Reinforcement Learning Based Anti-jamming with Wideband Autonomous Cognitive Radios

Stephen Machuzak, *Student Member, IEEE*, and Sudharman K. Jayaweera, *Senior Member, IEEE*
Communications and Information Sciences Laboratory (CISL)
Department of Electrical and Computer Engineering
University of New Mexico
Albuquerque, NM 87131-0001, USA
Email: {smachuzak29, jayaweera}@unm.edu

***Abstract*—This paper presents a design and an implementation of a wideband autonomous cognitive radio (WACR) for anti-jamming. The proposed anti-jamming scheme is aimed at evading a jammer that sweeps across the whole wideband spectrum range in which the WACR is expected to operate. The WACR makes use of its spectrum knowledge acquisition ability to detect and identify the location of the sweeping jammer. This information and reinforcement learning is then used to learn the optimal communications mode to avoid the jammer. In this paper, we discuss a specific reinforcement learning mechanism based on Q-learning to successfully learn such an anti-jamming operation over a several hundred mega-Hz of wide spectrum in real-time. We describe a cognitive anti-jamming communications protocol that selects a spectrum position with enough contiguous idle spectrum uninterfered by both deliberate jammers and inadvertent interferers and transmits till the jammer catches up to it. When the jammer interferes with the cognitive radio's transmission, it switches to a new spectrum band that will lead to the longest possible uninterrupted transmission as learned through Q-learning. We present results of an implementation of the proposed WACR for cognitive anti-jamming and discuss its effectiveness in learning the surrounding RF environment to avoid both deliberate jamming and unintentional interference.**

***Index terms*— Anti-jamming, q-learning, reinforcement learning, sub-band selection problem, wideband autonomous cognitive radios, wideband spectrum knowledge acquisition, wideband spectrum sensing.**

I. Introduction

Wideband autonomous cognitive radios (WACRs) are radios that have the ability to make their own operating decisions in response to the perceived state of the spectrum, network and radio itself [1]. The key to such autonomous operation is the radio's ability to sense and comprehend its operating environment. In general, it is desired that the radio have the ability to operate over a wide frequency range making the problem of sensing all frequencies of interest to the radio in real-time a challenging problem. However, assuming that this is achieved, such WACRs provide an excellent technological option to achieve cognitive communications desired in many application scenarios. A situation in which cognitive communications can be a great asset is when reliable communications is needed in the presence of unintentional interference and deliberate jammers.

In this paper, we present a design of a WACR architecture to achieve cognitive anti-jamming and interference avoidance. We present a general approach that may be used to scan and sense a wide spectrum range in order to achieve real-time spectrum awareness. A cognitive anti-jamming and interference avoidance communications protocol that uses this spectrum knowledge is then developed. There is a strong justification for basing cognitive communications protocols on machine learning so that they can both be autonomous and responsive to dynamic channel and network conditions. In this paper, we employ reinforcement learning to aid our proposed anti-jamming and interference avoidance communications protocol. Reinforcement learning (RL) has the advantage of facilitating unsupervised learning of an optimal decision-making policy under reasonable spectrum dynamics.

There have been a few previous attempts at using machine learning techniques, in particular reinforcement learning, to achieve anti-jamming in cognitive radio networks (CRN). For example, [2] has proposed a modified Q-learning technique for jammer avoidance in a CRN. This ON-policy synchronous Q-learning algorithm was shown to converge faster than the standard Q-learning algorithm in learning the behavior of both a sweeping jammer and an intelligent jammer. Two other reinforcement learning approaches, namely SARSA and QV-Learning algorithms, were investigated in [3] to develop an anti-jamming policy against a smart jammer in a CRN. However, reinforcement learning has found many other applications in cognitive radios than being limited to anti-jamming operation [4]. In fact, there are many examples of use of reinforcement learning in dynamic spectrum sharing (DSS) systems. For instance, in [5] so-called secondary users employed Q-learning to learn optimal transmission powers in channels with unknown parameters. Similarly, in [6] minimax-Q learning was used by secondary users in an anti-jamming stochastic game to learn the spectrum-efficient throughput optimal policy to avoid jammers.

Reinforcement learning is, of course, not the only machine learning tool that can be useful for modeling and implementing anti-jamming cognitive communications. Two promising alternatives are the game theoretic learning and artificial neural networks (ANNs). For example, in [7] anti-jamming and jamming strategies were modeled in a game-theoretic framework allowing radios to learn good policies using a variant of fictitious play learning algorithm. In another study [8], the friend-or-foe detection technique was used to detect intelligent malicious users, acting as jammers, in a CRN. Reinforcement learning techniques can also be used in conjunction with game-theoretic models to help learn good policies. For example, Q-learning based strategies are used in [9] and [10] for anti-jamming and jamming games to find the optimal channel-access strategies. The authors in [9] have shown that Nash-Q and friend-or-foe Q-learning can be effective in aggressive jamming environments and in mobile ad-hoc networks, respectively. In [10], the authors presented a game-theoretic anti-jamming scheme (GTAS) that used a modified Q-learning algorithm to evade jammer attacks.

Most of the above referenced contributions, however, have only been limited to either analysis or simulations. In this research, however, we have developed a comprehensive cognitive anti-jamming communications protocol and implemented on a hardware-in-the-loop (HITL) simulation of a WACR prototype. We show results for a cognitive radio that operates over about 200MHz-wide spectrum in real-time in the presence of common wireless interferers as well as a deliberate jammer. Importantly, we demonstrate that a simple reinforcement learning algorithm can indeed learn the behavior of the jammer to achieve effective cognitive anti-jamming and interference avoidance.

The remainder of this paper is organized as follows: Section II details our proposed WACR architecture and the wideband spectrum knowledge acquisition framework. Section III discusses a cognitive communications protocol for anti-jamming and interference avoidance and its implementation using a reinforcement learning algorithm. Section IV presents our hardware-in-the-loop WACR prototype implementation of anti-jamming and the results observed in the presence of both deliberate jammer and unintentional interference. Finally, the paper is concluded in Section V by drawing a few final conclusions and discussing possible further work.

II. Wideband Spectrum Knowledge Acquisition

The most unique aspect of a cognitive radio is the ability to be aware of its RF environment (spectrum state) [1]. In dynamic spectrum sharing applications, this is achieved by what is called spectrum sensing [1], [11]. In the case of wideband autonomous cognitive radios, on the other hand, spectrum sensing can be more involved than simply finding so-called spectrum white-spaces [1]. Indeed, the potential of WACRs lies in their ability to sense and fully comprehend the wide spectrum of interest to the radio. Such comprehension normally includes not just finding active signals, but also determining the characteristics of these signals so that they can properly be identified. Hence, we define a wideband spectrum knowledge acquisition framework consisting of 3 steps as shown in Fig. 1 [1].

The first step in spectrum knowledge acquisition framework is the wideband spectrum scanning. By definition, WACRs are wideband radios that may operate over a large frequency range. However, due to hardware constraints [1], at any given time, it may be able to observe and process only a portion, called a sub-band, of its operating spectrum range of interest. To gain knowledge of the complete spectrum range, thus, a WACR needs to follow an efficient algorithm to determine which sub-band to be sensed at any given time. Clearly, this choice will depend on the performance objectives of the radio. Wideband spectrum scanning step can, thus, closely be coupled with the communications protocol itself.

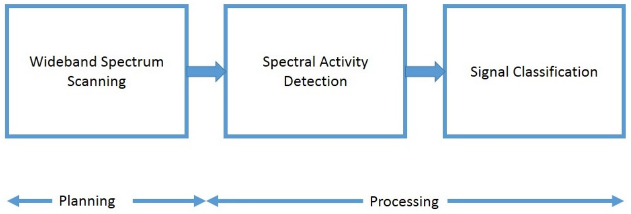

Figure 1. Spectrum knowledge acquisition consists of a planning stage and a processing stage [1].

In the second step of the spectrum knowledge acquisition process, the WACR detects active signals present in the sensed sub-band. For this, our proposed design uses Neyman-Pearson thresholding of an estimated power spectrum of the sub-band signal. Note that, this is very different from spectrum sensing in a DSS cognitive radio in which only a single channel is sensed at a time and a particular type of primary signal is to be detected. Instead, all active signals present in a sub-band is to be detected. This step, thus, allows the WACR to extract carrier frequencies of detected active bands but not necessarily other specific information about the signal [1]. Thus, the wideband spectrum knowledge acquisition framework consists of a third step of signal classification and identification. In this final step, detected signals are classified to identify their origin and, in particular, what systems they may belong to. Often, classification is better performed on certain features extracted from the detected signals [1].

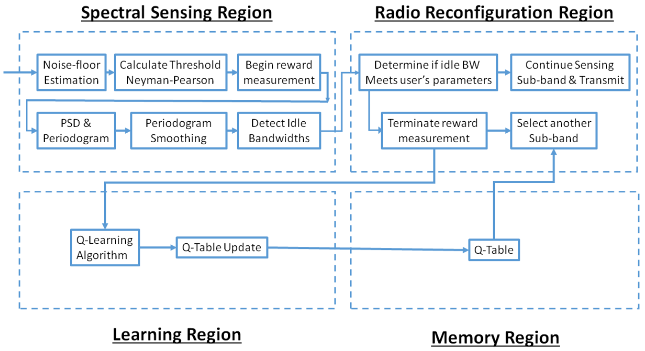

Figure 2. Block Diagram of the Cognitive Engine and its signal processing tasks.

Figure 2 shows a cognitive engine implementation of the above spectrum knowledge acquisition framework especially detailing the steps associated with the spectral activity detection step. First, the noise floor of each of the sub-bands is estimated. This is used to compute the Neyman-Pearson threshold for spectral activity detection subjected to a given false-alarm probability. Next, an estimate of the power spectral density (PSD) of the sensed sub-band signal is computed. In the absence of any a priori knowledge on possible signals in a sub-band, a possible spectrum estimator is the periodogram of the sensed signal, defined as:

$$\hat{S_y}(F) = \frac{1}{N}\left|\sum_{n=0}^{N-1} y[n]^{-j2\pi Fn}\right|^2 \tag{1}$$

where $y[n]$ is the time-domain signal of the sensed sub-band and N is the number of signal samples [1].

The periodogram, however, can be very erratic and noisy even when a large number of samples, N, is used. To reduce the effects of such noisy fluctuations on spectral activity detection, in our approach we apply frequency-domain smoothing to the periodogram estimate of the sub-band spectrum as shown below:

$$T(\mathbf{Y}) = \frac{1}{LN}\sum_{l=-(L-1)/2}^{(L-1)/2} |Y[k+l]|^2 \tag{2}$$

where L denotes the length of the rectangular smoothing window, Y denotes the FFT of the sensed sub-band signal, k is the sample in the spectrum where the rectangular window is centered at and $T(\mathbf{Y})$ is the smoothed periodogram [1]. It is imperative to smooth the periodogram to reduce the possibility of noise causing the PSD estimator to exceed the detection threshold while it should not, and vice versa. The Neyman-Pearson threshold, is then, applied to the smoothed periodogram to detect any active signals in the sub-band.

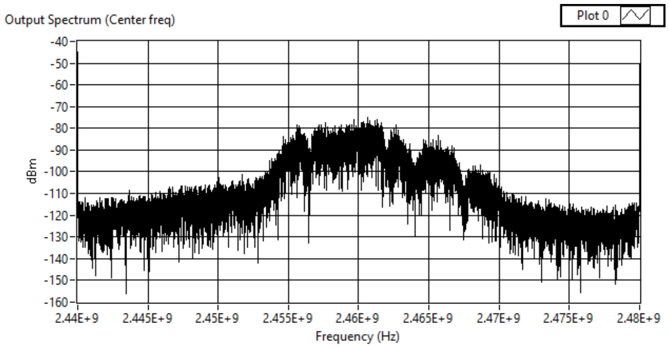

Figure 3. Periodogram estimate of the sub-band spectrum for a 40MHz-wide sub-band centered at 2.46GHz.

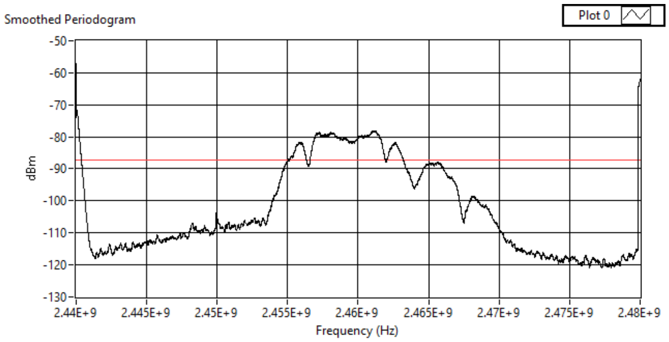

Figure 4. Smoothed periodogram estimate of the sub-band spectrum, as given by (2), for a 40MHz-wide sub-band centered at 2.46GHz.

Figures 3 and 4 show actual real-time periodogram and smoothed PSD estimators for a system that uses 40MHz wide sub-bands. By thresholding the smoothed periodogram estimate (2), the WACR determines the locations and bandwidths of the active signals. This information is then utilized by the radio reconfiguration region (see Fig.2) to determine the idle frequency bands within the just sensed sub-band. These are next used to determine whether there is enough idle bandwidth to satisfy the user's desired minimum idle bandwidth requirement.

III. Cognitive anti-jamming communications

The proposed cognitive anti-jamming communications protocol avoids both deliberate and unintentional interference by learning when to switch its transmission to a new sub-band and when to continue to transmit in the current sub-band. This is called the sub-band selection problem [1]. In this paper, we develop a reinforcement learning based decision policy based on which a WACR selects the sub-bands for sensing and transmission to meet a given user performance criterion. Specifically, our performance objective is anti-jamming and interference avoidance.

For effective sub-band selection, the WACR needs to be able to predict the sub-band that will most likely have desired conditions to meet the performance objectives set by the user [1]. This can effectively be achieved if we were to have a good predictive model for the state dynamics of sub-bands. A commonly used, and a reasonable, model is to assume that the state dynamics are Markov. A cognitive radio learns its environment by sensing one sub-band at a time. Hence, this is a decision-making problem in a partially observable environment leading to a Partially Observable Markov Decision Process (POMDP). Although the POMDP model is elegant in its formulation, optimal policy computation for POMDPs can be computationally too demanding except in the case of small-size problems [1].

In this work, we get around the computational complexity issue by developing a low-complexity reinforcement learning technique to learn an optimal policy for sub-band selection for anti-jamming and interference avoidance. The WACR will select a sub-band that has a portion of the sub-band idle for transmission and has not been interfered with, deliberately or unintentionally, for the longest amount of time. Note that, the type of communications will determine the minimum contiguous length of idle bandwidth a sub-band must have for it to be a candidate for selection. Once the desired idle bandwidth condition is violated in the current sub-band due an interferer or a jammer, the WACR will select another sub-band from among all available sub-bands.

Based on the assumed communications objectives, in this work we have developed a novel, and simple, definition for the state of a sub-band. In particular, each sub-band can be in one of two possible states: Either it contains a contiguous idle bandwidth of a required length (state 1) or it does not (state 0). With this state definition, a WACR will have to select a new sub-band if and when the state of the current sub-band changes to state 0. For efficient operation with effective anti-jamming, of course, the selected new sub-band must have low interference with high probability. When interference is due to a deliberate jammer, efficient selection can be achieved if the WACR can learn the pattern of behavior of the jammer. Our proposal employs an autonomous learning algorithm to achieve this.

An approach to learn an effective sub-band decision policy, as mentioned earlier in this section, is to use reinforcement learning techniques such as Q-Learning. Q-Learning is utilized in this application due to its low computational complexity. Moreover, it does not require the knowledge of transition probabilities of the underlying Markov model. Essentially, Q-Learning is a reinforcement learning technique in which for each state and action pair, what is called a Q-value is computed. The Q-value is a quantification of the merit of taking a particular action when in a given state [4]. After each execution of an action, the WACR updates the Q-table based on a certain observed reward. In our approach, we use a reward function that depends on the amount of time it takes the jammer or interferer to interfere with the WACR transmission once it has switched to a new sub-band.

Let us denote the Q-value associated with selecting action a in state s by $Q(s, a)$. After each execution of an action, the WACR updates the Q-table entries as below, where $0 < \alpha < 1$ and $0 \leq \gamma < 1$ denote the learning rate and the discount factor, respectively [1]:

$$\begin{aligned} Q(s[n-1, a_{n-1}]) \leftarrow {} & (1-\alpha)Q(s[n-1], a_{n-1}) \\ & + \alpha[r_n(s[n-1], a_{n-1}) + \gamma \max_{a}(s[n], a)]. \end{aligned} \tag{3}$$

Our Q-Learning based sub-band selection algorithm selects sub-bands for sensing and transmission based on the Q-table. However, in RL literature, it is well known that a certain amount of exploration of state-action space is required for effective learning. Hence, the sub-band selection policy is defined as:

$$a^* = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(s, a) & \text{with probability } 1-\epsilon \\ \sim \mathcal{U}(\mathcal{A}) & \text{with probability } \epsilon \end{cases} \tag{4}$$

where $\mathcal{U}(\mathcal{A})$ denotes the uniform distribution over the action set and ϵ is the exploration rate (or the exploration probability). Note that, an exploration rate of ϵ implies that the learner randomly selects an action with probability ϵ (explores an action) and it selects the best action, as implied by the learnt Q-table, with probability 1-ϵ (exploitation). The exploration rate needs to be carefully selected so as to strike an acceptable balance between exploration and exploitation [1]. A high exploration rate may help the WACR to quickly understand the environment but it could reduce the performance due to excessive exploring and not exploiting what it has learned. In contrast, a low exploration rate could make the WACR take far more time to learn the environment and converge to the optimal solution, when that is indeed possible [1].

IV. Simulation Results

The hardware-in-the-loop setup is implemented on a LabVIEW program using an NI-USRP software-defined radio. Signal processing tasks of the cognitive engine are performed by the LabVIEW program running on a laptop in real-time.

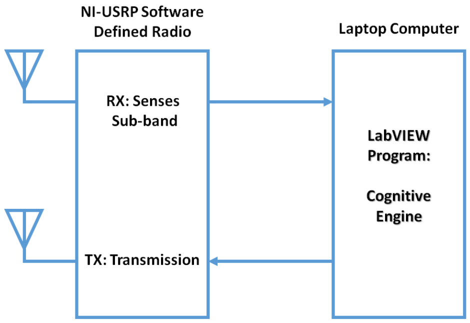

Figure 5. The setup of the hardware and a general top-layer overview of the hardware-in-the-loop setup.

Figure 5 shows the general hardware-in-the-loop simulation setup. The hardware portion collects real-time data, and passes them to the cognitive engine for processing. In addition, it also transmits the radio's own signals as instructed by the cognitive engine.

Our WACR prototype operates over a spectrum range of 200MHz in real-time and scans 40MHz-wide sub-bands at a time. In this case, the Q-table is a 5x5 matrix. Specifically, there are 5 states and 5 actions: the rows are the states and the columns are the actions. Note that, the action is the sub-band it selects for sensing during the next time instant in an attempt to escape the jammer.

To demonstrate our prototype's ability to learn a good sub-band selection policy, our field test used a continuous sweeping signal acting as the jammer which sweeps the 200MHz-wide spectrum within a period of 35 seconds. We tested our learning algorithm in two spectrum ranges: the 2GHz-2.2GHz band that usually contained unintentional outside interferers in addition to our sweeping jammer signal and the 3GHz-3.2GHz band that was mostly free of additional unintentional interferers.

The jammer sweeps these frequency bands from the lower to the higher frequency. Hence, in the absence of any other interference the optimal sub-band selection policy to avoid the jammer is intuitive: The WACR should cyclically shift to the sub-band that is adjacent to the current sub-band from the lower frequency side. For example, if the WACR is currently sensing sub-band 5, it should choose sub-band 4 in order to avoid the jammer for the longest amount of time possible. Table I shows this intuitive pattern of the optimal sub-band selection policy that the WACR needs to learn in order to effectively avoid the sweeping jammer (under the assumption that there are no other interferers except the sweeping jammer). Results from our field tests show that our WACR can indeed learn the above optimal sub-band selection policy to avoid deliberate jamming. Tables II and III show the Q-tables learned by the WACR, while operating in the 3GHz-3.2GHz band and the 2GHz-2.2GHz band, respectively. In these experiments, user defined minimum required bandwidth in a sub-band is 30MHz. Note that, the difference between Tables II and III is that in Table II, the WACR operated in a frequency band that was free of unintentional interference whereas in Table III the WACR operated in a band with unintentional interference.

Table I
Q-TABLE WITH OPTIMAL POLICY ANTI-JAMMER AVOIDANCE PATTERN.

a s	1	2	3	4	5
1	0	0	0	0	max Q-value
2	max Q-value	0	0	0	0
3	0	max Q-value	0	0	0
4	0	0	max Q-value	0	0
5	0	0	0	max Q-value	0

Table II
LEARNED Q-TABLE IN THE 3GHZ TO 3.2GHZ BAND

a s	1	2	3	4	5
1	0.0461	0.0956	0.2907	0.4676	4.6945
2	4.8770	0.0830	0.2008	0.2872	0.9495
3	0.8342	4.6882	0.1628	0.2097	0.2882
4	0.3272	0.7844	4.5411	0.0645	0.2087
5	0.2048	0.7756	0.7705	4.5520	0.0851

Table III
LEARNED Q-TABLE IN THE 2GHZ TO 2.2GHZ BAND

a s	1	2	3	4	5
1	0.0971	0.3677	0.4801	0.4254	1.0584
2	1.5785	0.2964	0.1780	0.3003	0.6007
3	0.4680	1.4561	0.0940	0.1792	0.30792
4	0.3332	0.2704	1.4148	0.1881	0.1898
5	0.3323	0.5728	0.4249	1.2130	0.1328

Clearly, these Q-tables show that our proposed reinforcement learning based sub-band selection algorithm can indeed learn the sweeping jammer's behavior and perform as an effective cognitive anti-jamming and interference avoidance protocol. The Q-tables in Tables II and III show that if the system were to exploit (choose the actions resulting in the greatest reward), it will indeed choose the optimal sub-band that follows our intuition as previously mentioned and as shown in Table I. Another observation from these results is that our proposed learning scheme is relatively robust against unintentional interference. For example, Table III shows that despite the presence of both unintentional interference and the deliberate jammer, the WACR is successful at learning a good action selection policy to avoid the jammer.

V. CONCLUSION

In this paper, we have presented an anti-jamming wideband autonomous cognitive radio and demonstrated it is indeed capable of evading both deliberate jammers and unintentional interference. In addition, we have also demonstrated that reinforcement learning can be an effective approach for a WACR to learn the optimal communications mode to avoid a deliberate jammer. Results obtained from an HITL simulation showed that it was able to successfully infer the jamming pattern and learn the optimal sub-band selection policy for jammer avoidance.

A possible future work is to use an expanded Q-table that can also include states in which unintentional interferers are present alongside the jammer. Implementing a game-theory based approach to defeating jammer interference can also be an additional future goal.

ACKNOWLEDGMENT

This work was funded in part by the Air Force Research Laboratory (AFRL), Space Vehicles Directorate, under grant FA9453-15-1-0314 and in part by the National Aeronautics and Space Administration (NASA) under the contract NNX15CC80P.

REFERENCES

[1] S. K. Jayaweera, *Signal Processing for Cognitive Radios*. Hoboken, New Jersey: Wiley, 2015.

[2] F. Slimeni, B. Scheers, Z. Chtourou, and V. L. Nir, "Jamming mitigation in cognitive radio networks using a modified q-learning algorithm," in *Military Communications and Information Systems (ICMCIS), 2015 International Conference on*, May 2015, pp. 1–7.

[3] S. Singh and A. Trivedi, "Anti-jamming in cognitive radio networks using reinforcement learning algorithms," in *Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*, Sep. 2012, pp. 1–5.

[4] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1136–1159, Third 2013.

[5] T. Chen, J. Liu, L. Xiao, and L. Huang, "Anti-jamming transmissions with learning in heterogenous cognitive radio networks," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2015 IEEE*, Mar. 2015, pp. 293–298.

[6] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 877–889, Apr. 2011.

[7] K. Dabcevic, A. Betancourt, L. Marcenaro, and C. S. Regazzoni, "A fictitious play-based game-theoretical approach to alleviating jamming attacks for cognitive radios," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 8158–8162.

[8] S. R. Sabuj, M. Hamamura, and S. Kuwamura, "Detection of intelligent malicious user in cognitive radio network by using friend or foe (FoF) detection technique," in *Telecommunication Networks and Applications Conference (ITNAC), 2015 International*, Nov. 2015, pp. 155–160.

[9] Y. Gwon, S. Dastangoo, C. Fossa, and H. T. Kung, "Competing mobile network game: Embracing anti-jamming and jamming strategies with reinforcement learning," in *Communications and Network Security (CNS), 2013 IEEE Conference on*, Oct. 2013, pp. 28–36.

[10] C. Chen, M. Song, C. Xin, and J. Backens, "A game-theoretical anti-jamming scheme for cognitive radio networks," *IEEE Network*, vol. 27, no. 3, pp. 22–27, May 2013.

[11] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb. 2005.