IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 11, 2018, accepted September 12, 2018, date of publication September 20, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871514

Physical Layer Authentication Enhancement Using a Gaussian Mixture Model

XIAOYING QIU[1], (Student Member, IEEE), TING JIANG[1], SHENG WU[1], (Member, IEEE), AND MONSON HAYES[2], (Life Fellow, IEEE)

[1]Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030, USA

Corresponding author: Xiaoying Qiu (qxy@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Project 61671075 and in part by the Major Program of National Natural Science Foundation of China under Grant 61631003.

ABSTRACT Wireless networks strive to integrate information technology into every corner of the world. This openness of radio propagation is one reason why holistic wireless security mechanisms only rarely enter the picture. In this paper, we propose a physical (PHY)-layer security authentication scheme that takes advantage of channel randomness to detect spoofing attacks in wireless networks. Unlike most existing authentication techniques that rely on comparing message information between the legitimate user and potential spoofer, our proposed authentication scheme uses a Gaussian mixture model (GMM) to detect spoofing attackers. Probabilistic models of different transmitters are used to cluster messages. Furthermore, a 2-D feature measure space is exploited to preprocess the channel information. Training data for a spoofer operating through an unknown channel, a pseudo adversary model is developed to enhance the spoofing detection performance. Monte Carlo simulations are used to evaluate the detection performance of the GMM-based PHY-layer authentication scheme. The results show that the probability of detecting a spoofer is higher than that obtained using similar approaches.

INDEX TERMS PHY-layer authentication, Gaussian mixture model, spoofing detection, wireless security.

I. INTRODUCTION

While open radio propagation channels allow for ''anybody, anywhere, anytime'' wireless access, they are subject to security vulnerabilities [1]. Wireless communication systems, for example, are prone to spoofing attacks where an adversary masquerades as a legitimate device. In order to secure such communication channels, signal authentication is necessary before a received message is processed. Although conventional upper-layer security mechanisms have been proposed to foil spoofer attacks, they typically require significant resources and computational power. Digital key distribution and management in dynamic networks is difficult, and current security protocols do not take into account the novel avenues for intrusion. When physical layer attributes are not considered, it becomes more difficult to detect an intruder using unauthorized digital keys because the device's identification and access rights are only verified through security keys [2].

Physical (PHY)-layer authentication may be used to complement security-based approaches by exploiting the dynamic characteristics of the physical layer. Since PHY-layer authentication takes advantage of the unique communication environment of a device, it is more difficult for the spoofer to masquerade as the authentic device. According to the well-known Jakes uniform scattering model, received signals are rapidly decorrelated over a distance of approximately half a wavelength [3]. As a result of this decorrelation, it is possible for a receiver to authenticate a sender from the received channel information vector. Traditional cryptography-based authentication is ill-suited for wireless networks where many terminals have limited energy and computing power, such as cellular Internet of Things networks [4] and body area networks [5]. However, PHY-layer authentication provides an effective solution to simplify authentication without incurring additional computational overhead by taking advantage of the signatures of the wireless channels.

There are three general approaches for PHY-layer authentication: waveform embedding [6], hardware-based authentication [7], and channel-based authentication [8], [9]. In this paper, we introduce a new approach for channel-based authentication that is based on a Gaussian Mixture Model (GMM) for environment-dependent radiometric features. The following section provides a brief overview of previous approaches to channel-based authentication and an overview of our approach for PHY-layer authentication.

II. BACKGROUND

Over the past several years, there has been a considerable amount of research on PHY-layer authentication [9]–[15]. Received signal strengths [2], channel impulse responses [14] and channel state information [3] have been used as the fingerprints of wireless channels to detect spoofing attacks. The feasibility of PHY-layer spoofing detection based on sparse signal processing has also been demonstrated in [11]–[13]. Feature extraction and fusion were used to make a distinction between legitimate transmitters and spoofers. The preprocessing of channel information reinforces the characteristics of the signal. Signal processing and feature recognition, if done successfully, will greatly enhance PHY-layer authentication. A channel-based PHY-layer authentication method was developed in [14] that quantizes the channel amplitude and path delay into one-bit features, exploiting the spatial independence of these features for different transmitter-receiver pairs. A privacy-preserving proximity-based security system was also proposed for mobile users that exploits the packet arrival time and the received signal strengths [28]. In [3] power spectral densities have also been used to distinguish different users, thereby denying access to potential attackers. In addition, multiple antennas were exploited to provide further authentication security for wireless systems and to reduce the spoofing detection error rate [15], [16].

Although many channel-based authentication approaches show promising results, they typically rely on a simple threshold to detect a spoofer, and it is a challenge for a device to flexibly and efficiently choose a threshold for effective security. Moreover, the communication environment and node mobility are factors that significantly affect the test threshold. A PHY-layer spoofing detection method has been proposed that is based on reinforcement learning to achieve an optimal test threshold [17]. The interaction between a legitimate receiver and a spoofing transmitter is formulated as a zero-sum PHY-layer authentication game. Learning-based spoofer detection developed in [23] improves the secure capacity in the presence of smart jamming. In [24], model learning based on tree structure is proposed to increase the learning speed in stochastic environments, and in [18] a classification algorithm based on extreme machine learning is presented. Machine learning is a powerful mathematical tool that enables PHY-layer authentication without requiring a fixed threshold, and there has been a variety of important studies on the use of machine learning techniques in wireless network intrusion detection [19], [20].

In this paper, we propose a Gaussian Mixture Model (GMM)-based PHY-layer authentication enhancement scheme that exploits the characteristics of the channel to detect a spoofer at an intended receiver. Although a Gaussian Mixture Model for spoofer detection using channel estimates has been considered before [21], channel estimation errors significantly impact the reliability of this approach. Therefore, in order to mitigate the effects of these estimation errors, we base our approach on pre-processing the channel variations along with using multi-dimensional features. One of the major differences between the proposed scheme and existing work [17]–[19], [21] is that the characteristics of the radio channels for the legitimate user and a pseudo adversary are exploited and modeled mathematically. More specifically, in the following a PHY-layer authentication scheme is proposed that exploits channel state information to detect spoofing attacks and formulate the authentication process between the intended receiver and the transmitter as a hypothesis test problem. In contrast to previous work using channel information [19], [21], the PHY-layer authentication problem is formulated as a comparison to determine whether random signals have similar two-dimensional feature vectors. A Gaussian Mixture Model is used in the PHY-layer authentication method to measure the *similarity* of channel information vectors and to determine an output target label. A soft decision is used to make a decision on received data packets from which transmitter. In order to model the channel of the potential spoofer, a pseudo adversary model is proposed. The Gaussian Mixture Model for legitimate user and pseudo adversary is established via learning. The proposed probabilistic model requires a small amount of training data from legitimate users and can achieve high detection accuracy.

The rest of the paper is organized as follows. In Section III, we present the system model and formulate the problem that is to be solved. In Section IV, the feature vector that is used to discriminate between the legitimate user and the spoofer is defined along with the pseudo adversary model. In Section V, the GMM-based PHY-layer authentication for spoofing detection is presented. In Section VI, simulations are given that illustrate the proposed GMM-based authentication approach via simulations. Finally, in Section VII, conclusions are presented.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

The basic channel model is shown in Fig. 1. In this model, there are three different parties: a legitimate transmitter Alice, an intended receiver Bob, and a malicious transmitter Eve who masquerades as Alice with a fake MAC address. Based on information determined about the channel, the goal is to authenticate a message that is received by Bob and determine whether or not it is from Alice. In our approach, it is assumed that the initial transmission from Alice to Bob has been authenticated prior to Eve's arrival [1]. This may be done using a standard high-layer protocol that confirms that the first message is sent by Alice [6]. It is also assumed that the spoofing adversary knows when to begin sending false messages to Bob.

The channel state information that is used for authentication is estimated from the pilot or preamble symbols.

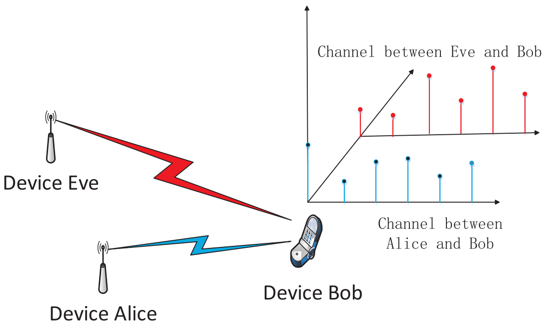


FIGURE 1. Illustration of a PHY-layer authentication scheme consisting of the legitimate transmitter Alice, the spoofing attacker Eve, and the intended receiver Bob.

Let $\boldsymbol{H}_A(k)$ denote the channel state information vector for the channel between Alice and Bob at time k, which may be written as

$$\boldsymbol{H}_A(k) = [h_{A,0}(k), h_{A,1}(k), \cdots, h_{A,M-1}(k)] \tag{1}$$

where M is the number of subcarrier frequency samples, and $h_{A,m}(k)$ are the samples. Similarly, for the channel between Eve and Bob, the channel state information at time k will be denoted by $\boldsymbol{H}_E(k)$.

When measuring the channel vector from the received symbols, the measurements will generally contain estimation errors. Therefore, the measured channel vector is modeled as

$$\boldsymbol{H}_A(k) = \underline{\boldsymbol{H}}_A(k) + \boldsymbol{\Delta}(k) \tag{2}$$

where $\underline{\boldsymbol{H}}_A(k)$ is the "true" channel vector and $\boldsymbol{\Delta}(k)$ is the channel estimation error, which is assumed to be a vector of uncorrelated complex zero-mean Gaussian random variables with variance σ_Δ^2 [17], [29]. The true channel $\underline{\boldsymbol{H}}_A(k)$, is modeled as a Rayleigh fading channel,

$$\underline{\boldsymbol{H}}_A(k) \sim \mathcal{CN}(0, \sigma_A^2 \boldsymbol{I}) \tag{3}$$

where $\mathcal{CN}(\cdot)$ is a complex Gaussian random vector and σ_A^2 is the average power gain of Alice's channel [15]. Therefore, the estimated channel vector is

$$\boldsymbol{H}_A(k) \sim \mathcal{CN}(0, (\sigma_A^2 + \sigma_\Delta^2)\boldsymbol{I}) \tag{4}$$

When Bob receives a new message at time $k+1$, the channel information vector for this transmission, $\boldsymbol{H}_T(k+1)$, is estimated. Given $\boldsymbol{H}_T(k+1)$, Bob must then determine whether the received message is from Alice or from Eve. In the PHY-layer authentication approach that is proposed, the following assumptions are made:

Assumption 1: Alice's and Eve's channels are uncorrelated.

This assumption is based on the property that in an environment full of scatterers and reflectors, the channel response decorrelates rapidly as the terminal positions change by the order of a wavelength or more [22]. For 5G wireless networks, this corresponds to 6 cm. Since, in any practical communication environment the spoofer will not be very close to the legitimate user (in units of wavelength), then the two channel information vectors will be uncorrelated. Therefore, it is assumed that $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_E(k)$ are uncorrelated.

Assumption 2: The channel information vector for two successive transmissions (packets) from the same transmitter are correlated.

In a typical cellular system such as 3G networks, the channel coherence time is on the order of tens of milliseconds [30], and in wireless body area networks the average channel coherence time is 48 milliseconds when walking and 31 milliseconds when running [31]. With a time slot of duration 1.67 milliseconds in the 3G system standard, multiple time slots can be grouped together to form a frame with a duration consistent with the coherence time of the network [30]. Therefore, it is reasonable to assume that $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_A(k+1)$ will change little from one time slot to the next over a frame, and will therefore be correlated throughout the frame. In the case of static channels and static transmitters, $\boldsymbol{H}_A(k+1) = \boldsymbol{H}_A(k)$.

B. PROBLEM FORMULATION

Given the channel information vector $\boldsymbol{H}_A(k)$ at time k that corresponds to a transmission from Alice to Bob, with the next transmission the channel information vector is estimated, $\boldsymbol{H}_T(k+1)$, and the goal is to determine whether $\boldsymbol{H}_T(k+1)$ corresponds to a transmission from Alice or Eve. In other words, Bob seeks to authenticate the identity of the transmitter based on the characteristics of the channel. Given the two assumptions above, the PHY-layer authentication process may be formulated as a hypothesis test problem as follows:

$$H_0 : \boldsymbol{H}_T(k+1) = \boldsymbol{H}_A(k+1) \tag{5}$$

$$H_1 : \boldsymbol{H}_T(k+1) = \boldsymbol{H}_E(k+1) \tag{6}$$

where the null hypothesis H_0 is that the message belongs to Alice, and the alternative hypothesis H_1 is that the received message comes from the adversary, i.e., there is a spoofing attacker trying to send messages masquerading as Alice. Since Alice and Eve use different channels, then $\boldsymbol{H}_A(k+1) \neq \boldsymbol{H}_E(k+1)$ and the hypothesis test involves determining whether $\boldsymbol{H}_T(k+1)$, is closer to $\boldsymbol{H}_A(k+1)$ or to $\boldsymbol{H}_E(k+1)$. In order to measure the *similarity* between a pair of channels, a two-dimensional feature vector described in the following section is used to construct the test statistics that will be used to solve the hypothesis test problem.

IV. FEATURE EXTRACTION

In this section, a two-dimensional feature vector is defined that is used for PHY-layer authentication. A pseudo adversary model for the spoofer is also presented, which will be important for the authentication approach that is proposed.

A. TWO-DIMENSIONAL FEATURE SPACE

Given the channel information vector $\boldsymbol{H}_A(k)$ for a legitimate transmitter at time k, and $\boldsymbol{H}_T(k+1)$ from an unknown transmitter at time $k+1$, two features are measured that will

TABLE 1. Summary of important symbols.

Symbol	Definition
M	Number of subcarrier frequency
$\underline{\boldsymbol{H}}_i(k)$	Channel information vector for transmitter i at time k
$\boldsymbol{H}_i(k)$	Estimated channel vector
$\boldsymbol{\Delta}(k)$	Channel estimation error
$\overline{\boldsymbol{H}_i(k)}$	Mean of the vector $\boldsymbol{H}_i(k)$
σ_i^2	Average power gain from the transmitter i
T	Transmitter under test
D	Distance between two channel vectors
R	Sample Pearson correlation coefficient
$\boldsymbol{F} = [D, R]$	Two-dimensional channel feature vector
ς	Channel correlation coeffficient
v	Speed of the node
λ	RF wavelength
t	Symbol duration
$\mathcal{N}(\cdot)$	A multivariate Gaussian density
$\mathcal{CN}(\cdot)$	Complex multivariate Gaussian density
μ_j, Σ_j	Mean and covariance
π_j	Weights in Gaussian Mixture Model
$\boldsymbol{x}$	Feature vector extracted from $\boldsymbol{F}$
$\boldsymbol{X}$	Training samples of the channel feature vectors
$p(j\|x_i, \lambda)$	Probability that x_i belongs to the j^{th} component

be used for authentication. The first feature is the Euclidean distance between the vectors $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_T(k+1)$,

$$D_T(k) = \|\boldsymbol{H}_A(k) - \boldsymbol{H}_T(k+1)\| \tag{7}$$

and the second is the sample Pearson correlation coefficient,

$$R_T(k) = \frac{\left\langle \boldsymbol{H}_A(k) - \overline{\boldsymbol{H}_A(k)}, \boldsymbol{H}_T(k+1) - \overline{\boldsymbol{H}_T(k+1)} \right\rangle}{\left\|\boldsymbol{H}_A(k) - \overline{\boldsymbol{H}_A(k)}\right\| \left\|\boldsymbol{H}_T(k+1) - \overline{\boldsymbol{H}_T(k+1)}\right\|} \tag{8}$$

where $\overline{\boldsymbol{H}_A(k)}$ and $\overline{\boldsymbol{H}_T(k+1)}$ are the means of $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_T(k+1)$, respectively. These two features form a two-dimensional feature vector,

$$\boldsymbol{F}_T(\boldsymbol{H}_A(k), \boldsymbol{H}_T(k+1)) = [D_T(k),\ R_T(k)]\ . \tag{9}$$

that will be used to make a decision on whether $\boldsymbol{H}_T(k+1)$ corresponds to a transmission from Alice or Eve.

B. CHANNEL VARIATIONS

A key to the success in any machine learning algorithm is the acquisition of training data. With a channel-based PHY-layer authentication method, training data is necessary to develop channel models for the legitimate user as well as for a potential spoofer. For legitimate users, training data is relatively easy to obtain since the channel information obtained in the previous frame may be used [17]. According to Assumption 2, the channel information between Alice and Bob is correlated for successive pilots tones. In the ideal case of a static channel, $\underline{\boldsymbol{H}}_A(k+1)$ will be constant,

$$\underline{\boldsymbol{H}}_A(k+1) = \underline{\boldsymbol{H}}_A(k) \tag{10}$$

but in real-world applications, the channel will not be static due to the communication environment and node mobility. To account for this, the channel model is [18]

$$\underline{\boldsymbol{H}}_A(k+1) = \varsigma \underline{\boldsymbol{H}}_A(k) + \sqrt{(1-\varsigma^2)\sigma_A^2}\,\boldsymbol{w}(k) \tag{11}$$

where $\boldsymbol{w}(k)$ is a zero-mean unit variance complex Gaussian random vector that is independent of $\underline{\boldsymbol{H}}_A(k)$. The parameter ς is the correlation coefficient, which is a function of the RF wavelength λ, the speed of the terminal, v, and the symbol duration t. More specifically,

$$\varsigma = J_0(2\pi vt/\lambda) \tag{12}$$

where J_0 is a zeroth-order Bessel function of the first kind [22]. Note that when Alice moves slowly, $vt/\lambda \approx 0$ and $\varsigma \approx 1$, and the channel is approximately constant. It is clear that the amount of mobility in Alice is reflected in the value of ς, i.e., the faster that Alice moves the smaller that ς becomes.

Since the location of the spoofing attacker is unknown, and there is no information available about the spoofer's channel, then a PHY-layer classifier would make a decision as to whether $\boldsymbol{H}_T(k+1)$ is from Alice or Eve based on how similar it is to $\boldsymbol{H}_A(k+1)$. Therefore, to improve the performance of the PHY-layer classifier, a pseudo adversary is used to generate training samples for the spoofer's channel [18]. More specifically, the channel of a pseudo adversary at time $k+1$, denoted by $\boldsymbol{H}_E^P(k+1)$, is modeled as

$$\boldsymbol{H}_E^P(k+1) \sim \mathcal{CN}(0, \sigma_E^2 \boldsymbol{I}) \tag{13}$$

where it is assumed that $\boldsymbol{H}_E^P(k+1)$ is independent of $\underline{\boldsymbol{H}}_A(k)$ [15]. Equation (13) is then used to generate two-dimensional channel features of the spoofing attacker.

Since it is assumed that that $\underline{\boldsymbol{H}}_A(k)$ and $\boldsymbol{H}_E^P(k+1)$ are independent Gaussian random vectors, then the statistic

$$D_E^P(k) = \left\|\underline{\boldsymbol{H}}_A(k) - \boldsymbol{H}_E^P(k+1)\right\| \tag{14}$$

has a chi distribution with M degrees of freedom [18]. Therefore, it follows that the expected value of $D_E^P(k)$ is

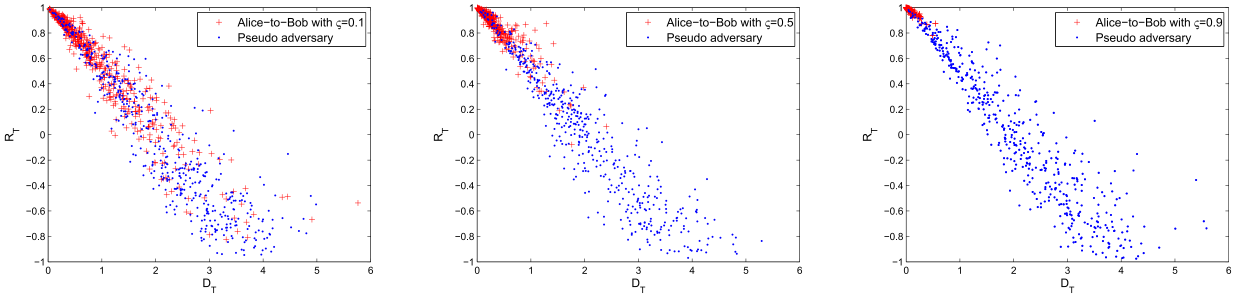
$$E\left\{D_E^P(k)\right\} = \sqrt{2}\frac{\Gamma((M+1)/2)}{\Gamma(M/2)} \tag{15}$$

where $\Gamma(\cdot)$ is the Gamma function, and the variance is

$$\text{var}\left\{D_E^P(k)\right\} = M - E^2\left\{D_E^P(k)\right\} \tag{16}$$

For the sample Pearson correlation coefficient between $\underline{\boldsymbol{H}}_A(k)$ and $\boldsymbol{H}_E^P(k+1)$, the density function for $R_E^P(k)$ can be written as [18]

$$f(R_E^P(k)) = \alpha\left[1 - (R_E^P(k))^2\right]^{(M-4)/2} \tag{17}$$



(a) (b) (c)

FIGURE 2. Training samples of the channel feature vector for the legitimate and pseudo adversary channels for three different correlation coefficients with $\sigma_A^2 = 1$ and $\sigma_E^2 = 0.8$. (a) Channel data for $\varsigma = 0.1$. (b) Channel data for $\varsigma = 0.5$. (c) Channel data for $\varsigma = 0.9$.

where α is a constant. Note that since $R_E^P(k)$ is bounded by one,

$$-1 \leq R_E^P(k) \leq 1, \tag{18}$$

and $f(R_E^P(k))$ is symmetric, then it follows that the expected value of $R_E^P(k)$ is zero,

$$E\left\{R_E^P(k)\right\} = 0 \tag{19}$$

It may be shown [18] that the variance of $R_E^P(k)$ is

$$\text{var}\left\{R_E^P(k)\right\} = \frac{1}{M-3}. \tag{20}$$

Examples of the two components of the feature vector $\boldsymbol{F}$ for the legitimate and pseudo adversary channels are shown in Fig. 2 for three different values of the correlation coefficient in Eq. (11), and with $\sigma_A^2 = 1$ and $\sigma_E^2 = 0.8$. As described in the next section, these correspond to training samples that would be used to train a GMM to cluster the data and perform authentication.

V. THE PROPOSED SECURITY FRAMEWORK

In this section, a GMM-based security authentication technique is proposed for the detection of spoofers. Since the method of spoofing detection is based on clustering using a GMM, we begin first with a description of the Expectation Maximization (EM) algorithm that is used to find the parameters of the GMM. Then we describe how the mixture model is used to identify deviations in the received data packets in order to detect a potential spoofer.

A. EM ALGORITHM

In practical communication scenarios, machine learning may be used as an effective approach to cluster messages that are received in a wireless network, and to determine the originator of received data packets. Here, a GMM is used to estimate the probability density function and the posterior probability of the feature vector $\boldsymbol{F}(\boldsymbol{H}_A(k), \boldsymbol{H}_T(k+1))$. Once these are known, it is then possible to determine how likely it is that $\boldsymbol{H}_T(k+1)$ represents the channel from Alice to Bob or that of a spoofer.

A GMM is a weighted sum of K Gaussians,

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{21}$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, and where the weights (mixture coefficients), π_k, are non-negative and sum to one,

$$\sum_{k=1}^{K} \pi_k = 1. \tag{22}$$

Given a set of training samples, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, that are drawn independently from $\boldsymbol{F}(\boldsymbol{H}_A(k), \boldsymbol{H}_A(k+1))$ and $\boldsymbol{F}(\boldsymbol{H}_A(k), \boldsymbol{H}_E^P(k+1))$ the goal is to estimate the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ along with the mixture coefficients π_k. This is done by maximizing the log-likelihood function,

$$\begin{aligned} \ln P(\boldsymbol{X}|\lambda) &= \ln \prod_{n=1}^{N} P(\boldsymbol{x}_n|\lambda) \\ &= \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \end{aligned} \tag{23}$$

where λ is the collection of all parameters into a single parameter,

$$\lambda = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, \quad i = 1, 2, \cdots, K \tag{24}$$

Note that (23) is a non-linear function of the parameter λ. In our PHY-layer authentication approach, the parameters are estimated by maximizing the log-likelihood function using the Expectation Maximization (EM) algorithm [25]. Given the training samples and an initial set of model parameters, the EM algorithm involves two steps:

1) EXPECTATION STEP

In this step, the membership probabilities for all N training samples are computed,

$$p(j|\boldsymbol{x}_i, \lambda) = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum\limits_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{25}$$

where $p(j|\boldsymbol{x}_i, \lambda)$ is the probability that sample $\boldsymbol{x}_i$ belongs to the j^{th} mixture component.

2) MAXIMIZATION STEP

In this step, the parameters π_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are updated. Specifically, the weighting coefficients are updated as follows,

$$\pi_j^{new} = \frac{1}{N}\sum_{i=1}^{N} p(j|\boldsymbol{x}_i, \lambda). \tag{26}$$

Then the mean and covariance values are updated:

$$\boldsymbol{\mu}_j^{new} = \frac{\sum_{i=1}^{N} p(j|\boldsymbol{x}_i, \lambda)\boldsymbol{x}_i}{\sum_{i=1}^{N} p(i|\boldsymbol{x}_i, \lambda)} \tag{27}$$

and

$$\boldsymbol{\Sigma}_j^{new} = \frac{\sum_{i=1}^{N} p(j|\boldsymbol{x}_i, \lambda)(\boldsymbol{x}_i - \boldsymbol{\mu}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_i)^T}{\sum_{i=1}^{N} p(j|\boldsymbol{x}_i, \lambda)}. \tag{28}$$

The EM algorithm iteratively updates the parameter estimated, and at each step of the iteration the log-likelihood function may be shown to be increasing. In addition, under some mild continuity conditions, the EM algorithm is guaranteed to converge to a local maximum [25], [26].

The idea now is that for each new message that is received, the channel vector $\boldsymbol{H}_T(k+1)$ is estimated, the feature vector $\boldsymbol{F}(\boldsymbol{H}_A(k), \boldsymbol{H}_T(k+1))$ is computed, and a decision is made as to whether the received message is from Alice or a spoofer by using the GMM to choose the one that maximizes the posterior probability.

B. GMM-BASED PHY-LAYER AUTHENTICATION

We now illustrate how the GMM is used to make a decision on which transmitter generates a received message. The proposed GMM-based PHY-layer authentication mechanism includes three phases: channel-based feature vector generation, GMM parameters initialization and model training, and GMM-based spoofing detection. Fig. 3 shows the process and all of the steps in our two-dimensional feature-based GMM authentication approach. In the following, each step is described in more detail.

1) CHANNEL-BASED FEATURE VECTOR GENERATION

After a transmission is received by Bob, and the channel parameter vector $\boldsymbol{H}_T(k+1)$ is estimated, the next step is to extract features that will be used to authenticate the received message. Specifically, a two-dimensional feature vector is formed that consists of the Euclidean distance between $\boldsymbol{H}_T(k+1)$ and $\boldsymbol{H}_A(k)$ and the sample Pearson correlation between these same two vectors. Distilling the channel information down to these two features helps to mitigate the effects of channel estimation errors and, as we will see, forms an effective characterization of the channel for authentication.

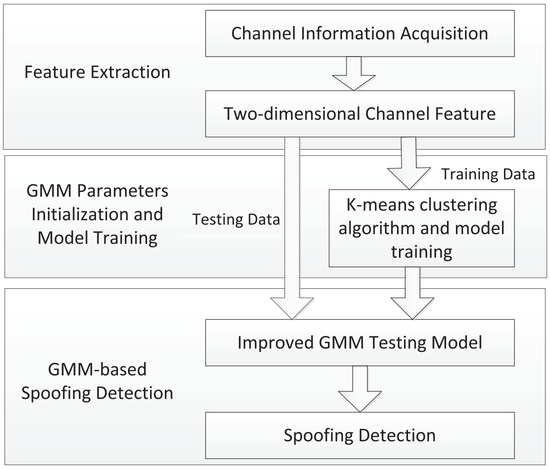

Feature Extraction
Channel Information Acquisition
Two-dimensional Channel Feature
GMM Parameters Initialization and Model Training
Testing Data
Training Data
K-means clustering algorithm and model training
GMM-based Spoofing Detection
Improved GMM Testing Model
Spoofing Detection

FIGURE 3. The authentication framework.

2) GMM PARAMETERS INITIALIZATION AND MODEL TRAINING

As described earlier, the EM algorithm is used to find GMM parameters for the feature vectors $\boldsymbol{F}_T(\boldsymbol{H}_A(k), \boldsymbol{H}_T(k+1))$ that come from one of two sources: a legitimate channel corresponding to a transmission from Alice to Bob, and a channel that corresponds to a transmission from a spoofer, Eve to Bob. The training data for the legitimate channel is obtained from the previously known transmissions from Alice to Bob, whereas the training data for the spoofer's channel is generated using a pseudo adversary model. In theory, it is difficult to determine the number of classifications of the model. In order to speed up the training of the GMM, the parameters for the EM algorithm are initialized using the k-means clustering algorithm [27]. The cluster centers are used as the Gaussian means in the GMM and the variance of the samples in each cluster is used as the Gaussian variances in the GMM. For the weights π_i of the GMM, the relative number of targets in each cluster is used. Once the GMM has been found, it is then used to calculate the posterior probability of the channel data that is extracted from a received message.

3) GMM-BASED SPOOFING DETECTION

The GMM is used to determine the "similarity" of $\boldsymbol{H}_T(k+1)$ to the channel information vectors for Alice and for Eve, and then to determine an output target label. This is done by determining the target model that has the maximum a posteriori probability for the given $\boldsymbol{H}_T(k+1)$. The results are used to identify the corresponding sender of the received message. The pseudocode of GMM-based PHY-layer authentication algorithm is shown in Algorithm 1.

VI. RESULTS AND ANALYSIS

A. SIMULATION SCENARIOS

In this section, M ATLAB simulations are used to evaluate the performance of the GMM-based PHY-layer authentication method. For these experiments, an OFDM system is used

Algorithm 1 GMM-Based PHY-Layer Authentication

```
Initialization: Maximum iteration $L$.
for each case of node mobility do
    generate channel information vectors
    obtain two-dimensional channel feature via (9)
    initialize GMM parameters using $k$-means algorithm
    training GMM model via (25), (26), (27) and (28)
    for $k = 1, 2, \cdots, M$ (for each packet) do
        obtain channel feature vector $\boldsymbol{F_T}$
        get the posterior probability with trained GMM
        if the recognition result is Alice then
            update $\boldsymbol{H}_A(k) \leftarrow \boldsymbol{H}_A(k+1)$
            receive the data
        else
            keep $\boldsymbol{H}_A(k) \leftarrow \boldsymbol{H}_A(k-1)$
            send an alarm
        end if
    end for
end for
```

with 1024 subcarriers, QPSK modulation for each subcarrier, and a cyclic prefix of length 256. There were 64 comb-type pilots inserted into each OFDM symbol for channel estimation, and the channel model that was used is a random Rayleigh fading channel. The channel information is sampled for different signal-to-noise ratios (SNRs) that vary from 0 dB to 26 dB. The symbol duration, t, that corresponds to the channel coherence time, is $t = 10$ ms, and the RF wavelength is set to $\lambda = 6$ cm. With node speeds of $v = 100, 80, 65, 60, 50, 45, 40, 35, 20$ km/h, the resulting channel correlation coefficients are $\varsigma = 0.1, 0.2, \ldots, 0.9$, respectively. For training the GMM, 500 samples were used, and the termination criteria set in the EM algorithm is that the error value is less than 10^{-15}.

The performance of our authentication method was compared with that proposed by Xiao *et al.* [19] and the clustering scheme with one-dimensional estimated channel conditions proposed by Weinand *et al.* [21]. To compare the performance of these two approaches with ours, the probability of detecting a spoofer, P_{pd}, when an actual spoofing attack exists was evaluated. The performance of our method was also compared with that of Xiao et. al. in terms of the probability of authentication error, P_e, or minimum Bayes risk, which is the probability of either choosing hypothesis H_0 when H_1 is true or choosing hypothesis H_1 when H_0 is true. This probability is given by

$$P_e = P(\text{Choose } H_0|H_1)P(H_1) + P(\text{Choose } H_1|H_0)P(H_0) \tag{29}$$

In our experiments, we exploited the posterior probability calculated in GMM-based PHY-layer authentication scheme to analyze spoofing detection performance. Therefore, a soft decision is used in the proposed scheme, without depending on a fixed threshold. At each procedure, when Bob receives the new signal, P_{pd} and P_e are calculated to obtain the average results based on 10, 000 runs.

B. NUMERICAL RESULTS

Fig. 4 shows a comparison of the probability of spoofing detection for our GMM-based authentication approach with that of Xiao et. al. [19] and Weinand et. al. [21] for SNRs that vary from 0 dB to 25 dB. In these experiments, the channel correlation coefficient was set to $\varsigma = 0.5$. From the figure, it is clear that the GMM-based authentication approach performs better in being able to detect a spoofer. For example, note that with a SNR of 20 dB, while the detection probability of the method of Weinand et. al. is 0.89 and for the method of Xiao et. al. it is 0.90, the detection probability of our GMM-based method is 0.97. Even when the SNR is less than 10 dB, our GMM-based approach has a detection probability of more than 0.90. The proposed scheme achieves a relatively high system security when SNRs are relatively low. This illustrates the benefits of using a multi-dimensional feature in the GMM-based scheme. Different channel features mean that different explanatory factors behind the data can be hidden more or less. Similar gains were noted for other values of the channel correlation coefficient. What is interesting to note is that the probability of detection does not change much when SNR > 12 dB.

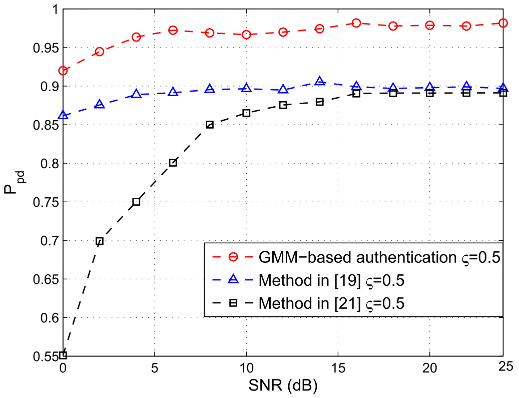


FIGURE 4. Probability of detection versus SNR for a channel correlation coefficient $\varsigma = 0.5$.

A plot of the spoofer detection probability as a function of the channel correlation coefficient, ς, for our GMM-based method and the method of Xiao et. al. is shown in Fig. 5 for two different SNRs. This figure shows that as ς increases, the detection probability increases significantly for both methods. The reason for this is that as the correlation coefficient of the legitimate channel increases, the correlation between $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_A(k+1)$ tends to become larger (See Eq. (11)), and the difference between $\boldsymbol{H}_A(k)$ and $\boldsymbol{H}_E(k+1)$ tends to become larger. As a result, the two channels between Alice-to-Bob and spoofer-to-Bob are easier to classify, which results in a lower P_{pd}. The probability of detection remains approximately constant for $\varsigma > 0.6$. From Fig. 5, it is clear that the GMM-based approach outperforms the method

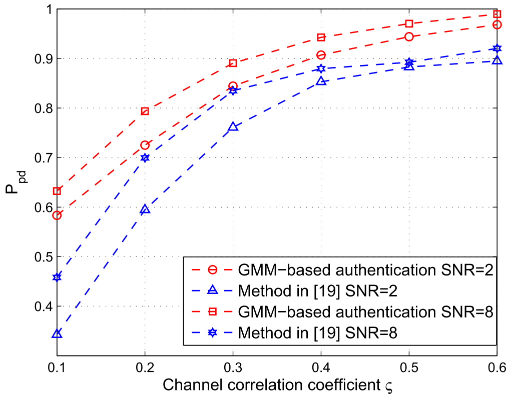


FIGURE 5. Probability of detection versus the channel correlation coefficient ς.

of Xiao. For example, note that when $\varsigma = 0.6$ and SNR $= 2$ dB, the detection probability of the GMM-based method is approximately 0.98 compared to 0.90 for Xiao's method. It should again be noted that similar gains were observed for other SNRs.

Fig. 6 shows how the minimum Bayes risk varies as a function of the channel correlation coefficient for a SNR of 2 dB for our method and the approach of Xiao. Note that as ς increases, the minimum Bayes risk for both methods decreases significantly, with the GMM-based method having a lower Bayes risk, particularly when $\varsigma < 0.7$ As a specific example, note that for $\varsigma = 0.5$, the minimum Bayes risk is 0.05 in our method while the risk is 0.3 for Xiao's method.

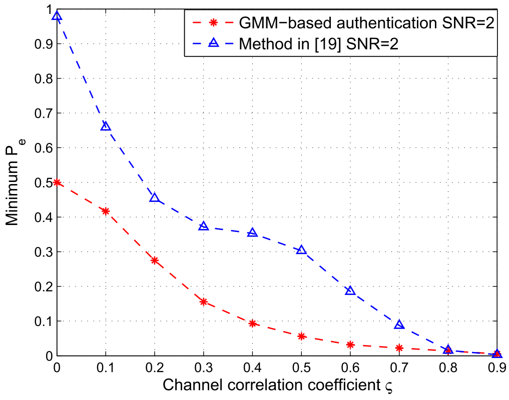


FIGURE 6. The minimum Bayes risk as a function of the channel correlation coefficient ς.

The effect of the channel correlation coefficient ς on the detection performance of the GMM-based PHY-layer authentication scheme is shown in Fig. 7. Note that when the channel correlation coefficient is large, e.g., $\varsigma = 0.9$, the minimum Bayes risk value is very small. From this figure we see that as the correlation coefficient increases the minimum Bayes risk becomes small, which in turn leads to a high probability of spoofing detection as was shown in Fig. 5.

Finally, Fig. 8 illustrates how the minimum Bayes risk varies as a function of SNR for the GMM-based PHY-layer method and Xiao's method [19]. For both methods, the minimum Bayes risk decreases as the SNR increases, which again is as to be expected. Consistent with previous results, this figure shows that the performance of our proposed approach for authentication is better than the method of Xiao. For example, with a SNR of 25 dB, the minimum Bayes risk of the GMM-based scheme is close to 0.07, which is 0.05 lower than that of the method of Xiao. It is also important to note that the minimum Bayes risk is less than 0.1 when the SNR is greater than 8 dB. Compared with the results in Fig. 7, the minimum Bayes risk in Fig. 8 is slightly increased due to smaller value of the channel correlation coefficient.

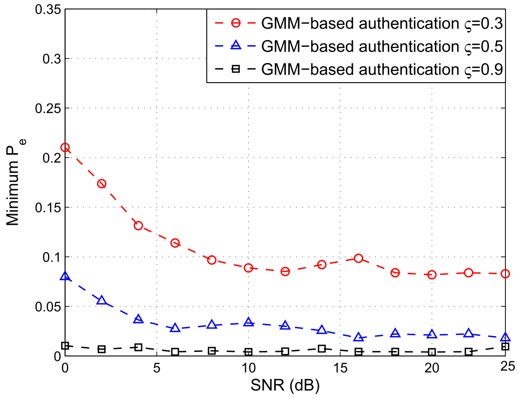


FIGURE 7. The minimum Bayes risk versus SNR for three different channel correlation coefficients.

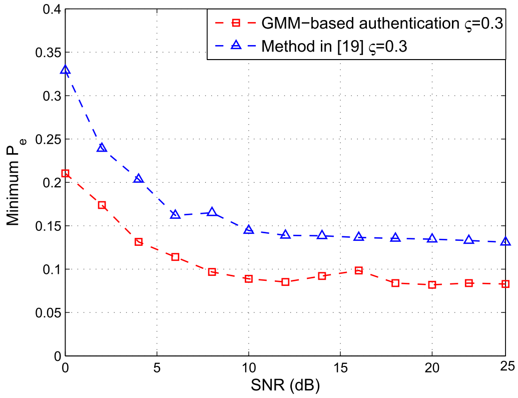


FIGURE 8. The minimum Bayes risk versus SNR at $\varsigma = 0.3$.

C. ROBUSTNESS ANALYSIS

In any practical communication system, it is important for an authentication method to be robust to different channel models and characteristics. Fig. 9 shows the effect of channel estimation errors on our proposed approach to authentication. The channel estimation error is controlled by the value of σ_Δ^2 (See Eq. (2) - Eq. (4)) and for our evaluation, we considered values of R between -2 dB and -20 dB where

$$R = 10\log_{10}\frac{\sigma_\Delta^2}{\sigma_A^2} \tag{30}$$

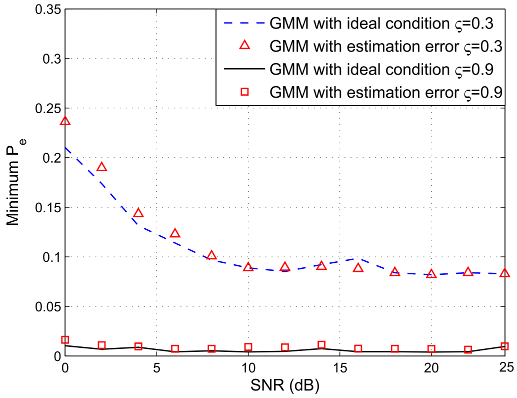


FIGURE 9. **Performance comparison between using the ideal channel and using the channel with estimated error $R = -2$ dB.**

TABLE 2. **Detection probability (%) for different channel estimation errors.**

	$\varsigma = 0.4$	$\varsigma = 0.8$
Ideal channel	86.29	97.28
$R = -20$ dB	85.91	97.16
$R = -10$ dB	84.78	97.11
$R = -2$ dB	83.67	97.23

In Fig. 9, we set $R = -2$ dB, and consider channels with a correlation coefficient $\varsigma = 0.3, 0.9$. This figure shows that the channel estimation error does not have a significant effect on the performance of the proposed GMM method. Table 2 shows the results of using GMM-based method when the correlation coefficient $\varsigma = 0.4, 0.8$ for different channel estimation errors. What these results show is that the mobility of the node, the communication environment, and the channel estimation error all influence the authentication performance of the system. However, note that the effect of channel estimation error becomes negligible when the channel correlation coefficient becomes large.

VII. CONCLUSION

In this paper, a novel channel-based PHY-layer authentication method was presented that is based on training a GMM using a two-dimensional feature vector extracted from an estimated channel state vector. The GMM is used as a classifier to determine whether a new message is being sent from a legitimate transmitter or a spoofer. Although training data is relatively easy to obtain for a legitimate transmitter, this is not the case for a spoofer who is transmitting messages through an unknown channel. Therefore, a pseudo-adversary model was used as a model for the potential spoofer.

This GMM-based authentication method was shown to achieve a very low Bayes risk compared to other methods. Although the maximum spoofing detection probability is high, $P_{pd} = 0.98$, there is room for future work. One is in the area of feature engineering and feature selection with the goal of creating better learning-based models. Although a two-dimensional feature vector that is based on the distance and correlation between two channel state vectors has been used successfully, other features are possible, and the use of more than two features could be considered. Another important study would be to implement this authentication algorithm in a real wireless system in order to evaluate its performance under real conditions and in different scenarios.

REFERENCES

[1] X. Wang, P. Hao, and L. Hanzo, "Physical-layer authentication for wireless security enhancement: Current challenges and future developments," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 152–158, Jun. 2016.

[2] C. Pei, N. Zhang, X. S. Shen, and J. W. Mark, "Channel-based physical layer authentication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 4114–4119.

[3] J. K. Tugnait, "Wireless user authentication via comparison of power spectral densities," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1791–1802, Sep. 2013.

[4] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[5] H. Moosavi and F. M. Bui, "Delay-aware optimization of physical layer security in multi-hop wireless body area networks," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 9, pp. 1928–1939, Sep. 2016.

[6] P. L. Yu, G. Verma, and B. M. Sadler, "Wireless physical layer authentication via fingerprint embedding," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 48–53, Jun. 2015.

[7] P. L. Yu, J. S. Baras, and B. M. Sadler, "Physical-layer authentication," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 38–51, Mar. 2008.

[8] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and privacy analysis of mobile health applications: The alarming state of practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018.

[9] F. J. Liu, X. Wang, and H. Tang, "Robust physical layer authentication using inherent properties of channel impulse response," in *Proc. IEEE Conf. MILCOM*, Baltimore, MD, USA, Nov. 2011, pp. 538–542.

[10] X. Qiu and T. Jiang, "Safeguarding multiuser communication using full-duplex jamming receivers," in *Proc. IEEE PIMRC*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[11] N. Wang, S. Lv, T. Jiang, and G. Zhou, "A novel physical layer spoofing detection based on sparse signal processing," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Orlando, FL, USA, Dec. 2015, pp. 582–585.

[12] N. Wang, W. Li, T. Jiang, and S. Lv, "Physical layer spoofing detection based on sparse signal processing and fuzzy recognition," *IET Signal Process.*, vol. 11, no. 5, pp. 640–646, Jul. 2017.

[13] N. Wang, T. Jiang, W. Li, and S. Lv, "Physical-layer security in Internet of Things based on compressed sensing and frequency selection," *IET Commun.*, vol. 11, no. 9, pp. 1431–1437, Jun. 2017.

[14] J. Liu and X. Wang, "Physical layer authentication enhancement using two-dimensional channel quantization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4171–4182, Jun. 2016.

[15] L. Xiao, L. J. Greenstein, N. B. Mandayam, and W. Trappe, "Channel-based spoofing detection in frequency-selective Rayleigh channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5948–5956, Dec. 2009.

[16] L. Xiao, T. Chen, G. Han, W. Zhuang, and L. Sun, "Game theoretic study on channel-based authentication in MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7474–7484, Aug. 2017.

[17] L. Xiao, Y. Li, G. Liu, Q. Li, and W. Zhuang, "Spoofing detection with reinforcement learning in wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–5.

[18] N. Wang, T. Jiang, S. Lv, and L. Xiao, "Physical-layer authentication based on extreme learning machine," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1557–1560, Jul. 2017.

[19] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10037–10047, Dec. 2016.

[20] G. Huang, S. Song, J. N. D. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.

[21] A. Weinand, M. Karrenbauer, J. Lianghai, and H. D. Schotten, "Physical layer authentication for mission critical machine type communication using Gaussian mixture model based clustering," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Sydney, NSW, Australia, Jun. 2017, pp. 1–5.

[22] W. C. Jakes and D. C. Cox, *Microwave Mobile Communications*. Piscataway, NJ, USA: IEEE Press, 1993.

[23] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[24] K.-S. Hwang, W.-C. Jiang, and Y.-J. Chen, "Model learning and knowledge sharing for a multiagent system with dyna-Q learning," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 978–990, May 2015.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[26] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2016.

[28] L. Xiao, Q. Yan, W. Lou, G. Chen, and Y. T. Hou, "Proximity-based security techniques for mobile users in wireless networks," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2089–2100, Dec. 2013.

[29] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[30] I. R. Baran and B. F. Uchoa-Filho, "Exploiting time coherence in opportunistic beamforming for slow fading channels," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Las Vegas, NV, USA, Apr. 2006, pp. 1753–1758.

[31] R. Zhang, N. F. Timmons, and J. Morrison, "Opportunistic relay scheme exploiting channel coherence time in IEEE 802.15.6 wireless body area networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–7.

XIAOYING QIU (S'18) received the B.S. and M.S. degrees from Shandong Normal University in 2013 and 2015, respectively. She is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. Her current research interests include physical layer security and signal processing.

TING JIANG received the B.S., M.S., and Ph.D. degrees in communication and information system from Yanshan University, China, in 1982, 1988, and 2003, respectively. He is currently a Professor with the Beijing University of Posts and Telecommunications. His research interests include the wireless broadband interconnection, the information theory, the short distance wireless communication technological theory and application, and the wireless sensor network.

SHENG WU (S'13–M'14) received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, in 2014. He was a Post-Doctoral Researcher with the Tsinghua Space Center, Tsinghua University, Beijing. He is currently an Assistant Professor with the Beijing University of Posts and Telecommunications. His research interests are mainly in iterative detection and decoding, channel estimation, massive MIMO, and satellite communications.

MONSON HAYES (M'81–SM'86–F'92–LF'14) received the B.A. degree in physics from the University of California at Berkeley, and the M.S.E.E. and Sc.D. degrees in electrical engineering and computer science from M.I.T. He was a Professor of electrical and computer engineering at the Georgia Institute of Technology until 2011, and served as an Associate Chair with the School of ECE, Georgia Tech, and as Associate Director of Georgia Tech Savannah, and is currently a Professor Emeritus at Georgia Tech. From 2011 to 2014, he was a Distinguished Foreign Professor at Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University, Seoul, South Korea. Since 2014, he has been a Professor and the Chair of the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA. He has served the Signal Processing Society of the IEEE in numerous positions, including General Chairman of ICASSP 1996, ICIP 2006, and ICASSP 2018. He has published over 200 papers, and is the author of the textbook *Statistical Digital Signal Processing and Modeling* (John Wiley and Sons, 1996).