A Measurement Source Authentication Methodology for Power System Cyber Security Enhancement

Yi Cui, *Member, IEEE*, Feifei Bai, *Member, IEEE,* Yong Liu, *Member, IEEE,* Yilu Liu, *Fellow IEEE*

***Abstract*—This letter proposes a spatial signature-based power system measurement source identification and authentication methodology. A Mathematical Morphology (MM) method is used to decompose the power system measurement signals and obtain the intrinsic components, which sparsity trends and roughness values are further derived to establish the Time-Frequency (TF) sparsity mapping. Then Random Forest Classification (RFC) is utilized to correlate the correct measurement source for each measurement signal based on the derived TF sparsity mapping. Experiment results using five phasor measurement units (PMU) has validated the effectiveness of this methodology.**

***Index Terms*— Cyber security, measurement, source authentication, spatial signature, phasor measurement unit (PMU).**

I. Introduction

Due to the increasing high-speed communication network architectures, the exposure of power system measurements to malicious cyber-attacks has increased dramatically. For example, source ID mix has emerged as a new type of highly-deceiving data spoofing attacks [1]. Such attack can mix the source IDs of measurement data from different locations without tampering the measurement values, thus deceiving the traditional cyber-attack detection techniques and paralyzing most measurement-based applications. Obviously, power system cyber security can be significantly enhanced if the source of a chuck of measurements data can be authenticated.

In this letter, a methodology to extract and utilize the spatial signatures of power system measurements for source identification or authentication will be introduced. Experiment results will also be presented to demonstrate the effectiveness of the proposed methodology. This letter is organized as follows: Section II introduces the spatial signatures of power system measurements; Section III explains the proposed source identification and authentication methodology; Section IV presents the experimental verification and further discussions on the performance of the proposed method are presented in Section V. The letter is concluded in Section VI.

II. Spatial Signature of Power System Measurements

Due to the stochastic variations of local grid conditions and unique local grid characteristics, the power system measurement at each location has a unique spatial signature, which can be used as a fingerprint for source authentication [2]. As illustrated in Fig. 1, though the frequency measured at three different locations share the same main trend due to the synchronicity required by the AC power system, slight differences can still be observed at the level of mHz if the measurement unit's accuracy and resolution are high enough. Considering the randomness of local grid condition variations and the uniqueness of local grid characteristics, this spatial signature can be practically considered to be unique for each location.

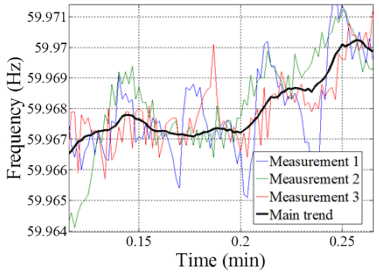
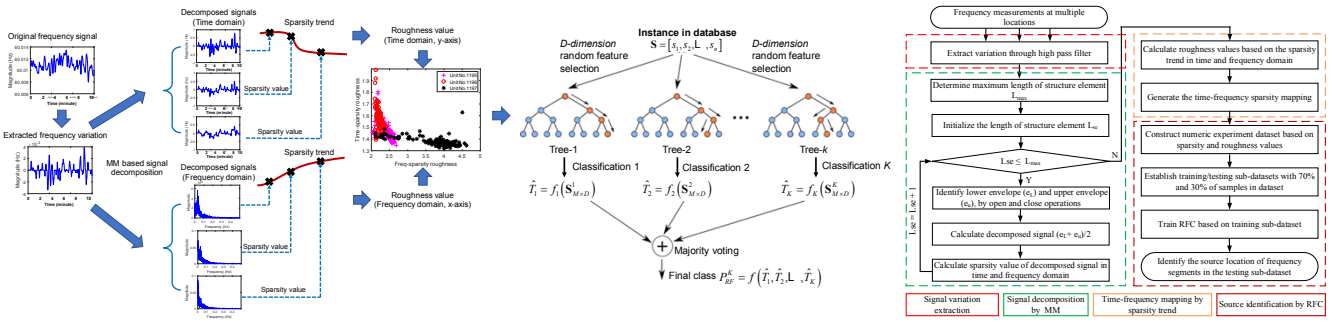


Fig. 1 Frequency measurements taken by FNET/GridEye at different locations

III. Source Authentication Methodology

Realizing the uniqueness of each power system measurement's spatial signature, a methodology has been developed to utilize it for measurement source authentication. Such methodology can automatically identify the source locations of the frequency measurements without the needs of concurrent power references. This methodology includes four major steps: firstly, a weighted high pass filter removes the common component of power system measurement signals, such as the main trend in Fig. 1; secondly, a Mathematical Morphology (MM) method [3] decomposes the measurement signals into a series of intrinsic components at multiple levels, where each component reflects the unique underlying nonlinearity and non-stationarity characteristics of the original signal; thirdly, the sparsity trend and roughness value of each intrinsic component are further calculated in both time and frequency domains to construct the Time-Frequency (TF) sparsity mapping, which will be used as the informative features of this measurement signal; finally, each measurement signal's TF sparsity mapping results are utilized by the Random Forest Classification (RFC) algorithm [4] to identify the correct source for each measurement signal. A flowchart of this methodology is presented in Fig. 2. Please note that there may exist other methods for signal decomposition, informative feature extraction and classification but MM, TF sparsity mapping and RFC are initially selected in this letter for good reasons. For example, the advantages of RFC are that it has a high robustness to the input data and it can overcome the over-fitting problem during the training process, thus producing a higher classification accuracy.



(a) (b)

Fig. 2 Flowchart of the proposed source authentication methodology (a) schematic diagram (b) flowchart.

IV. Methodology Verification

A. Experiment Setup

To verify the proposed source authentication methodology, five distribution-level phasor measurement units (PMUs) were installed in the same metropolitan area as shown in Fig. 3. The geographic distance between any two measurement locations is no more than a couple of miles. Thus, if the sources of the five measurement signals can be correctly identified in this experiment, the proposed methodology will prove to be accurate enough for almost all the transmission and distribution measurement source authentication purposes.

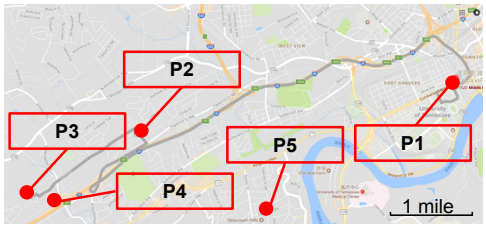


Fig. 3 Locations of installed distribution-level PMUs in Knoxville, TN, USA

B. Verification Results

The source identification of these five measurement signals was performed by the following steps. Firstly, for each measurement signal, it was divided into non-overlapping segments with a 10 minutes data window. The reporting rate of the data is 120Hz. Then 1000 segments from each measurement signal were randomly collected and processed by the proposed methodology to construct a TF sparsity mapping database. In this database, 70% samples were used as the training sub-dataset and the left 30% were used as the testing sub-dataset. Then ten-fold cross validation was conducted on the training sub-dataset to determine the optimal parameters of the RFC algorithm, such as the total tree number of the forest, minimal leaves of each decision tree and number of features which were randomly selected by each tree. Once the optimal values of the above parameters were obtained, RFC was trained on the whole training sub-dataset. Subsequently, the trained RFC algorithm was implemented to identify the source of samples in the testing sub-dataset. Please note that the above procedures (dataset separation, cross validation, training/testing) were repeated for 100 times and the match accuracies of the testing sub-dataset were recorded as the evaluation metrics in TABLE I.

From TABLE I, it can be seen that the overall source identification accuracy of these five measurement signals were as high as 96%. The overall identification accuracy was significantly higher compared with the author's previous experiment in [2] where the accuracy was less than 50%. It means the spatial signatures of measurement signals are unique enough even when two measurement locations are only several miles away and proves that the proposed methodology is accurate enough to identify or authenticate a measurement signal's source. But please note that this is under the assumption that the original measurement signal has high enough measurement accuracy and resolution so that the spatial signatures are captured in the original signals. For example, based on our observations so far, the grid frequency measurements should have at least an accuracy of 0.5mHz in order to successfully record the spatial signatures.

TABLE I
Confusion Matrix From Five Locations in the Same City

Accuracy (%)		P1	P2	P3	P4	P5	Overall
Identified by TF sparsity mapping	P1	**92**	8	0	0	0	**96.4**
	P2	5	**95**	0	0	0	
	P3	0	0	**100**	0	0	
	P4	0	0	5	**95**	0	
	P5	0	0	0	0	**100**	

V. Further Discussions

A. Discussions on the Training Dataset Construction

In this section, the RFC was trained by using the spatial characteristics database constructed from the 5-location (P1 to P5) frequency segments measured in June (700 segments for each location, each segment has 10 minutes data length). After the training, it was further implemented to recognize the source locations of the frequency segments measured in October (300 segments for each location, each segment is 10 minutes). TABLE II summarized the confusion matrix of the frequency segments in the testing sub-dataset. By comparing the results from TABLE II and TABLE I, it appeared the overall identification accuracy did not change significantly (above 95%), which demonstrated the spatial characteristics of the frequency measurements in the above five locations were stable that can be used for source identification.

TABLE II
Confusion Matrix From Five Locations in the Same City (Trained on June, Tested on October)

Accuracy (%)		P1	P2	P3	P4	P5	Overall
Identified by TF sparsity mapping	P1	**100**	0	0	0	0	**95.7**
	P2	10	**90**	0	0	0	
	P3	0	0	**100**	0	0	
	P4	0	0	5	**95**	0	
	P5	0	0	0	6.6	**93.4**	

B. Discussions on the Length and Reporting Rate for Source Authentication

In this section, the impact of length and reporting rate of the frequency measurement on the source identification was investigated. The measurements were manually down sampling from 120Hz to 60Hz, 30Hz, 10Hz and the length of segments varied from 1 second to 15 minutes. Fig. 4 showed the overall match accuracy of the measured frequency with different reporting rate and segment length by using TF sparsity mapping. From Fig. 4 it can be seen there was an increase trend in the overall match accuracy when the reporting rate increased. At each reporting rate, the overall match accuracy generally showed an increasing trend when the segment length increased from 1 second to 10 minutes and the maximum overall match accuracy can be attained by using 10 minutes segments. Further increase in the segment length will result in a decrease in the match accuracy. This is because the features (sparsity trend and roughness values) extracted from the measured signals will be averaged out when the duration of data increases beyond a certain length, which may lead to a decrease in the match accuracy. On the contrary, reducing the signal duration to less than 10 minutes will also lead to a reduction in the match accuracy due to the insufficient number of samples available for the source identification.

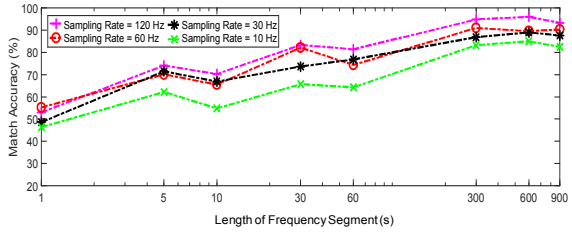


Fig. 4 Overall match accuracy of frequency measurements with different sampling rate and segment length.

C. Comparison with other Machine Learning Algorithms

In this comparative case study, another four computational intelligence algorithms, including k-nearest neighbour (kNN), support vector machine (SVM) with Gaussian kernel function, artificial neural networks (ANN) and C4.5 decision tree (DT), were applied to the same measurement database as mentioned in Section IV-A to identify the source locations. TABLE III summarized the overall identification accuracy among the four algorithms and the match accuracy for each location. Among the four algorithms, the SVM attained the highest identification accuracy (92.6%) while the ANN showed the minimal value (84.4%). It appeared the overall identification accuracy of the above four algorithms were all relatively lower than the RFC, which demonstrated the effectiveness of the proposed source authentication methodology.

TABLE III
COMPARISON OF IDENTIFICATION ACCURACY AMONG FOUR ALGORITHMS

Identification Accuracy (%)	P1	P2	P3	P4	P5	Overall
kNN	86	97	91	78	87	87.8
SVM	85	96	92	96	94	92.6
ANN	87	67	95	84	89	84.4
DT	94	90	83	94	92	90.6

D. Discussions on the Applicability of TF Sparsity Mapping on Real Time Measurement Authentication

Fig. 5 shows the framework of applying the proposed method for real time measurement authentication. It mainly contains three steps:

(1) Data pre-processing and feature extraction: Sufficient number of historical frequency segments from multiple locations were collected to construct the training database. Each measurement is pre-processed to check the continuity of the raw data and eliminate the outliers in the original segment. Then the TF sparsity mapping was applied to extract the spatial signatures from the segments in the training dataset.

(2) Offline training: In this step, the RFC algorithm makes use of a historic frequency measurements dataset to construct a mathematical model that approximates the relationship between the spatial characteristics (i.e., features) and the source locations of the frequency segments. It should be mentioned that since the spatial signatures are usually stable, the training process does not have to be performed frequently or in real time manner.

(3) Real time measurement authentication: Upon the arrival of new frequency measurements, the above trained model can make a quick classification (in several milliseconds) on the source location of the new frequency signals of interest into one of the categories in the above historic training dataset.

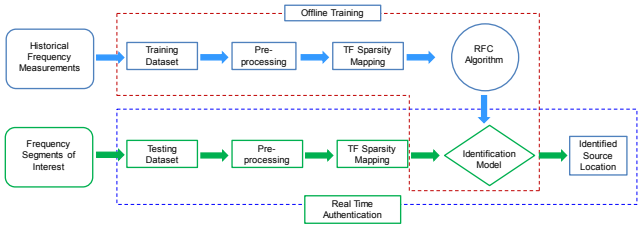


Fig. 5 Framework of real time application for measurement authentication.

VI. CONCLUSION

In this letter, a spatial signature based methodology, which is a combination of MM signal decomposition, TF sparsity mapping and RFC algorithm, was proposed to extract the informative features of each measurement signal's spatial signatures and use them for source identification. The experimental results showed that the accuracy and robustness of the proposed methodology is high enough for the power system cyber security enhancement.

REFERENCES

[1] H. Lin, Y. Deng, S. Shukla, J. Thorp and L. Mili, "Cyber security impacts on all-PMU state estimator - a case study on co-simulation platform GECO," in *Proceedings of International Conference on Smart Grid Communications*, 2012,Tainan, Taiwan, pp. 587-592.

[2] W. Yao, J. Zhao, M. J. Till, S. You, Y. Liu, Y. Cui and Y. Liu, "Source location identification of distribution-level electric network frequency signals at multiple geographic scales," *IEEE Access*, vol.5, pp. 11166-11175, 2017.

[3] J. L. Wu, T. Y. Ji, M. S. Li, P. Z. Wu and Q. H. Wu, "Multistep wind power forecast using mean trend detector and mathematical Morphology-Based local predictor," *IEEE Trans. Sustainable Energy*, vol.6, Issue 4, pp. 1216-1223, 2015.

[4] L. Breiman, "Random forest," *Mach. Learn.*, vol.45, Issue 1, pp. 5-32, 2001.