논문 2010-47TC-11-10

Triple-state 보상 함수를 기반으로 한 개선된 DSA 기법

(An Improved DSA Strategy based on Triple-States Reward Function)

타사미아*, 구 준 롱*, 장 성 진*, 김 재 명**

(Tasmia Ahmed, Junrong Gu, Sung Jeen Jang, and Jae Moung Kim)

요 약

본 논문은 보상함수 수정을 통해 보다 완벽한 DSA(Dynamic Spectrum Access)를 수행하는 새로운 방법을 제시하였다. POMDP(Partially Observable Markov Decision Process)는 미래의 스펙트럼 상태를 예측하는데 사용되는 알고리즘으로서, 그 중 보상함수는 스펙트럼을 예측하는데 있어 가장 중요한 부분이다. 그러나 보상함수는 Busy 및 Idle의 두 가지 상태만 갖고 있기 때문에 채널에서 충돌이 발생하게 되면 보상함수는 Busy를 반환함으로써 2차 사용자의 성능을 감소시키게 된다. 따라서 본 논문에서는 기존의 Busy를 Busy 및 Collision 의 두 상태로 구분하였고, 이렇게 추가된 Collision 상태를 통해 2차 사용자의 채널 접근 기회를 보다 향상시킴으로서 데이터 전송율을 증대시킬 수 있도록 하였다. 또한 본 논문은 새로운 알고리즘의 신뢰도 벡터를 수학적으로 분석하였다. 마지막으로 시뮬레이션 결과를 통해 개선된 보상함수의 성능을 검증하고, 이를 통해 새로운 알고리즘이 CR 네트워크에서 2차 사용자의 성능을 향상시킬 수 있음을 보인다.

Abstract

In this paper, we present a new method to complete Dynamic Spectrum Access by modifying the reward function. Partially Observable Markov Decision Process (POMDP) is an eligible algorithm to predict the upcoming spectrum opportunity. In POMDP, Reward function is the last portion and very important for prediction. However, the Reward function has only two states (Busy and Idle). When collision happens in the channel, reward function indicates busy state which is responsible for the throughput decreasing of secondary user. In this paper, we focus the difference between busy and collision state. We have proposed a new algorithm for reward function that indicates an additional state of collision which brings better communication opportunity for secondary users. Secondary users properly utilize opportunities to access Primary User channels for efficient data transmission with the help of the new reward function. We have derived mathematical belief vector of the new algorithm as well. Simulation results have corroborated the superior performance of improved reward function. The new algorithm has increased the throughput for secondary user in cognitive radio network.

Keywords : Dynamic Spectrum Access; Partially Observable Markov Decision Process; Reward Function.

Ⅰ. Introduction

The proliferation of a wide range of wireless devices has resulted in an overly crowded radio spectrum. In contrast to this scarcity in spectrum availability is the pervasive existence of spectrum opportunities. Dynamic Spectrum Access (DSA) is one of the approaches envisioned for dynamic spectrum management in cognitive radio networks[1~2]. The basic idea of DSA is to allow secondary users to identify and exploit spectrum opportunities under the constraint that they do not cause harmful interference to primary users. Most of the existing works on DSA strategies assume the presence of a

* 학생회원, ** 평생회원, 인하대학교 정보통신대학원
(INHA-WiTLAB, Graduate School of IT & Telecommunication)
※ 이 논문은 2010년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2010-0008000)
※ 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2010-C1090-1011-0007)
접수일자: 2010년6월23일, 수정완료일: 2010년11월10일

primary network[3]. A basic component of DSA is a sensing strategy at the MAC layer for spectrum opportunity tracking. Since the secondary user may not be able to sense all channels in the spectrum simultaneously. A sensing strategy for intelligent channel selection is crucial to track the rapidly varying spectrum opportunities. By modeling primary users' channel occupancy as a Markov Process, the design of sensing strategies is formulated as a Partially Observable Markov Decision Processes (POMDPs)[1, 3]. POMDPs are extensions of Markov Decision Processes (MDPs) in which the system states are not completely observable. In POMDPs, a secondary user interacts with a stochastic environment at discrete time steps. The secondary user takes actions and as a result, receives observations and rewards. The user then has to find a way of choosing actions, or policy, which maximizes the total reward received over time. POMDP method tries to construct a Markovian-state information using a dynamic scheme and the history of actions and observations experienced by the user. This information is called a belief vector. Then POMDP method uses reward information in order to associate an action to belief vector[4]. For maximizing the throughput of secondary users while limiting the probability of collision with primary users, the joint PHY-MAC design of Opportunistic Spectrum Access as a constrained POMDP is formulated in [5]. The decision-theoretic approach integrates the design of spectrum access protocols at the MAC layer with spectrum sensing at the physical layer and traffic statistics determined by the application layer of the primary network based on the theory of Partially Observable Markov Decision Process[6].

In communication system, Collision happens when secondary user senses the channel idle and starts transmission and primary user returns to the band before the secondary user finishes its transmission. The collision happens between the primary user and secondary user which generated by imperfect sensing during the sensing period. In collision case, the reward function of POMDP formulation shows busy. This busy state may be caused by false alarm or actual existence of primary user. The false alarm indicates the idle channel as busy state.

In this paper, we propose a new algorithm by modifying reward function of POMDP to get better communication opportunity for secondary users. We have separated collision from busy state. The secondary users with the new reward function are able to detect collided channel using collision state. The busy state and collision state are not the same because the causes are different from each other. The behaviors of imperfect sensing are the false alarm and miss detection. False alarms result in wasted spectrum opportunities while miss detections lead to collisions between users.

There are another differences between collision and busy state. In collision case neither secondary users can transmit data while primary user can send data successfully in busy condition. The proposed reward function is able to reduce the wasted time and decrease complexity of the spectrum access. Our simulation results also show collision does not affect throughput of the secondary user if collision state can be detected.

The paper is organized as follows. In section Ⅱ, the Partially Observable Markov Decision Process and our algorithm are introduced. The modified reward function is analyzed in section Ⅲ. In section Ⅳ, the simulation results are given. Section Ⅴ concludes the paper.

Ⅱ. System Model

In this section, we introduce how the collision happens in a PU (Primary User) network and SU (Secondary User) network coexistence scenario. The we introduce the famous POMDP model. Afterwards, we propose our model specified by a characterized reward function.

1. Background Introduction

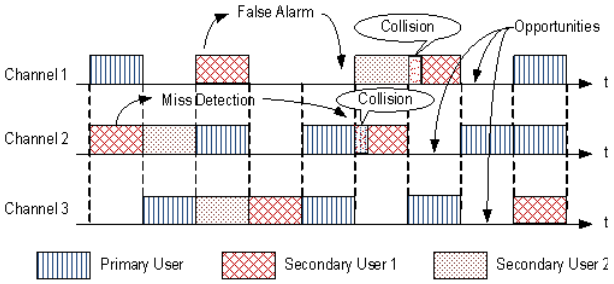

그림 1. 다중 사용자간 충돌 시나리오
Fig. 1. Scenarios of collision between multi-users.

Figure 1 shows the Scenarios of different condition cases. In this multi-channel environment, There happens successful data transmission and failed transmission (collision) between both primary user and secondary users. False alarm and miss detection are shown in this Figure. Collision occurs when secondary user senses the channel idle and starts transmission while primary user returns back to the channel before the secondary user completes its transmission. A secondary user is not able to know the belief vector of other secondary user, and then unexpected collision happens. Here, error sensing is responsible for collision while secondary user completely trusts the sensing outcomes in making access decisions.

In channel 1, the user misses the accessing opportunity because of false alarm and on the other hand the collision happens between two secondary users by detecting busy channel as an idle channel. The channel 2 has miss detection that creates collision between primary user and secondary user. The channel 3 has no sensing error of secondary users' detection. This channel is a good accessing channel. There is no sensing error of secondary channel that may cause the accessing miss-opportunity. A collision establishes between two secondary users because they have different belief vectors due to their different observation histories. A secondary user does not know the other secondary user's belief vector unless through explicit cooperation.

The model of Figure 2 shows the network of the primary users and secondary users. Assume the

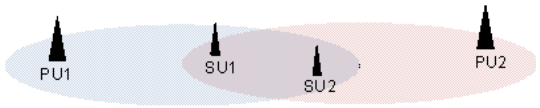

그림 2. 우선사용자와 2차 사용자간 채널에 대한 시나리오
Fig. 2. Scenario of channel between primary user and secondary user.

spectrum is divided into L independent channels that are allocated to a time synchronized slot t based primary network with multiple primary users. In the secondary network, there are N users; each chooses one channel to sense at the beginning of each time slot and transmits if an unused channel exists.

Figure 2 shows the scenario considering two secondary users (N = 2) contend with each other but perceive different primary users. Spectrum opportunities for secondary users (SU1 and SU2) are determined by primary users (PU1 and PU2).

2. A Constrained POMDP Formulation

From [4~7], we have formulated the opportunistic channel access sequence as a POMDP represented by $(S, A, P_{s,s'}, O, a, R)$ given below.

State Space S The system state is given by the State Space S of each channel at the beginning of each slot. The state space is $S = \{0,1\}^L$.

Past observation λ At the beginning of slot t, our knowledge of the system state based on all past decisions and observations can be summarized by a belief vector.

Sensing Action A is the sensing action profile for all the users that sense at the beginning of each time slot.

State Transition Probabilities $P_{s,s'}$ is a set of Markovian state transition probabilities. $P_s(s,s') = \Pr\{s(t+1) = s' | s(t) = s\}$, denotes the probability of

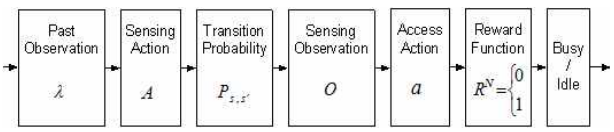

그림 3. 기존의 POMDP 알고리즘 순서도
Fig. 3. Conventional POMDP algorithm sequence.

being at state $\acute{s}$ at time slot $t+1$ when given that at time slot t.

Observation Space O is the observation space. For each user, there are two kinds of observation in a time slot t: Busy, Idle. "Busy" means the channel is occupied by the primary user in this slot. So the secondary user must defer to the next slot. "Idle" means if the channel is sensed as empty, the secondary user transmits a packet.

Probability of Observation P_o represents the probability that all action A for state s at time slot t will give observation O, i.e. $P_o(s,A,O)=\Pr\{O(t)=O|s(t)=s,A(t)=A\}$ *Reward Function* R represents the reward function mapping from the observation space O to real numbers. $R^N(t)$ is the reward for secondary user N in time slot t defined as follows:

$$R^N(t)=\begin{cases}0 \text{ if } O^N(t) is\ Busy \text{ or } Collision\\ 1 \text{ if } O^N(t) is\ Idle\end{cases} \quad (1)$$

Policies A sensing policy π is a policy to decide for each secondary user what action to take in each time slot.

$$\pi=\max_{\pi} E_{\pi}\left[\sum_{t=1}^{T} R(t)|\Lambda(1)\right] \quad (2)$$

Where E_{π} represents the expectation given policy π is employed and $\Lambda(1)$ is the initial belief vector.

Myopic policy in belief vector which ignores the impact of the current action on the future reward, focusing solely at maximizing the immediate reward. Moreover, a straightforward solution to the channel selection problem is to employ the greedy policy, i.e., the policy of maximizing the expected instantaneous reward[7]. The myopic policy under information state of channel, $\overline{\omega}=[\omega_1,\ldots,\omega_N]\in\{0,1\}^N$ is given by $\hat{a}(\overline{\omega})=\max_{a=1,..,N} E\left[R_a(\overline{\omega})\right]$. Here, $R_a(\overline{\omega})$ is the reward collected under state $\overline{\omega}(t)$ when channel $a(t)=\pi(\overline{\omega}(t))$ is selected by the secondary user with sensing policy(π). $\hat{a}(\overline{\omega})$ is the belief vector based on the access action $\hat{a}(t)$ under myopic policy. In general, obtaining the myopic action in each time slot requires the successive update of the information state, which explicitly relies on the knowledge of the state transition probabilities $\{P_{s,\acute{s}}\}$ as well as the initial condition or belief vector $(\overline{\omega}(t))$.

3. Reward Function Characterization Algorithm

To overcome the unwanted collision, we propose a new algorithm to detect the collision and maximize the throughput. We introduced a new variable alpha (α) that indicates collision which is used in reward function. Hence, the expected total reward of the POMDP represents overall throughput of secondary user, the expected total number of bits that can be delivered by the secondary user in T slot.

In Figure 4, at the beginning of data transmission, the secondary user tries to choose any set of channels for sensing using all past decisions and observations to sense. The current state of underlying Markov process is $\acute{s}$,the secondary user observes (observation O) which indicates the availability of each sensed channel. Based on the observation O, these secondary user chooses a channel to access. Based on this access action a, the secondary user gets there reward function R.

$$R^N(t)=\begin{cases}0 \text{ if } O^N(t) is\ Busy\\ 1 \text{ if } O^N(t) is\ Idle\\ \alpha \text{ if } O^N(t) is\ Collision\end{cases} \quad (3)$$

In (3), the collision state is not same as the busy state. There are three observation results in slot t: Busy, Collision, and Idle. In this paper, we especially focus on this formulation. "Busy" means the channel

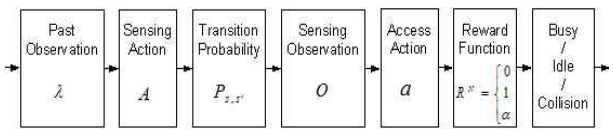

그림 4. 보상함수 특징이 부가된 POMDP 알고리즘

Fig. 4. POMDP algorithm with characterized reward function.

is occupied by the primary user in a slot. The secondary user can choose another channel or wait for the busy slot to be free[8]. "Idle" means the channel is free and the secondary user can transmit a packet on it. If a primary user returns back to the channel that is being used by a secondary user, collision happens. In the "Collision" state, both of the primary user and secondary user are unable to transmit data. If a secondary user finds a channel in collision state, it should leave for another channel.

To distinguish the busy and collision state, we can employ the new reward function in the sensor. The secondary user gets the knowledge about the channel condition while the secondary user is trying to sense a chosen channel. When secondary user finds collision happened, secondary user pause their data transmission and switch to another channel. Because, failed communication occurs in collision state. The channel switching can bring a great advantage to increase throughput for secondary user. In busy state user can switch to other channel or wait for the slot to be free[8]. We have separated collision and busy state in reward function by adding new variable α that represents collision. We have got 0.5 as a proper value by testing different a values in the simulation section.

Ⅲ. Analysis of Reward Function Algorithm

1. Imperfect Sensing

Due to hardware limitations and energy constraints, a secondary user may not be able to sense all the channels in the spectrum simultaneously. In this case, a sensing strategy for intelligent channel selection to track the rapidly varying spectrum opportunities is necessary. When sensing error occurs, the state space and the sensing outcome are not same $S(t) \neq O_A(t)$. The purpose of the sensing strategy is two folds: Catch a spectrum opportunity for immediate access and obtain statistical information on spectrum occupancy so that more rewarding sensing decisions can be made in the future. A tradeoff has to be reached between these two often conflicting objectives[3]. This collision state happens for miss detection in sensor. The lost opportunity namely "busy" is caused by false alarm. The limitations of MAC layer induce the false alarm and miss detection. If the secondary user completely trusts the sensing outcomes in making access decisions, false alarms result in wasted spectrum opportunities while miss detections lead to collisions with primary users[5].

2. Reward Function of Proposed Algorithm

We got these three equations based on the reward function under the policies. These equations are the combine form of (2) and (3).

Here, R_T^{π} denotes the system reward of the policy π, which is defined as the expected reward or the

$$R_T^{\pi}(\overline{\omega}) = \max_{\pi} \sum_{t=1}^{T} \lambda^{t-1} E_{\pi}\left[R_{\pi_t}(\overline{\omega}(t)) | \overline{\omega}(t) = 1\right]; \quad R(\overline{\omega}) = 1 \tag{4a}$$

$$R_T^{\pi}(\overline{\omega}) = \max_{\pi} \sum_{t=1}^{\infty} \lambda^{t-1} E_{\pi}\left[R_{\pi_t}(\overline{\omega}(t)) | \overline{\omega}(t) = 0\right]; \quad R(\overline{\omega}) = 0 \tag{4b}$$

$$R_T^{\pi}(\overline{\omega}) = \max_{\pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \lambda^{2t-1} \cdot E_{\pi}\left[R_{\pi_t}(\overline{\omega}(t)) | \overline{\omega}(t) = \alpha\right]; 0 < R(\overline{\omega}) < 1 \tag{4c}$$

expected total number of bits that can be delivered by the secondary user in T slot. Equation (4a) represents idle channel. This equation has two rewards with only past observation and decision histories (λ^{t-1}). Equation (4b) indicates the reward function of the busy channel and the last reward function (4c) indicates the collision case and this equation has two observations and decisions histories: the past all observations and decisions (λ^{t-1}) and other is the instant observation and decision (λ^{t}) in a slot t by which other secondary user can understand that the slot is occupied. The past observation histories are the result of the sensing and the instant observation and decision is for the result

of the miss detections. They would stop transmitting data and search for another channel.

3. Mathematical Analysis of Belief Vector

In this section, we have showed mathematically the difference between busy and collision state. We have analyzed that the belief vector values of busy and collision states as well as reward are not the same. In the busy state, secondary user has zero reward. In the collision state, the secondary user has reward that associated the condition of channel. The authors in [5] used this formula for maximizing the throughput of secondary users while limiting the probability of collision with primary users using the joint MAC-PHY design. At the beginning of each slot t, a sensing policy specifies a set $A(t)$ of channels to be sensed based on the current belief vector $\Lambda(t)$ and the sensing outcomes $O_A(t)$. The system state based on all past decisions and observations can be summarized by a belief vector, $\Lambda(t)=[\lambda_1(t),\lambda_2(t),..]$ where $\lambda(t)$ is the decision and observation history in slot t. The reward function can be accumulated starting from slot t consists of two parts: the immediate reward $R_{\pi_t}=\sum_{a\in A}k(t)B_a$ and the maximum expected future reward $V_{t+1}(\Lambda(t+1))$ where $\Lambda(t+1)=\tau(\Lambda(t)|A,K_A)$ that represent the updated belief vector for slot $t+1$ after action and observation acknowledgement[5]. The belief vector value of Figure 3 with the all possible observation acknowledgement K_A with maximizing over all actions A under the myopic policy is representing as

$$V_t(\Lambda(t))=\max_A \sum_{s\in S}\sum_{s'\in S}\lambda_{s'}(t)P_{s,s'}\Pr\{K_A=k_A|S=s\}\cdot\left[\sum_{a\in A}k_aB_a+V_{t+1}(\tau(\Lambda(t)|A,K_A))\right] \quad (5)$$

In the idle channel, the number of events is moved to the queue during the slot T. Since such an event represents gain of one reward. In the idle channel, equation (6) is modified form of (5). We get the belief vector value in the idle channel,

$$V_t(\Lambda(t))=\max_A \sum_{t=1}^{T}\lambda^{t-1}E\left[R_{k_t}+V_{t+1}(\tau(\Lambda(t)|A,1))\right] \quad (6)$$

In the busy channel, the number of events is moved to the end of the queue. Since such an event represents gain of zero reward. We get (7) that is modified form with past observation and decision history(λ^{t-1}) of (5) in the busy channel. The belief vector value of secondary user in busy channel,

$$V_t(\Lambda(t))=\max_A \sum_{t=1}^{\infty}\lambda^{t-1}E\left[R_{\pi_t}+V_{t+1}(\tau(\Lambda(t)|A,0))\right] \quad (7)$$

Derivation: See Appendix A.

In the collided channel, the number of events is moved to the slot $\lim_{T\to\infty}\frac{1}{T}$ of the queue because the secondary user does not need to sense that collided channel again and move to other unused channel. Since such an event represents gain of valued reward. In collision state, the failed transmission may occur, acknowledgements are necessary to ensure the result of the transmission outcome. It is seemed that the process has a limiting distribution. As a consequence, the limit in (8) exists. There have two observation and decision histories: past observation and decision history (λ^{t-1}) and instant observation and decision history (λ^t) for taking decision to search the other channel. The belief vector value of secondary user and we also use (α) to indicate the outcome in the collided channel,

$$V_t(\Lambda(t))=\max_A \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\lambda^{2t-1}E\left[R_{k_t}|A,\alpha\right] \quad (8)$$

Derivation: See Appendix B.

These three equations (6), (7) and (8) are the belief vector of (4a), (4b) and (4c) respectively.

Ⅳ. Simulation

In this section, We have tested our new algorithm with different collision degrees (different a values). we also have compared the performance of traditional reward function and new reward function by the throughput of secondary user.

1. α value test

Figure 5 shows the throughput of secondary user with different a values. We have considered one independent channel with bandwidth normalized to one $(B_a = 1)$ during 15 seconds. We have regarded collision (α) values (0.3, 0.4, 0.5, 0.6, 0.7, and 0.8) of different reward function with greedy policy.

Figure 5 shows that α with 0.5 has better throughput than all the other values. It shows that 0.3 and 0.4 alpha values give lower throughput than that of 0.5 in every second. Though 0.6, 0.7, 0.8 give better throughput in the initial stage. The collision happens in 12 second. After 12 th second, 0.5 (alpha value) offers highest throughput.

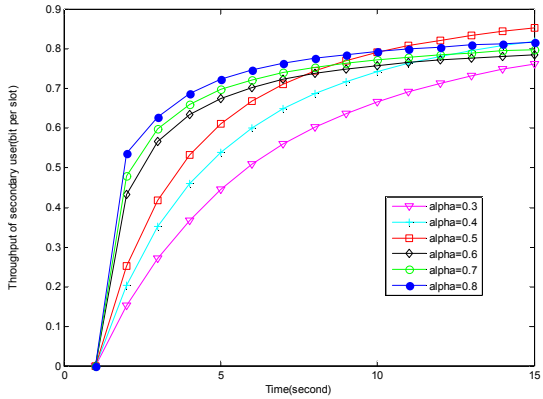

그림 5. α값에 따른 2차 사용자의 성능
Fig. 5. Comparison of different alpha values.

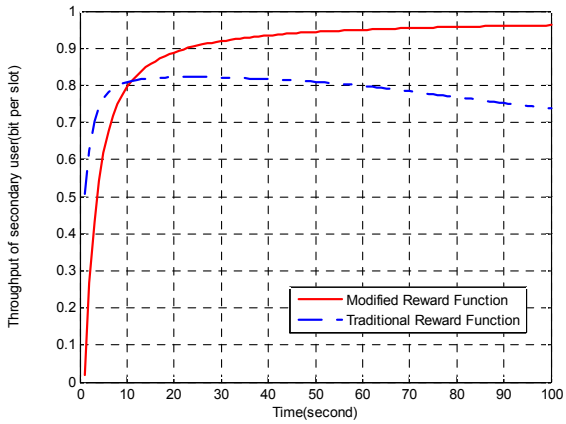

그림 6. 시간에 따른 2차 사용자의 성능
Fig. 6. Comparison of throughput (bit per slot).

2. Throughput comparison of different algorithms

In Figure 6, we have considered one independent channel with bandwidth $(B_a = 1)$ in 100 seconds under greedy policy. In case of traditional reward function, there are two states in reward function (Idle=1 and Busy=0) while in case of modified reward function, the reward function with additional collision state (Idle=1, Busy=0 and Collision=0.5).

At the beginning of Figure 6, there is no collision so both curves increase with time. At 10th second, a collision occurs in the channel. Form this time on the traditional reward function decreases slowly with time but the one of modified reward function increases without being changed. The noticeable point is at 50th second. From this point on the traditional reward function is decreasing rapidly while the modified one is increasing without being changed. In Figure 6, the traditional reward function cannot detect collision so it needs to wait for next slot. The secondary user is not able to sense the entire communication channel. So the user has to wait for vacancy among the sensed channels. This is the main reason for the throughput decreasing. However, the secondary user with modified reward function can detect collision. So the users are looking for other channel without wasting time by waiting for the next slot.

Ⅴ. Conclusion

In this paper, we have analyzed a new reward function algorithm with additional collision state. This new reward function algorithm makes it easier for secondary user to access spectrum in cognitive radio

network. The algorithm with new reward function keeps down the complexity of transmission data for secondary users and also allows identifying and exploiting spectrum opportunities without any decrease in throughput. The other secondary user does not need to wait for the next available slot in the collided channel. We also show the difference between busy state and collision state. So we cannot use one indicator for both busy and collision state.

In future, based on this new reward function, we also can use (busy, collision) tone when two users try to contend for the same slot. If one user of them accesses the channel successfully, the succeeded user can use busy tone. When collision happens the collided secondary user use the collision tone, other secondary users can understand that slot is congested by secondary users. They will not transmit data and search for another channel.

Appendix

In this section, we have derive the belief vector in the Idle, Busy and Collided channel. In appendix A, we have derived the belief vector of the equation (4a) and (4b). similarly we have derived the belief vector of the equation (4c) in appendix B.

Appendix A: Derivation of the belief vector in the idle channel and busy channel

At the beginning of the data transmission, the secondary user tries to choose a set of channels for sensing using all past decisions and observations to sense. With the current belief vector and the sensing outcomes the secondary user observes which indicates the availability of each sensed channel. Based on the observation , the secondary user chooses a channel to access. The system state based on all past decisions and observations can be summarized by a belief vector. The belief vector value $V_t(\Lambda(t))$ is defined as,

$$V_t(\Lambda(t))= \max_{A} \sum_{s\in S}\sum_{s'\in S}\lambda_{s'}(t)P_{s,s'}\Pr\{K_A=k_A|S=s\}\cdot \left[\sum_{a\in A}k_aB_a+V_{t+1}(\tau(\Lambda(t)|A,K_A))\right] \quad (9)$$

Here, the maximum expected future reward $V_{t+1}(\tau(\Lambda(t)|A,K_A))$ which represents the updates belief vector for slot $t+1$ after action and observation acknowledgement.

The conditional distribution $U_{s,k_A}(\Lambda(t))$ of the acknowledgement of all actions can be calculated as $U_{s,k_A}(\Lambda(t))\triangleq Pr\{K_A=k_A|S=s\}$ is the conditional distribution of the acknowledgement given current state s as well as action A and the reward function $R_{k_A(t)}=\sum_{a\in A}K_a(t)B_a$, the channel bandwidth is B_a [5]. Here, $\sum_{s\in S}\sum_{s'\in S}\lambda_{s'}(t)P_{s,s'}$ is a constant for given belief vector.

$$V_t(\Lambda(t))= \max_{A} \sum_{s\in S}\sum_{s'\in S}\lambda_{s'}(t)P_{s,s'}\sum_{K_a\in\{0,1\}}U^{(L)}_{s,k_A}(\Lambda(t))\cdot \left[R_{k_t}+V_{t+1}(\tau(\Lambda(t)|A,K_A))\right] \quad (10)$$

The sensing action a, the expected immediate reward $E\left[R_{k_A(t)}|\Lambda(t)\right]$ may happens based on the all previous channel observation and decisions as λ^{t-1},

$$V_t(\Lambda(t))= \max_{A} \sum_{t=1}^{T}\lambda^{t-1}E\left[R_{k_t}+V_{t+1}(\tau(\Lambda(t)|A,K_A))\right] \quad (11)$$

In the idle channel, There has valued acknowledge to ensure the transmitter about the idle channel and the one unit of reward from the receiver which represents the success of the data transmission.

$$V_t(\Lambda(t))= \max_{A} \sum_{t=1}^{T}\lambda^{t-1}E\left[R_{k_t}+V_{t+1}(\tau(\Lambda(t)|A,1))\right] \quad (12)$$

In the busy channel, there has no acknowledgement and the zero reward represents the failed transmission happens in the busy channel.

$$V_t(\Lambda(t)) = \max_{A} \sum_{t=1}^{\infty} \lambda^{t-1} E\left[R_{k_t} + V_{t+1}(\tau(\Lambda(t)|A,0))\right] \quad (13)$$

Appendix B: Derivation of the belief vector in the collided channel

In the collision state, There have two histories: the past observation and decision history (λ^{t-1}) and the instant observation and decision history (λ^t). The total observation and decision history is $((\lambda^{t-1} \bullet \lambda^t) = \lambda^{2t-1})$. The steady-state value of belief vector under the myopic policy is defined as $U(\Lambda(t)) \triangleq \lim_{T\to\infty} \frac{V_T(\Lambda(1))}{T}$. $V_T(\Lambda(1))$ is the expected total reward obtained in $[t = 0,1,...,T]$ slots under the myopic policy when initial belief is $\Lambda(t)$. Here, $U(\Lambda(t))$ is determined by the Markov reward process$(S(t), R(t))$. The value of belief vector with the all past decision history in collision state is defined as,

$$V_t(\Lambda(t)) = \max_{A} \sum_{t=1}^{T} \lambda^{t-1} \lim_{T\to\infty} \frac{V_T(\Lambda(t))}{T} \bullet \left[B_A + V_{t+1}(\tau(\Lambda(t)|A,\alpha))\right] \quad (14)$$

The sequence of actions a taken at times $(t = 0,1,2,....T)$ and the initial state S and the initial observation λ^t for any action $V_T(\Lambda(t)) = \sum_{t=0}^{T} \lambda^t E\left[r(B_A, V_{t+1}(\tau(\Lambda(t))))\right]$. An event represents gain of one reward in collided channel. The reward accumulated in $[t,(t+1)]$ period; $R_{k_t} = r(B_A, V_{t+1}(\tau(\Lambda(t),A)))$. Equation (15) represents the value of belief vector in the collision.

$$V_t(\Lambda(t)) = \max_{A} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \lambda^{2t-1} E\left[R_{k_t}|A,\alpha\right] \quad (15)$$

Here, we get the three equations which will lead to the optimal solution.

Reference

[1] Qing Zhao, Bhaskar Krishnamachari, "Structure and Optimality of Myopic Sensing for Opportunistic Spectrum Access," *Communications, 2007. ICC '07 IEEE International Conference on*, pp.6476 - 6481, June 2007.

[2] Sachin Shetty, Min Song, Chunsheng Xin, E. K. Park, "A Learning-based Multiuser Opportunistic Spectrum Access Approach in Unslotted Primary Networks," *INFOCOM 2009, IEEE*, pp.2966-2970, April 2009.

[3] Qing. Zhao and B.M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal. Processing Magazine,* pp.79-89, May 2007.

[4] George E. Monahan, "State of the Art-A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms," *MANAGEMENT SCIENCE Vol.28, No.1,* pp.1-16, January 1982.

[5] Yunxia Chen, Qing Zhao, Swami, A., "Joint PHY-MAC design for opportunistic spectrum access in the presence of sensing errors," *Information Theory, IEEE Transactions on,* Volume 54, Issue 5, pp.2053-2071, May 2008.

[6] Q. Zhao, L. Tong, A. Swami, and Y. Chen "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework," *IEEE J. Select. Areas Commun.,* vol.25, no.3, pp.589 - 600, Apr. 2007.

[7] Jayakrishnan Unnikrishnan, Venugopal V. Veeravalli, "Algorithm for Dynamic Spectrum Access with Learning for Cognitive Radio," *Signal Processing, IEEE Transactions on,* pp.750-760, Feb. 2009.

[8] Alex Chia-Chun Hsu, David S. L. Weiand C.-C.JayKuo, "A Cognitive MAC Protocol Using Statistical Channel Allocation for Wireless Ad-hoc Networks," *Wireless Communications and Networking Conference*, pp.105-110, 2007.

[9] Md. Imrul Hassan, Ju Bin Song, Young-Il Kim, "Throughput Capacity of a Wireless Multi-hop Relay Network using Cognitive Radio," *Journal of IEEK* Volume 44, Issue 5, pp.33-39, May 2007.

[10] I. Hassan, Chang-Bae Rho, Ju Bin Song, "System Throughput of Cognitive Radio Multi-hop Relay Networks," *Journal of IEEK* Volume 46, Issue 4, pp.29-39, Apr. 2009.

저 자 소 개

타사미아(학생회원)
2005년 Daffodil International대학교 전자통신공학과 학사
2009년~현재 인하대학교 정보통신대학원 석사 과정
<주관심분야 : 무선인지기술>

구 준 롱(학생회원)
2005년 중국 동북대학교 통신공학 학사
2008년 중국 대련이공대학교 통신정보계통공학 석사
2008년~현재 인하대학교 정보통신대학원 박사 과정
<주관심분야 : 이동통신, 무선인지기술>

장 성 진(학생회원)
2007년 인하대학교 전자공학과 학사
2009년 인하대학교 정보통신대학원 석사
2009년~현재 인하대학교 정보통신대학원 박사 과정
<주관심분야 : 이동통신, 무선인지기술>

김 재 명(평생회원)-교신저자
1974년 한양대학교 전자공학과 학사
1981년 미국남가주대학교(USC) 전기공학과 석사
1987년 연세대학교 전자공학과 박사
1974년 3월~1979년 6월 한국과학기술연구소, 한국 통신기술연구소 근무
1982년 9월~2003년 3월 한국전자통신연구원 위성통신연구단장/무선방송연구소 소장
2003년 4월~현재 인하대학교 정보통신대학원 교수, 통신위성 우주산업연구회 회장 외 기술 자문으로 다수 활동 중
<주관심분야 : 차세대 무선이동통신 및 Cognitive Radio, UWB