithms, aims to partition a set of observations into k clusters, where each observation belongs to the cluster with the nearest cluster centroid. When the number of clusters k is determined, the algorithm minimizes within-cluster variances by iterating: *i*) assigning observations (i.e. the RMS energy) to clusters based on the current centroids, and *ii*) choosing centroids based on the current assignment of data points to clusters. In his work, an RMS computation is considered to belong to a particular cluster if it is closer to that cluster's centroid than any other centroids.

The details are given as follows. Assume there are a set of observations $x^{(1)}, \ldots, x^{(m)}$, where each observation $x^{(i)} \in \Re^n$ is a vector of n features. After partitioning them into k cohesive clusters, each observation is assigned with a label $c^{(i)}$ indicating the associated cluster. The steps of the algorithm are as follows:

Step 1: Choose the number of clusters k. Initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \Re^n$ randomly.

Step 2: For each observation, set the label as:

$$c^{(i)} = argmin_j \left\| x^{(i)} - \mu_j \right\|^2 \tag{5}$$

where $x^{(i)}$ is the i-th RMS energy value computed at the desired coarsness.

Step3: For each cluster, set the centroid as:

$$\mu_j = \frac{\sum_{i=1}^{m} \{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \{c^{(i)} = j\}} \tag{6}$$

Step 4: Repeat Step 2 and 3 until convergence.

As K-means clustering does not guarantee to achieve an global optimal solution, the algorithm needs to be excecuted multiple times with randomly selected initial centroids. The best result minimizes the within-cluster-sum of squared errors (WSS):

$$WSS = \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2 \tag{7}$$

Another factor for obtaining the best clustering is to determine the optimal value of k in order to avoid either under or over estimating the number of clusters. The so-called elbow method provides a systematic way to achieve this outcome. This method starts by calculating the WSS for different values of k, and then, choosing the k where WSS does not substantially decrease. Plotting the variation of WSS versus k makes the elbow obvious via visual inspection, as shown in Fig. 8.

Up to this point, K-means clustering can be readily used to capture different energy levels, recall that each observation includes the RMS energy at the frequency of interest. Hence, when a new observation is obtained, it can be assigned to the cluster whose centroid is closest. However, it is possible that the new observation corresponds to a new energy level that has not been observed and is far away from any of the centroids. In that case, the analyst needs to mark the observation as undetermined and re-run the clustering algorithm with the new observation to find out if a new potential cluster has emerged.

IV. Case Studies

The oscillation observed in the current magnitude measurement around a STATCOM device deployed on the Dominion Energy's grid is investigated by applying spectral estimation on the statistical summaries of synchrophasor data from BTrDB, and then, applying the K-means clustering algorithm to identify the RMS energy levels and clusters. All synchrophasor data is stored in the PingThings platform, which also handles all computational processes. The original sampling rate of the data is at 30 Hz. As indicated by the spectrogram in Figure 4 for an entire day, the oscillation of interest is ~1 Hz and the value of the PSD for this mode is related to the reactive power output of the STATCOM as shown in Figure 5. Note that as the reactive power output increases, so does the value of the PSD at ~1 Hz.

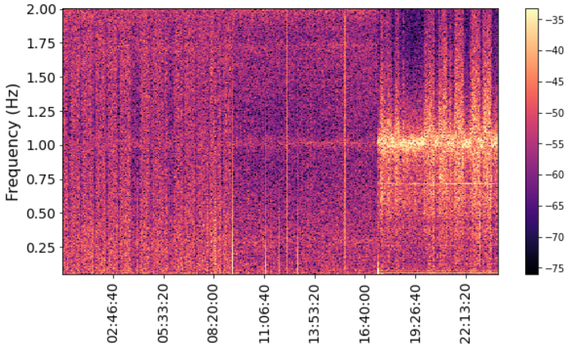

Figure 4. Spectrogram for one day of a single PMU data stream.

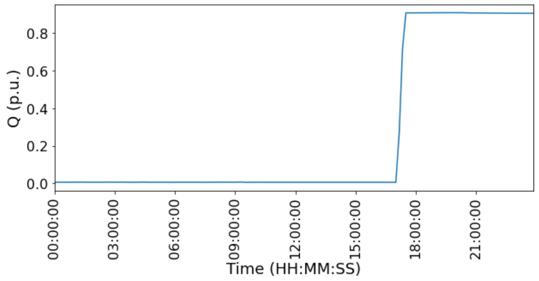

Figure 5. Reactive power output of the STATCOM

When investigating the RMS energy level of the 1 Hz mode, the reporting rate f_s of the data is decreased to 3.75 Hz. Observe that this is sufficient to capure the ~1 Hz mode in the frequency domain according to Shannon's sampling theorem.

Nex, we identify different oscillation levels for the ~1 Hz mode using thirty days of historical data. The PSD are calculated for every 10 minutes of data generating 4320 PSD estimates in total. A narrow band [0.98 Hz, 1.02 Hz] is selected for detecting the RMS energy. The variation of both the RMS energy and the reactive power output of the STATCOM are shown in Figure 6, which shows the high correlation between them.

The distribution of the samples of RMS energy is shown in the histogram of Figure 7, which also shows two potential clusters, i.e. $k = 2$. These were confirmed by applying the Elbow method, whose results are shown in Figure 8 where the "elbow" is at two. Setting $k = 2$ in the K-means algorithm, the identified clusters obtained are shown in Figure 7 marked in different colors (orange and blue), and the centroid of each cluster is indicated by the red bar.

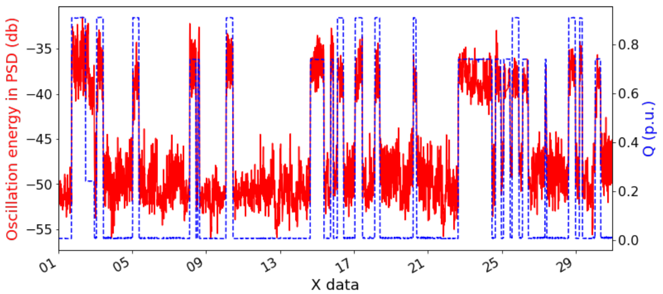

Figure 6. Variation of RMS energy of 1 Hz mode (in red curve) and the reactive power output of the STATCOM (in blue dashed curve).

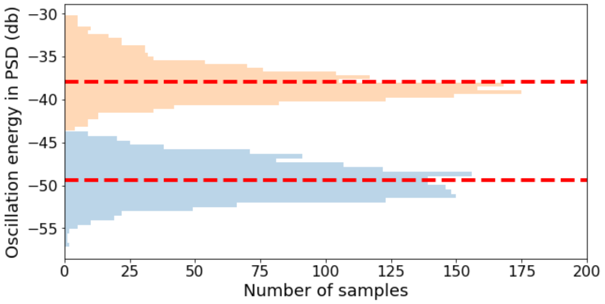

Figure 7. Distribution of the RMS energy and the identified centroids. The plot is transposed to show the PSD on the y-axis.

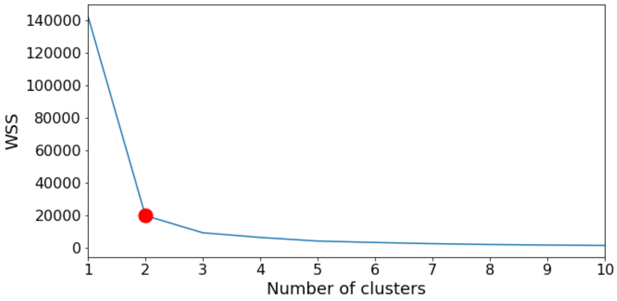

Figure 8. Elbow curve.

The result shows that when the reactive power output of the STATCOM changes to a non-zero value, the 1 Hz mode will appear and the RMS energy will be likely to stay around -48 db regardless the specific value of the reactive power. Hence, in this case it is reasonable to set a fixed threshold, say -44 db, to detect the appearance of the 1 Hz mode.

V. Conclusions and Future Works

This paper showed how fast data analytics applied to PredictiveGrid's summary statistics can allows to label data for the machine learning applications. A new method for labeling the presence of oscillation modes detected using RMS energy was developed using the K-means clustering. The proposed method was successfully applied to the measurement data from a STATCOM in Dominion Energy's grid. Future work includes the application of machine learning techniques leveraging the results presented in this paper and by taking advantage of the features of the PredictiveGrid platform.

References

[1] C. Mishra, A. Pal, J. S. Thorp, and V. A. Centeno, "Transient Stability Assessment (TSA) of Prone-to-Trip Renewable Generation (RG)-Rich Power Systems using Lyapunov's Direct Method," *IEEE Trans. Sustain. Energy*, pp. 1–1, 2019, doi: 10.1109/TSTE.2019.2905608.

[2] C. Mishra, L. Vanfretti, and K. Jones, "Power System Frequency Domain Characteristics for Inertia Estimation from Ambient PMU Data," presented at the 2021 IEEE Power & Energy Society General Meeting, Jul. 2021. doi: 10.13140/RG.2.2.16404.63363.

[3] D. J. Trudnowski, "Estimating Electromechanical Mode Shape From Synchrophasor Measurements," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1188–1195, Aug. 2008, doi: 10.1109/TPWRS.2008.922226.

[4] C. Wang, C. Mishra, K. D. Jones, and L. Vanfretti, "Identifying Oscillations Injected by Inverter-Based Solar Energy Sources in Dominion Energy's Service Territory," Apr. 13, 2021. [Online]. Available: https://naspi.org/sites/default/files/2021-04/D1S1_02_wang_dominion_naspi_20210413.pdf

[5] L. Wehenkel, T. Van Cutsem, and M. Ribbens-Pavella, "An artificial intelligence framework for online transient stability assessment of power systems," *IEEE Trans. Power Syst.*, vol. 4, no. 2, pp. 789–800, May 1989, doi: 10.1109/59.193853.

[6] J. F. Hauer and F. Vakili, "An oscillation detector used in the BPA power system disturbance monitor," *IEEE Trans. Power Syst.*, vol. 5, no. 1, pp. 74–79, Feb. 1990, doi: 10.1109/59.49089.

[7] M. Donnelly, D. Trudnowski, J. Colwell, J. Pierre, and L. Dosiek, "RMS-energy filter design for real-time oscillation detection," in *2015 IEEE Power Energy Society General Meeting*, Jul. 2015, pp. 1–5. doi: 10.1109/PESGM.2015.7286192.

[8] M. Donnelley, "Implementation and Operating Experience with Oscillation Detection at Bonneville Power Administration," p. 36.

[9] L. Vanfretti, M. Baudette, J.-L. Domínguez-García, M. S. Almas, A. White, and J. O. Gjerde, "A Phasor Measurement Unit Based Fast Real-time Oscillation Detection Application for Monitoring Wind-farm-to-grid Sub–synchronous Dynamics," *Electr. Power Compon. Syst.*, vol. 44, no. 2, pp. 123–134, Jan. 2016, doi: 10.1080/15325008.2015.1101727.

[10] J. Follum, J. Holzer, and P. Etingov, "A statistics-based threshold for the RMS-energy oscillation detector," *Int. J. Electr. Power Energy Syst.*, vol. 128, p. 106685, Jun. 2021, doi: 10.1016/j.ijepes.2020.106685.

[11] "BTrDB: Optimizing Storage System Design for Timeseries Processing | USENIX." Online:https://tinyurl.com/BTrDBpub (accessed Jul. 28, 2021).

[12] C. Mishra, L. Vanfretti, and K. D. Jones, "Synchrophasor Phase Angle Data Unwrapping Using an Unscented Kalman Filter," *IEEE Trans. Power Syst.*, pp. 1–1, 2021, doi: 10.1109/TPWRS.2021.3089027.

[13] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun. 1967, doi: 10.1109/TAU.1967.1161901.