Reduction of Computational Time for Cooperative Sensing Using Reinforcement Learning Algorithm

S.A. Olatunji[1], T.O. Fajemilehin[2] and J.F. Opadiji[1]
[1]Department of Computer Engineering
[2]Department of Electrical and Electronics Engineering,
University of Ilorin, Ilorin, Nigeria
olatunji.sa@unilorin.edu.ng, fajemilehin.to@unilorin.edu.ng,jopadiji@unilorin.edu.ng

Abstract

Cooperative spectrum sensing in cognitive radio systems is characterized by high computational time for decision making due to the fusing of individual decisions of cognitive radios involved in the cooperative scheme. This increases the communication overhead of the network. In this paper, an adaptive cooperative spectrum sensing algorithm is developed with improved detection algorithm. Reinforcement learning is thenincorporated to improve the decision making efficiency of the cooperative spectrum sensing such that less time is required to make a decision at the fusion centre. Three temporal difference learning techniques were compared in order to select the most efficient to reduce sensing and decision delays. Appropriate learning rate was utilized in the sensing and decision making algorithm to enhance the performance ofthe adaptive cooperative spectrum sensing. Results reveal significant reduction in the computation time required in cooperative spectrum sensing and decisions. This permits greater efficiency in dynamic spectrum management as the limited electromagnetic spectrum is being utilized for telecommunication services.

Keywords: Cognitive Radio, Cooperative Spectrum Sensing, reinforcement learning, computational time.

1. INTRODUCTION

Traditional wireless networks use fixed spectrum allocation policies for licensed users. Recent studies on the measurement of the spectrum show that by the conventional spectrum allocation policy, the average utilization of the spectrum is low [1]. Therefore, the real problem is not the spectrum scarcity but the inefficient spectrum usage. This inefficiency results from static spectrum allocations which cannot support the recent advancement and growth in wireless devices and services. This inevitable increase has therefore limited the suitability of FCA in managing the available useable radio spectrum effectively and efficiently. In addition to this, the world is advancing towards an era where usersrequire dynamic spectrum usage for optimized connectivity and capacity [2].

For this purpose, cognitive radio is proposed to efficiently harness spectrum opportunities leading to effective spectrum usage while improving the quality and efficiency of newer applications and services. Cognitive radio is a software defined radio that can autonomously modify the spectrum usage after sensing its operating environment to satisfy user and network demands [3]. It therefore provides the opportunity for frequency re-use and a more efficient technology to allocate channels and bandwidth for reliable communication [4]. The attention in this paper is focused on spectrum sensing function of cognitive radio which is the process of

detecting unused frequency bands and allocating them intelligently and autonomously without interference to needing users.

Energy Detection among all other spectrum sensing techniques despite its shortcomings under low SNR has been observed through research to possess the lowest computational and implementation costs which usually makes it the preferred technique. The cooperative sensing technique was developed to improve on the single user detection techniques which includesthose utilizing energy detection algorithm.Cooperative sensing helps to overcome the severe challenges faced by cognitive radios such as multipath fading and shadowing. However, there is increased delay incurred through the processing of the information received from the cognitive radios that make up the cooperative spectrum sensing scheme. This rise in time consumption is a major drawback for the cooperative sensing technique.

Several research efforts had been channelled to tackle this challenge as seen in [2], [5]–[7], but these were implemented using the conventional cognitive radios which perform poorly in low SNR conditions. Research efforts made to improve the sensing accuracy imposed further computational demands on the system and ultimately increases the overhead in terms of processing capacity, speed and memory demands [7]. The proposed method in this study uses an improved sensing algorithm. It possesses an adaptive mechanism that places very minimal computational demands on the system while increasing the sensing accuracy of the cooperative cognitive radios and minimizing the communication delay using reinforcement learning (RL). This defines its unique contribution to cognitive radio research.

The aim of this research, therefore, is to reduce the time consumed during cooperative spectrum sensing through the development of an adaptive cooperative spectrum scheme and the introduction of reinforcement learning algorithm. This aim is intended to be achieved while maintaining accurate spectrum sensing. It begins with the development of an adaptive cooperative spectrum scheme which utilizes improved energy detection sensing algorithm. Then various temporal difference learning techniques were compared to ascertain the most applicable to reduce the delay. Consideration of different learning rates was also done to determine the most suitable. The most applicable RL technique along with the most appropriate learning rate were then applied to the adaptive cooperative spectrum sensing algorithm initially developed.

The scope of the work is limited to cooperative spectrum sensing algorithm of cognitive radio systems. It is particularly focused on the reduction of time consumed during the processing of the multiple data received from the cognitive radios in the cooperative spectrum sensing process.

The rest of the paper continues with a summary of the recent researches conducted to implement reinforcement learning in cognitive radio systems. Details on the development of the adaptive sensing algorithm with improved sensing will be described in the methodology along with the reinforcement learning implementation. Simulation conducted and results are then presented, followed by discussions, conclusion and suggestions for future work.

2. RELATED WORKS

In the quest to improve the detection quality of energy detectors in cognitive radio systems, [8]proposed and evaluated an improved version of the energy detection algorithm.It outperformed the classical energy detection scheme while preserving a similar level of algorithm

complexity and computational cost. This reduced the time involved in detection compared to other more sophisticated methods of sensing as well as the sensitivity of the cognitive radios.

Another form of improvement was made for energy detection spectrum sensing by [9]. An optimum power operation was chosen to replace the squaring operation in the classical energy detector. This provided useful guidance on techniques to improve the performance of energy detectors in cognitive radio spectrum sensing. [10] further studied and tested various spectrum sensing schemes which included the adaptive threshold energy detector. The results were evaluated using several performance metrics and hardware requirements to confirm the theoretical basis of these techniques. This provided a solid foundation for the improvements in the area of sensitivity improvements and adaptation of the sensing algorithms.

A similar but improved energy detector was developed for wideband spectrum sensing in cognitive radio networks [11] with the aim of determining the detection thresholds for non-overlapping sub-bands. This improved spectrum sensing and opportunistic access for secondary users. It was developed specifically for wideband spectrum sensing and the focus was on accurate detection.[12] proposed an augmented spectrum sensing algorithm where the energy detector's detection is augmented by cyclostationary detection. This, however requires some information about the primary users' transmission characteristics which is not always available.

[13] had earlier discovered a means to improve the performance of the schemes employed for sensing using multiple antenna techniques. Several combining techniques for the cognitive radio users in cooperative spectrum sensing were considered while utilizing different modulation schemes to arrive at the Equal Gain Combining (EGC) which gave the highest gain. This paved the way formore efficient cooperative sensing regarding report combining techniques that would aid performance. [14] also proposed a scheme to improve the utilization of idle spectrum while ensuring fairness in channel selection. Even though, this work did not focus on reduction of time spent in sensing, it is worth noting as one of the researches in cognitive radio sensing that solved the problem of collision among cognitive radios.

The improved energy detection technique proposed by [8] was further used by [15]but enhanced with a *p*-norm energy detector to improve the sensing algorithm of individual cognitive radios used in cooperative spectrum sensing. This resulted in improved performance gain in fading channels when carrying out cooperative spectrum sensing. The focus was on sensing accuracy and learning was not included in the algorithm.

Distributed algorithms for learning was applied in cognitive radio systems by [16] while [17] applied a combinatorial multi-armed bandit formulation as a learning multiuser channel allocation technique. Both applied some form of learning in channel allocation and management, which aided better spectrum access and management but did not specifically address spectrum sensing time. The research to incorporate reinforcement learning algorithm in cognitive radio systems was initially carried out by [7]. They examined the use of Q-learning as a fast means of allocating channels in a wireless network. [5] further expanded the work by comparing several temporal-difference learning methods to minimize the cooperation overhead and improve detection performance.This work utilized the conventional spectrum sensing approach where learning is incorporated. More recently, [2] proposed a two-stage reinforcement learning approach to improve the performance of the cooperative sensing. The method helped minimize the number of sensing operations and reduced the energy required for sensing. These papers [2], [5], [7] utilized RL to improve the channel sensing and allocation, which demands computational

resources and learning time. This paper focuses on reducing the learning and sensing time by resolving the challenge of accurate sensing particularly during instantaneous signal property changes before incorporating learning.Thisimproves the quality of the data utilized for learning while ensuring minimal sensing time. This is done by incorporating an adaptive system of sensing even in low signal to noise ratio conditions which the RL algorithm can learn to utilize in varying channel conditions.

3. SYSTEM MODEL

The radio spectrum can be modelled in the form of communication channels with occupancy status. A channel will be considered available for transmission if no communication agent is transmitting on that channel at that point in time. The channel will be considered unavailable if there is ongoing transmission on it. The communication agent could be a licensed user (Primary User - PU) or could be an unlicensed user (Secondary User - SU) seeking opportunistic use of the channel when available.Spectrum sensing aims to detect these occupancy states and in order for the right decision to be taken. The hypothesis testing problem then arises as:

$$H_0 : \hat{y}[z] = \check{n}[z] \qquad z = 0, 1, \ldots, Z-1 \tag{1}$$

$$H_1 : \hat{y}[z] = \hat{s}[z] + \check{n}[z] \qquad z = 0, 1, \ldots, Z-1 \tag{2}$$

where
$\hat{y}[z]$= sample to be analyzed at each instant z,
$\check{n}[z]$ = noise (not necessarily white Gaussian noise) of variance σ^2,
$\hat{s}[z]$ = is the signal the network wants to detect
H_0 = noise-only hypothesis
H_1 = signal plus noise hypothesis
z = the number of samples collected during the signal observation interval

An energy detector is proposed as the cognitive radio sensing system due to its low computational complexity and operational demands which lowers the overall communication overhead. It simply measures the energy on a particular channel on a narrowband portion of the spectrum and compares it to a pre-set threshold to determine the presence of the primary user in the channel [4]. This is premised on the assumption that the energy of a signal is usually higher than background noise. It usually does not require prior knowledge of the primary user's transmission characteristics to detect which makes it less computationally demanding compared to other methods. If the measured energy exceeds the threshold set, the channel is declared busy which means that the primary user is occupying the channel at that instance. Energy measured which falls below the threshold indicates a spectrum opportunity for a secondary user to harness. The normalized test (decision) statistic for this detector is formulated similar to [18]as:

$$T' = \left(\frac{1}{N_{02}}\right) \int_0^T y^2(t)dt \tag{3}$$

Where:
T'= test statistic in during sensing session
y= received signal input
T= sampling instant

N_{02} = two-sided noise power density spectrum

H_1hypothesis results if the test statistics exceeds a fixed decision threshold while H_0hypothesis occurs when the test statistics is less than the decision threshold. This is the conventional sensing algorithm of energy detectors which would be referred to in this paper as the Conventional Energy Detection (CED). It is the modelused in conventional cooperative sensing schemes. The improved cooperative sensing algorithm, which has an adaptive mechanism is developed to enhance the detection sensitivity of the multiple sensors.

Modelling the Adaptive Sensing Algorithm using Improved Energy Detection.

The Improved Energy Detection (IED) proposed in [8], is developed as an improvement over the CED. The rationale for this improvement is to forestall the misdetections which could occur due to instantaneous deviation in the power received fromaPrimary or secondary user transmitting on a channel.The CED algorithmmayraise false alarms and may also cause interferences if it gives a wrong sensing report due to such instantaneous deviations which are not uncommon with energy detectors. The improved scheme is proposed to manage such errors. This is doneby computing two additional checks if the test statistics indicates a value less than the threshold value. The first check, presented in (4) and (5)begins with computing the average signal energy over a certain period of time indicated as *p* during the i^{th} sensing interval:

$$T_i^{avg}(T_i) = \frac{1}{P}\sum_{l=1}^{P} T_{i-P+l}(x_{i-P+l}) \quad (4)$$

$$T_i = (T_{i-P+1}(x_{i-P+1}), (T_{i-P+2}(x_{i-P+2}), \dots, (T_{i-1}(x_{i-1},), T_i(x_i)) \quad (5)$$

where

$T_i^{avg}(T_i)$ = average test statistic in the *i*-th sensing event
T_i = test statistic vector
P = number of previous sensing events.

The firstcheck is to examine if $T_i^{avg}(T_i) > \lambda$ to determine if the PU is using the channel or not duringa series of pastsensing sessions. Thesecond check to ensure accurate detection is tocheck the immediate past sensing session $T_{i-1}(x_{i-1})$ to determinethe duration for which the channel has been vacant. This check, $T_{i-1}(x_{i-1}) > \lambda$, is to confirm the viability of the channel for allocation and is referred to as an Improved Energy Detection (IED) technique.λ is the decision threshold which in the number of samples $N \gg 1$, can be expressed as a Gaussian distribution as proposed in [8]:

$$\lambda = \sqrt{\frac{2}{NQ^{-1}}}(P_{fa}^{CED} + 1) \quad (6)$$

where:

$$P_{fa}^{CED} = Q\left(\frac{\lambda - 1}{\sqrt{\frac{2}{N}}}\right) \quad (7)$$

$$P_d^{CED} = Q\left(\frac{\lambda - (1+\gamma)}{\sqrt{\left(\frac{2}{N}\right)(1+\gamma)^2}}\right) \quad (8)$$

$$\gamma = \frac{\sigma_s^2}{\sigma_w^2} \tag{9}$$

σ_s^2 is the received average primary signal power and σ_w^2 is the noise variance.

$T_i^{avg}(T_i)$is normally distributed as an average of independent and identically distributed Gaussian random variables. It can be expressed as [15]:

$$T_i^{avg}(T_i) \sim N\left(\mu_{avg}, \sigma_{avg}^2\right) \tag{10}$$

$$\mu_{avg} = \frac{R}{p}(1+\gamma) + \frac{p-R}{p} \tag{11}$$

$$\sigma_{avg}^2 = \frac{R}{p^2}\left(\frac{2}{N}(1+\gamma)^2\right) + \frac{p-R}{p^2} \tag{12}$$

The probability of detection, probability of false alarm and threshold for the IED can therefore be modified as[8]:

$$P_d^{IED} = P_d^{CED} + P_d^{CED}\left(1 - P_d^{CED}\right)Q\left(\frac{\lambda_{IED} - \mu_{avg}}{\sigma_{avg}}\right) \tag{13}$$

$$P_{fa}^{IED} = P_{fa}^{CED} + P_{fa}^{CED}\left(1 - P_{fa}^{CED}\right)Q\left(\frac{\lambda_{IED} - \mu_{avg}}{\sigma_{avg}}\right) \tag{14}$$

$$\lambda_{IED} = \left(Q^{-1}\left(P_{fa,target}^{IED}\right)\sqrt{2N} + N\right)\sigma_w^2 \tag{15}$$

Adaptive Cooperative Spectrum Sensing using IED algorithm

The sensitivity of the individual cognitive radios is first enhanced using the IED algorithm described previously. This makes the cooperating sensing different from the conventional methods. It ensures accurate detection at the individual CR units before being organized into a cooperative network.Each CR user orthogonally sends its report to the fusion center which adaptively uses OR fusion rule (hard fusion rule) when the SNR status of the channel is reliable and equal gain combining (soft fusion rule) when the SNR status is unreliable. The rationale for this is to ensure that onlyvery minimal cognitive radios are involved in the sensing when the SNR is reliable. These cognitive radios use the IED algorithm which is sensitive enough to detect accurately while the OR fusion technique further helps to prevent interference. Research in energy detection based sensing reveals that energy detectors perform reliably when the SNR status of the channel is good. When OR fusion rule is used, the IED threshold is utilized as follows when there are *m* cooperative cognitive radios:

$$\lambda_{IED,OR} = \left(\sqrt{\frac{2}{N}}Q^{-1}\left(1 - \left(1 - P_{fa}^{IED}\right)^{\frac{1}{m}}\right) + 1\right) \tag{16}$$

While the probability of detection becomes:

$$P_{d,OR}^{IED} = \left(1 - \left(1 - Q\left(\left(\frac{\lambda_{IED,OR}}{\sigma_{avg}^2(1+\gamma)}\right)\sqrt{\frac{N}{2}}\right)\right)\right)^m \tag{17}$$

The aim of the adaptive scheme is to ensure that reliable detection is carried out at every session with minimal cognitive radios which would minimize the time required to collate sensing reports and take spectrum allocation decision. Reinforcement learning is then applied to the adaptive cooperative sensing algorithm to further reduce time consumed during cooperative sensing.

Reinforcement Learning Model for Cooperative Sensing

Reinforcement Learning (RL) is a type of machine learning technique that is usually utilized in systems that need to make decisions in unpredictable environments such as the spectrum sensing environment. RL is particularly included in this research work to minimize the processing time which is the time it takes for each of the cognitive radios to report to the fusion center. This cost (processing time) is particularly high due to the sensitivity enhancement and adaptation to SNR that had been initially done. The additional verification carried out by each of the cognitive radios consumes a lot time thereby increasing the processing time before decision is made. Since some spectrum opportunities are just short-lived, a fast detecting system is therefore required which would process the multiple reports speedily and exploit spectrum holes maximally. This is what inspired the inclusion of reinforcement learning in this work to cut down on the delay involved.

In this research, the fusion center acts as the learning agent which takes the speed of reports from each of the cognitive radios as input and learns over time the category of cognitive users which carry out speedy and accurate sensing in specific parts of the spectrum. It then categorizes each of the cognitive radios and labels each one in line with the aspect of the spectrum where it senses best since each of the cognitive radios are located differently. This therefore implies a virtual categorization different spectrum sections to reveal the cognitive radios performing optimally in specific portions of the radio environment [13] and also to learn optimal spectrum sensing decisions.

Reinforcement Learning Model in the Adaptive Cooperative Sensing

The model used is premised on Markov Decision Process (MDP) where the learning agent takes actions at each given state and improves its next action at the next state based on the reward obtained from the present state. A simplified form of the RL model in spectrum sensing is presented in **Figure 1**.

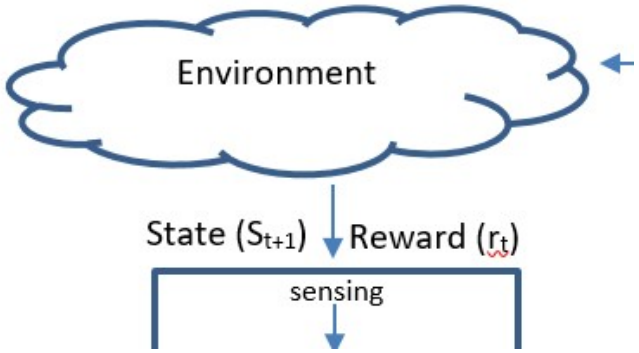

Figure 1: Reinforcement Learning Model in Spectrum sensing

Reinforcement learning was used for the cooperative spectrum sensing in [2], [5], [7], [14]. However, the conventional cooperative spectrum sensing technique was used. This paper differs from the cited papers due to the incorporation of reinforcement learning technique in an adaptive cooperative sensing scheme. In this scheme, each of the cognitive radios are embedded with the IED algorithm. This therefore improves the overall efficiency of the system.

Components of the RL model
Temporal difference prediction techniques are explored in this paper for the RL which consists of pairs of states and actions. The aim is to estimate the function that can represent the behavior policy of the cognitive radios during sensing. A representative series of state-action pairs and the rewards are presented in **Figure 2**.

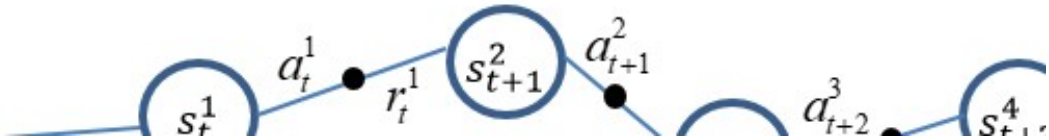

Figure 2: Representative model of state-action pairs with accompanying reward

The components of the RL model as are follows [19]:

State: this refers to the condition of a channel at every point in time. It would be denoted with

$$s_t^i \in S = \{1,2,\dots,m\} \tag{18}$$

where m represents the total number of channels available for transmission.
The value of the state of SU^i may be represented as:

$$s_{m,t}^i \in \{0 \le p_{idle,m}^i \le 1\} \tag{19}$$

This denotes the probability that the channel is idle PU is absent. A channel state could deteriorate by a value $\delta = 0.01$ if it has not been sensed recently. At each time stamp, each channel status should be updated as follows:

$$s_{m,t+1}^i = s_{m,t}^i - \delta \tag{20}$$

Action: this represents every step a_t^i taken out of a possible set of actions Ain each state.

$$a_t^i \in A = \{1,2,\dots,m\} \tag{21}$$

The actions that can be taken are: transmit, idle, sleep and sense

Delayed Reward: cost values received at time $t+1$for each action taken at time t

$$r_{t+1}^i(a_{t+1}^i) \in R = \{1,-1\} \tag{22}$$

Every successful transmission gets 1 while unsuccessful transmission gets a cost of -1.

Discounted Reward: This is a function of the discount factor γwhich reflects the reliance of the cognitive radio agent on the discounted future reward compared to the delayed reward. It is the representation of Q-values to determine future rewards. A Q-function can be presented as follows:

$$Q_{t+1}^i(a_t^i) \leftarrow (1-\alpha)Q_t^i(a_t^i) + \alpha.r_{t+1}^i(a_t^i) \tag{23}$$

The Q-function is used to update the Q-values in the network. It is a function of a state-action pair value computation $f(s_t^i, a_t^i)$. Higher values of the function $f(s_t^i, a_t^i)$ indicates desirable actions a_t^i of the

cognitive radio at specific state s_t^i. Lower values of the function indicate the converse. Learning the right values occurs as a product of the linear function $f(s_t^i, a_t^i)$ and matching values $\theta(s_t^i, a_t^i)$:

$$Q_t^i(s_t^i, a_t^i) = \theta(s_t^i, a_t^i) \,.\, f(s_t^i, a_t^i) \tag{24}$$

The values of the state-action matching pair $\theta(s_t^i, a_t^i)$ is updated as:

$$\theta_{t+1}(s_t^i, a_t^i) = \theta_t(s_t^i, a_t^i) + \alpha\left[r_{t+1}^i(s_t^i) + \gamma \,.\, Q_t^i(s_{t+1}^i, a_{t+1}^i) - Q_t^i(s_t^i, a_t^i)\right] .\, f(s_t^i, a_t^i) \tag{25}$$

Algorithm Description of Proposed Approach

This is the proposed procedure ofincorporating reinforcement learning process in the adaptive cooperative sensing.Action is taken at the fusion center. This action is to request for local sensing decisions from the cooperating secondary users. The approach for action selection is based on Boltzmann distribution for action selection which aids the utilization of all sensing information present. This approach has also been explored alongside other action selection strategies and has been proved efficient [20]. This action selection approach is a softmax approach which chooses an action a_n in state $s_n$based on weighted probabilities. It is presented as:

$$p(s_n, a_n = i) = \frac{e^{Q(s_n, a_n = i)/\tau_n}}{\sum_{j=1}^{N_a} e^{Q(s_n, a_j)/\tau_n}} \tag{26}$$

where

$i = 1, \ldots\ldots, N_a$

(s_n, $a_n = i$) = state-action value function that evaluates the quality of choosing action $a_n = i$ in state s_n,

N_a= number of actions, $|A|$

τ_n= temperaturewhich is a time varying parameter that affects the trade-off between exploration and exploitation.

The assumption made in Boltzmann action selection strategy is that the softmax over the output of the network is a measure of the learning agent's reliance on each action. Actions are selected based on the relativevalue of the sensing information. Actions that are therefore suboptimal are not considered. As more sensing sessions occur, the temperature parameter is reduced, intentionally to enable more exploitation of the spectrum holes already explored. The linear function to achieve this as used by [5] in is given as:

$$\tau_n = -(\tau_0 - \tau_N)(n/N + \tau_0) \tag{27}$$

where

N= total number of episodes,

τ_0 = initial value of temperature

τ_N= last value of the temperature.

The decisions sent back by each of the cooperating CR users are utilized to obtain the optimal decisions and to calculate the response of the action take as a function of time. The fusion point now selects an optimal set of cooperating users based on the speed of their response while sensing in specific areas and

attaches location tags such that the optimal group can be used in subsequent events that requires sensing in such areas. The cumulative reward is a function of the cumulative delay which is the fraction of time it takes each state to give its full report. This is presented given as [5]:

$$C_d = \frac{\sum_{j=0}^{n-1} t_d(s_j,a_j) + t_d(s_n,a_n=i)}{\sum_{j=0}^{m-1} t_d(s_j,a_j)} \tag{28}$$

where

C_d = delay cost factor

t_d = delay in report.

The cumulative delay can be modelled into a cost function which would be minimized as follows:

$$C_{s,a} = \min_{a \in A} \left(C_{d,i}(s_t, a_t)\right) \tag{29}$$

This constitutes the reward obtained for each action taken in each state. The aim is to maximize this reward by taking actions that give sensing report in a reduced time span. The goal is learning a policy that would achieve this aim.

This is achieved through the learning of the policy presented as:

$$W^{\pi*} = E\left[C^*_{s,a}(s_t, a_t)\right] \tag{30}$$

where

$W^{\pi*}$= Optimal policy to maximize the cumulative reward after all sensing episodes

$C^*_{s,a}$= cumulative rewardof the optimal cooperating CR users.

Three major temporal difference techniques are considered: Q-learning, Sarsa and Actor-Critic [19].
Utilizing decision epochs $t \in T = \{1,2,\dots,n\}$, the knowledge gained by a cognitive radio i based on sensing operations related to action a_t^i in state s_t^i at time t represented by Q-function is derived as follows:

$$Q_{t+1}^i(s_t^i, a_t^i) \leftarrow (1-\alpha)Q_t^i + \alpha\left[r_{t+1}^i(s_{t+1}^i) + \gamma \max_{a \in A} Q_t^i(s_{t+1}^i, a_{t+1}^i)\right] \tag{31}$$

where
$\gamma \max_{a \in A} Q_t^i(s_{t+1}^i, a_{t+1}^i)$ represents the discounted future reward

α is the learning rate which could be between 0 and 1.

The algorithms the temporal difference learning as well as the overall proposed approach is presented as follows:

Algorithm 1: Q-learning

1: Initialize Q(s,a) = 0 $\forall(\boldsymbol{s},\boldsymbol{a}) \in \boldsymbol{S} \times \boldsymbol{A}$
2: **for** each episode t, **do**
3: choose action $\boldsymbol{a}$in state $\boldsymbol{s}$ using policy based on Q-values in that state
4: $$\boldsymbol{a} = \arg\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i})$$
5: Take action $\boldsymbol{a}$, receive accompanying reward $\boldsymbol{r}$, proceed to next state $\boldsymbol{s_{t+1}^i}$
6: Update Q-value for action $\boldsymbol{a}$at state $\boldsymbol{s}$
7: $$\boldsymbol{Q_{t+1}^i}(\boldsymbol{s_t^i}, \boldsymbol{a_t^i}) \leftarrow (\boldsymbol{1}-\boldsymbol{\alpha})\boldsymbol{Q_t^i} + \boldsymbol{\alpha}\left[\boldsymbol{r_{t+1}^i}(\boldsymbol{s_{t+1}^i}) + \boldsymbol{\gamma}\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i})\right]$$
8: $\boldsymbol{s_t^i} \leftarrow \boldsymbol{s_{t+1}^i}$
9: **until** end of the states

Algorithm 2: Sarsa

1: Initialize Q(s,a) = 0 $\forall(\boldsymbol{s},\boldsymbol{a}) \in \boldsymbol{S} \times \boldsymbol{A}$
2: **for** each episode t, **do**
3: choose action $\boldsymbol{a}$in state $\boldsymbol{s}$ using policy based on Q-values in that state
4: $$\boldsymbol{a} = \arg\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i})$$
5: Take action $\boldsymbol{a}$, receive accompanying reward $\boldsymbol{r}$, proceed to next state $\boldsymbol{s_{t+1}^i}$
6: Update Q-value for action $\boldsymbol{a}$at state $\boldsymbol{s}$
7: $$\boldsymbol{Q_{t+1}^i}(\boldsymbol{s_t^i}, \boldsymbol{a_t^i}) \leftarrow \boldsymbol{Q_{t+1}^i}(\boldsymbol{s_t^i}, \boldsymbol{a_t^i}) + \boldsymbol{\alpha}\left[\boldsymbol{r_{t+1}^i}(\boldsymbol{s_{t+1}^i}) + \boldsymbol{\gamma}\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i}) - \boldsymbol{Q_{t+1}^i}(\boldsymbol{s_t^i}, \boldsymbol{a_t^i})\right]$$
8: $\boldsymbol{s_t^i} \leftarrow \boldsymbol{s_{t+1}^i}$; $\boldsymbol{a_t^i} \leftarrow \boldsymbol{a_{t+1}^i}$
9: **until** end of the states

Algorithm 3: Actor-Critic

1: Initialize Q(s,a) = 0 $\forall(\boldsymbol{s},\boldsymbol{a}) \in \boldsymbol{S} \times \boldsymbol{A}$
2: **for** each episode t, **do**
3: choose action $\boldsymbol{a}$in state $\boldsymbol{s}$ using policy based on Q-values in that state
4: $$\boldsymbol{a} = \arg\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i})$$
5: Take action $\boldsymbol{a}$, receive accompanying reward $\boldsymbol{r}$, proceed to next state $\boldsymbol{s_{t+1}^i}$
6: Update critic parameters
7: $$\boldsymbol{\Omega_{t+1}} \leftarrow \boldsymbol{\Omega_t} + \boldsymbol{\alpha}\left[\boldsymbol{r_{t+1}^i}(\boldsymbol{s_{t+1}^i}) - \boldsymbol{\rho} + \boldsymbol{\gamma}\max_{a \in A} \boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i}; \boldsymbol{\Omega_t}) - \boldsymbol{Q_{t+1}^i}(\boldsymbol{s_t^i}, \boldsymbol{a_t^i}; \boldsymbol{\Omega_t})\right]\boldsymbol{e_t}$$
8: Update average reward estimate
9: $$\boldsymbol{\rho_{t+1}} = \boldsymbol{\rho_t} + \boldsymbol{\alpha_t}(\boldsymbol{r_{t+1}^i} - \boldsymbol{\rho_t})$$
10: Update actor parameters
11: $$\boldsymbol{\Theta_{t+1}} = \boldsymbol{\Theta_t} + \boldsymbol{\beta_t}\boldsymbol{\Gamma}(\boldsymbol{\Omega_t})\boldsymbol{Q_t^i}(\boldsymbol{s_{t+1}^i}, \boldsymbol{a_{t+1}^i}; \boldsymbol{\Omega_t})$$
12: where $\boldsymbol{\beta_t}$ is positive step size parameter and $\boldsymbol{\Gamma}(\boldsymbol{\Omega_t})$ is a normalization factor
13: $\boldsymbol{s_t^i} \leftarrow \boldsymbol{s_{t+1}^i}$; $\boldsymbol{a_t^i} \leftarrow \boldsymbol{a_{t+1}^i}$
14: **until** end of the states

Algorithm 4: Algorithm of the Proposed Approach

1: Initialize $\boldsymbol{\alpha, \beta, \lambda, P} \in$ N, $\boldsymbol{t, i = 0, m = 10}$, Q(s,a) = 0 $\forall(\boldsymbol{s, a}) \in \boldsymbol{S \times A}$
2: **for** each episode t, **do**
3: $\mathbf{t \leftarrow t + 1}$
4: $\boldsymbol{T'(y_i)} \leftarrow$ test statistics of channel
5. compute $\boldsymbol{T_i^{avg}(T_i) = \frac{1}{P}\sum_{l=1}^{P} T_{i-P+l}(x_{i-P+l})}$
6.compute $\boldsymbol{\lambda_{IED} = \left(Q^{-1}\left(P_{fa,target}^{IED}\right)\sqrt{2N} + N\right)\sigma_w^2}$
7. **if** SNR>reliable threshold**then**
8. $\boldsymbol{\lambda_{IED} \leftarrow \lambda_{IED,OR}}$
9. **else**
10. $\boldsymbol{\lambda_{IED} \leftarrow \lambda_{IED,EGC}}$
11. **end if**
12: IED sensing report $\boldsymbol{R_i} \in \{\boldsymbol{H_1, H_0}\}$
13: select TD technique
14: choose action $\boldsymbol{a}$in state $\boldsymbol{s}$ using policy based on Q-values in that state
15: $$\boldsymbol{a = \arg\max_{a \in A} Q_t^i\left(s_{t+1}^i, a_{t+1}^i\right)}$$
16: Take action $\boldsymbol{a}$, receive accompanying reward $\boldsymbol{r}$, proceed to next state $\boldsymbol{s_{t+1}^i}$
17: Update learning parameters$\boldsymbol{\alpha, \tau_n, \gamma}$, and policy $\boldsymbol{W^{\pi *}}$based on TD technique selected
14: **until** end of the states

4. SIMULATION AND RESULTS

Simulation Setup

The cooperative spectrum sensing technique using RL will utilize the IED –based adaptive cooperative spectrum sensing technique described earlier in the paper.The network configuration consisting of PUs and SUs is presented in

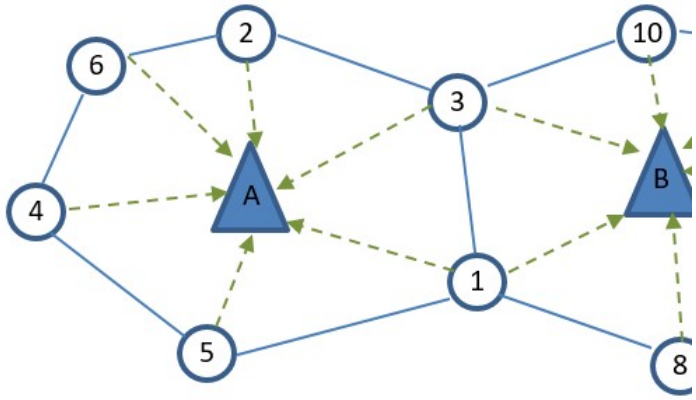

Figure 3. There 10 randomly deployed representative space $600m$ space. 2PUs coordinates (160,450). The time the CRs to report to at CR3 or CR1 are within the time $t < 20ms$

CRs which are within a of $300m \times$ are located at (150,150) and delay for each of the fusion points also randomized frame $0.1ms <$

Figure 3: Network configuration

Simulation parameters

Parameter	Value
Discount Factor	0.8
Actions	2
Simulation replications	30
Iterations for learning	10^6
SNR variation	-15dB – 15dB
Pf	0.05
No. of agents	10
Operating Frequency of PU	1 x 10^9Hz
Observation time	1 x 10^{-4}
Variance of the noise (σ^2n)	1 x 10^{-12}
Variance of the received signal (σ^2s)	(σ^2nx 10^{-1})^2

The number of cooperating CR users denoted $n$is selected from the total number of agents (m). The learning agent therefore can select a course of action in cooperation with other CR users from the states:

$$s_n \square S, 0 \leq n \leq m + 1 \quad (32)$$

where
s_n= number of states
n = number of cooperating CR users

The action and response of the starting and ending states are already predetermined as 'Begin' and 'Stop', respectively. The actual learning process therefore is within the other states where the actions are not predetermined but depends on the rewards obtained from previous actions. This invariably means the action can change and the sum total of all the actions and rewards in the cooperative sensing period forms the learning experience. MATLAB software was used as the is simulation tool. The IED algorithm, adaptation and RL incorporation was coded on the editor page of MATLAB software due easier implementation of the functions involved in the proposed algorithm.

Simulations were carried out using the aforementioned parameters to explore the performance of the energy detection –based sensing in various SNRs. This is to determine a reliable SNR threshold where the limited number of CRs can be utilized for sensing and reliable sensing results can still be obtained. The result of the performance of the improved energy detection algorithm across several SNRs is presented in **Figure 4**. This aided the selection of the threshold where the adaptive procedure could be implemented to minimize time for sensing report. In order to verify the reliability of the threshold which the adaptive cooperative scheme will utilize, a single energy detection procedure was compared with cooperative sensing. A receiver operations characteristics curve to compare the sensitivity performance of both algorithms is presented in **Figure 5**.

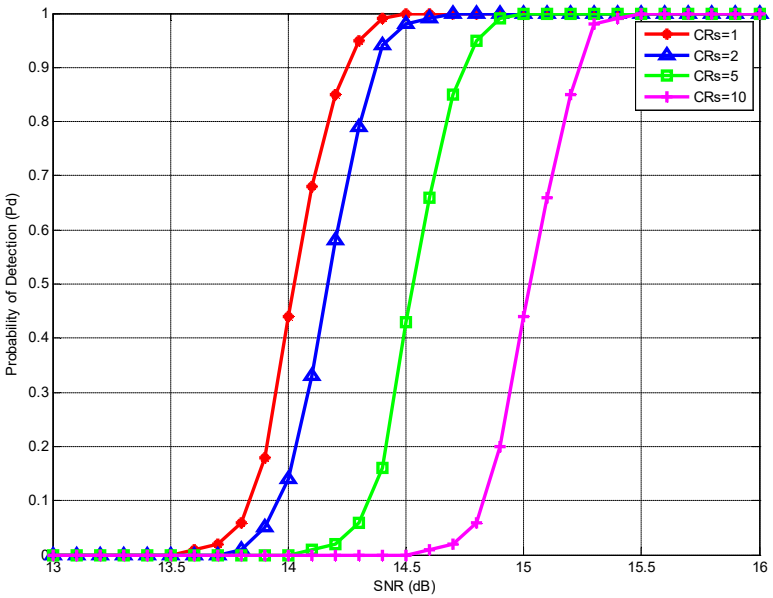

Figure 4: Performance of Cooperative Sensing using IED Scheme

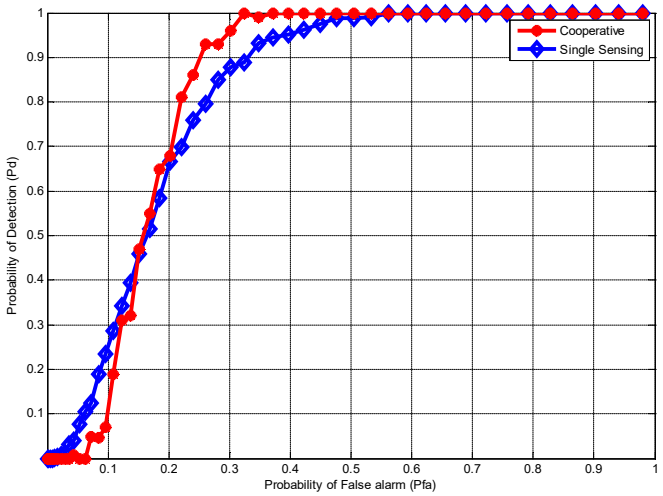

Figure 5: Comparison of Cooperative and Non-Cooperative Spectrum Sensing

The results comparing the various learning rates and the time consumption improvement is presented in **Figure 6**and **Figure 7** respectively.

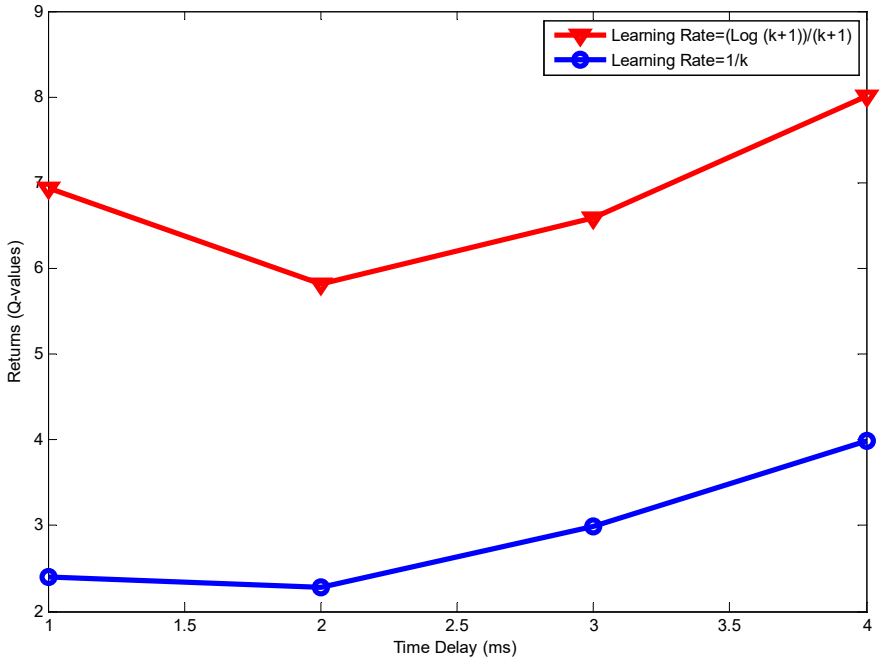

Figure 6:Results to Compare the Two Learning Rates Explored

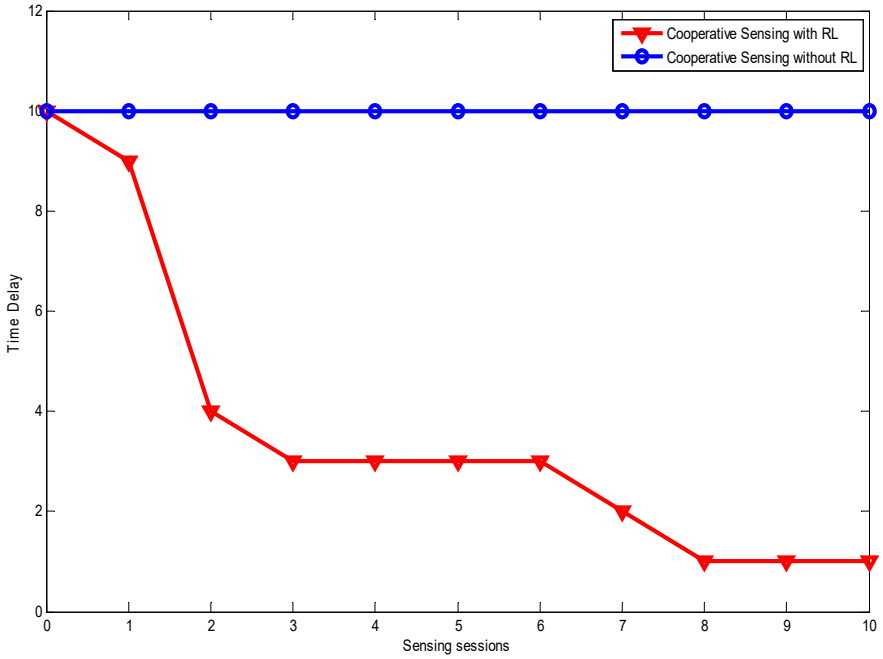

Figure 7: Comparison of Time Consumption for the Adaptive Cooperating Sensing Algorithm with Conventional Cooperating Sensing without RL

5. DISCUSSION

Performance of the IED-based cooperative sensing under varying SNR conditions

The result presented above in **Figure 4**shows the performance of the Improved Energy Detection (IED) scheme when several CRs are employed in sensing under varying SNR conditions. This is a scenariorepresenting a very noisy environment where with a high noise level. Yet, it is observed from the result that detections commence from about 13.6dB and increases rapidly to maximum detection probability within a range of about 1dB. This confirms the sensitivity of the IED-based cooperative sensing scheme.

It is also observed that the performance improves with increase in the number of cooperative cognitive radios. This reveals that multiple sensors give a more accurate result since the individual reports of each of the cognitive radios are collated before the final decision is taken. Hence, cooperative sensing done with more number of cognitive radios had more accurate detections than those with less cooperating users.However, as observed in **Figure 5**, both cooperative and non-cooperative sensing performed satisfactorily at about 15.5dB. This is therefore, used as a threshold for the adaptive cooperative spectrum sensing where the number of cooperating CRs could be reduced in order to minimize time spent on sensing reports.

Effects of the Various Temporal Difference (TD) learning techniques on performance of the Adaptive Cooperative Spectrum Sensing Scheme.

Three different Temporal Difference (TD) learning algorithms were considered namely: actor-critic, SARSA and Q-learning techniques. The learning time for the three TD techniques varied during the simulation trials conducted, but Q-learning constantly had higher cumulative reward each time and produced results at an average of 3% faster than SARSA and actor-critic. Previous researches such as [2][5], [7] also point to Q-learning in cooperative sensing as a technique with higher cumulative rewards. Q-Learning was therefore selected for the further simulations carried out in this research.

Comparison of Returns Obtained Using Different Learning Rates

Two learning rates were explored while the reinforcement learning was being developed for the adaptive cooperative sensing scheme. These are the normal learning rate α_{norm} and the logarithmic learning rateα_{log} given by the equations 7 and 8 respectively.

$$\alpha_{norm} = {}^{1}\!/_{m} \tag{33}$$

$$\alpha_{log} = {}^{\log(m+1)}\!/_{(m+1)} \tag{34}$$

where:

m = number of cognitive radios

α = learning rate

The result comparing both learning rates with respect to the returns (Q-values) obtained as presented in **Figure 6**shows that the logarithmic learning rate gave much higher Q-values than the conventional learning rate. The higher Q-values seen could be explained to be a difference resulting from the rate of convergence of the learning. A learning rate which results in higher Q-values is often synonymous with more exploration than exploitation of rewards which implies slower convergence. It can therefore be deduced that the algorithmic learning rate α_{log} which gave higher returns in terms of Q-values converged slower than the α_{norm}. This implies opportunity for more exploration than exploitation to produce better outcome. This result provided the basis for the use of the logarithmic learning rate in the process of reinforcement learning.

Optimal Outcome Using Proposed Approach

After employing Q-learning along with the logarithmic learning rate, a major difference in sensing report time was observed. It was observed in simulations of 1000 sensing episodes that the cooperative sensing delay time which initially hovered around a mean of 10msgradually dropped to a mean of 1ms after about 10 sensing sessions. This is the period it took the fusion point (learning agent) to learn the sensing report characteristics of the individual CRs.This is observed in **Figure 7**. This is simulated with reference to a conventional cooperative spectrum sensing algorithm where all 10 cooperative users are constantly in use.

This translates to a reduction in time by about 60 percent compared to the time spent by all the CR users in a conventional cooperative spectrum sensing technique which is assumed to utilized all 10 CRs during sensing. This is seen in **Figure 8**.This is an interesting result which fulfils the objective of this research - to cut down on the cumulative time consumed for individual sensing report and processing.

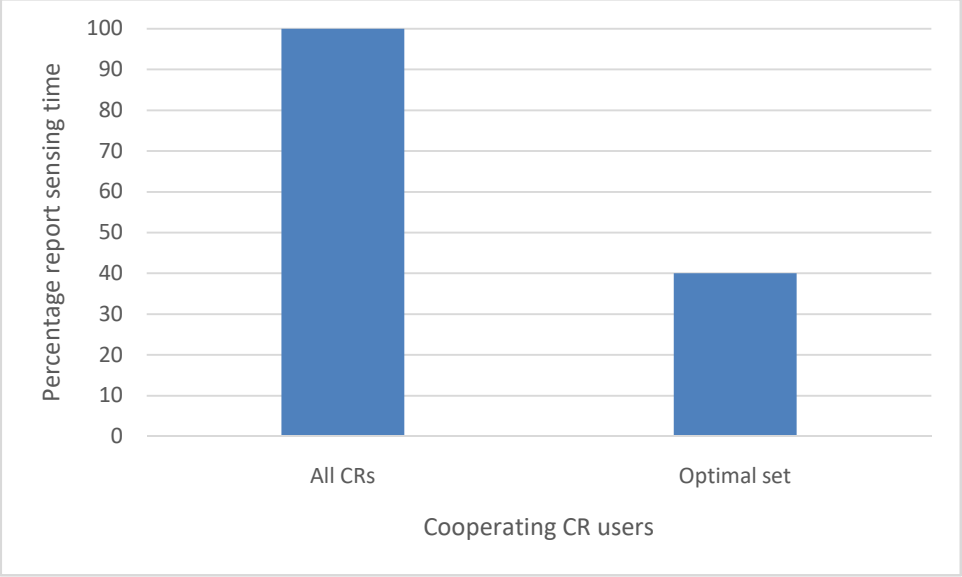

Figure 8: Percentage Time Consumption for the Adaptive Cooperating Sensing Algorithm and Conventional Cooperative Sensing Algorithm

6. CONCLUSION

In this paper, the sensitivity of energy detection based spectrum sensing was increased using Improved Energy Detection (IED) algorithm. This was then utilized in an Adaptive Cooperative spectrum scheme which was aimed at reducing the time consumed during the processing of spectrum sensing reports. The adaptive spectrum sensing scheme made it possible to utilize a few cognitive radios for sensing when the SNR status of the channel was reliable. This reduced the number of reports that needed to be processed during cooperative sensing. It also reduced the delay involved in report processing time. Reinforcement learning was incorporated in the adaptive spectrum sensing algorithm in order to further reduce sensing report collation delay. The introduction of reinforcement learning aided the decision of the fusion point which was the learning agent to learn the policy of selecting an optimal set of cognitive radios for cooperative decision. This further reduced the number of cooperative users that need to be considered when cooperative sensing reports need to be made. It alsominimized the time delay that would have been incurred when all the cognitive radios were required to give their sensing report irrespective of the quality and promptness of their report. This technique introduced helped improve on the time lapse between sensing and decision making in cognitive radio systems.This provides a solution to one of the major open issues in cooperative spectrum sensing using RL where the network performance has been improved at the expense of higher amount of control overhead[6]. This technique provides an effective solution to reduce the overhead of the cooperative sensing while preserving the simplicity of the network.

REFERENCES

[1] A. M. Mikaeil, "Machine Learning Approaches for Spectrum Management in Cognitive Radio Networks," in *Machine Learning - Advanced Techniques and Emerging Applications*, IntechOpen, 2018, pp. 117–139.

[2] R. Vishnu, I. Dias, T. Tholeti, and K. Sheetal, "Spectrum Access In Cognitive Radio Using a Two-Stage Reinforcement Learning Approach," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 20–34, 2018.

[3] Xiao Yang and Hu Fei, *Cognitive Radio Networks*. Taylor and Francis Group, LLC, 2009.

[4] C. Sachin, "Spectrum Sensing for Cognitive Radios: Algorithms, Performance, and Limitations," Aalto University, 2012.

[5] B. F. Lo and I. F. Akyildiz, "Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks," *IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, pp. 2244–2249, 2010.

[6] K. A. Yau, G. Poh, S. F. Chien, and H. A. A. Al-rawi, "Application of Reinforcement Learning in Cognitive Radio Networks : Models and Algorithms," *Hindawi Publ. Corp.*, vol. 2014, 2014.

[7] N. Hosey, S. Bergin, I. Macaluso, and O. Diarmuid, "Q-Learning for Cognitive Radios," in *China-Ireland Information and Communications Technologies Conference*, 2009.

[8] M. Lopez-Benitez and Casadevall F., "Improved energy detection spectrum sensing for cognitive radio," *IET Commun.*, vol. 6, no. 8, May 22 2012, pp. 785–796, 2012.

[9] Y. Chen, "Improved Energy Detector for Random Signals in Gaussian Noise," *IEEE Trans. Wirel. Commun.*, vol. 9, no. 2, 2010.

[10] F. D. C. Paisana, "Spectrum Sensing Algorithms for Cognitive Radio Networks Master in

Electrical and Computer Engineering Jury Members," 2012.

[11] Y. Eghbali, H. Hassani, A. Koohian, and M. Ahmadian-Attari, "Improved Energy Detector for Wideband Spectrum Sensing in Cognitive Radio Networks."

[12] M. Kanti, D. Barma, H. Singh, S. Roy, and S. K. Sen, "Augmented Spectrum Sensing in Cognitive Radio Networks," *IJCSN Int. J. Comput. Sci. Netw.*, vol. 4, no. 6, 2015.

[13] U. R. Fatih, "Spectrum Sensing Techniques for Cognitive Radio Systems With Multiple Antennas," 2010.

[14] J. Zhu, Y. Song, D. Jiang, and H. Song, "Multi-armed bandit channel access scheme with cognitive radio technology in wireless sensor networks for the Internet of Things," *IEEE Access*, vol. 4, pp. 4609–4617, 2016.

[15] M. Jain, V. Kumar, R. Gangopadhyay, and S. Debnath, "Cooperative Spectrum Sensing using Improved p-norm Detector in Generalized κ-μ Fading Channel," in *International Conference on Cognitive Radio Oriented Wireless Networks*, 2015.

[16] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, 2011.

[17] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation," *2010 IEEE Symp. New Front. Dyn. Spectrum, DySPAN 2010*, no. May, 2010.

[18] H. Urkowitz, "Energy Detection of Unknown Deterministic Signals," *Proc. IEEE*, vol. 55, no. 4, 1967.

[19] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, London, England, 2017.

[20] Juliani Arthur, "Simple Reinforcement Learning with Tensorflow Part 7: Action-Selection Strategies for Exploration," 2016. [Online]. Available: https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-7-action-selection-strategies-for-exploration-d3a97b7cceaf. [Accessed: 22-Feb-2019].