Docitive Networks – An Emerging Paradigm for Dynamic Spectrum Management

Lorenza Giupponi, Ana Galindo-Serrano, Pol Blasco and Mischa Dohler

***Abstract*—Prime design goals for next generation wireless networks to support emerging applications are spectral efficiency and low operational cost. Among a gamut of technical solutions, cognitive approaches have long been perceived as a catalyst of above goals by facilitating the coexistence of primary and secondary users by means of efficient dynamic spectrum management. Whilst most available techniques today essentially are opportunistic in nature, a truly cognitive device needs to exhibit a certain degree of intelligence to draw optimum decisions based on prior observations and anticipated actions. Said intelligence however comes along with a high complexity and poor convergence, which currently prevents any viable deployment of cognitive networks. We thus introduce an emerging and largely unexplored concept of docitive networks where nodes effectively teach other nodes with the prime aim to reduce cognitive complexity, speed up the learning process and draw better and more reliable decisions. To this end, we review some important concepts borrowed from the machine learning community for both centralized and decentralized systems, in order to position the emerging docitive with known cognitive approaches. Finally, we validate introduced concepts in the context of a primary digital television system dynamically coexisting with IEEE 802.22 secondary networks. For this scenario, we demonstrate the superiority of various unprecedented docitive over known opportunistic/cognitive algorithms.**

I. INTRODUCTION

Spectrum itself is not scarce. The spectral sub-5GHz region over which networking between wireless devices is facilitated at reasonable costs, however, is. Not only is it highly congested, but also used fairly inefficiently as of today [1]. A variety of techniques at physical, medium access and networking layers have hence emerged over past decades, which aim at improving upon this efficiency. One promising approach relies on cognition, a fairly abstract concept in itself, applicable to virtually all layers, algorithms and technologies [2]. Cognition, from “cognoscere” = “to know” in Latin, is generally defined as “a process involved in gaining knowledge and comprehension, including thinking, knowing, remembering, judging, and problem solving”. Promising numerous advantages over static or opportunistic approaches, cognition has been the focus of numerous disciplines in the past, such as biology, biomedicine, telecommunications, computer science, etc.

Whilst a rigorous and ubiquitous quantification of intelligence is still an open problem, it is generally recognized that intelligence is impacted by the degree of observation, the ability to learn, the amount of memory, among others. These concepts have been investigated by the artificial intelligence and machine learning communities, which propose numerous cognitive approaches capable of finding optimal decision policies in extremely dynamical scenarios characterized by only one decision maker. However, the wireless setting in general and the cognitive radio scenario in particular are not always characterized by a node centralizing the radio resource management (RRM) decision process.

As a matter of fact, there has been lately a clear trend towards decentralizing RRM functionalities, as it is the case for IEEE P1900.4 standardization efforts [3]. The implications of learning in decentralized settings are still not fully understood even by the machine learning community and constitute an ongoing research line referred to as “multi-agent learning”. The underlying interactive learning processes of the different nodes generate oscillating behaviors, which not always reach an equilibrium. The dynamics of learning may be long and complex in terms of required operations and memory, with complexity increasing with an increasing observation space. Other drawbacks are the slow rate of convergence and poor scalability, which essentially prevent these techniques from being used in dynamic large-scale decentralized systems [4].

Even as cognition and learning have received a considerable interest from various communities in the past, the process of knowledge transfer, i.e. teaching, over the wireless medium however has received fairly little attention to date. We thus aim at introducing an emerging framework referred to as docitive radios, from “docere” = “to teach” in Latin, which relates to radios (or general networking entities) which teach other radios. These radios are not (only) supposed to teach end-results (e.g. in form of “I sense the spectrum to be occupied”), but rather elements of the methods of getting there. This concept mimics well our society-driven pupil-teacher paradigm which generally acknowledges that inferior teachers teach end-result whereas good teachers facilitate learning. It will be shown to capitalize on the advantages but, most importantly, mitigate major parts of above-mentioned drawbacks of purely cognitive radios.

Whilst applicable to a variety of problems in communications, we will assume coexistence of a single primary system with several secondary systems. We will apply said docitive techniques to RRM algorithms of the secondary systems' dynamic spectrum management suites. More specifically, to demonstrate the superiority of docitive over cognitive techniques, we have chosen to concentrate on a cognitive power allocation strategy, such that capacity in secondary systems is maximized, without jeopardizing the primary system.

Generally, concepts and mathematical techniques related to cognition stem from the artificial intelligence and game theory communities [4], [5]. The teaching process, on the other

---

The authors are with the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Parc Mediterrani de la Tecnologia, Av. Carl Friedrich Gauss 7, Barcelona, Spain 08860; phone: +34-93-645-2900, fax: +34-93-645-2901, email: {name.surname}@cttc.es.

hand, involves transmission over the wireless medium which is core to the wireless communications community. It also requires cooperation between nodes which has successfully been used in all of the prior mentioned communities. The docitive approach thus constitutes a centroid of techniques and algorithms previously developed in these and other communities. Important and unprecedented questions thus arise when applied to dynamic spectrum management, part of which will be dealt within this paper.

To this end, the paper is structured as follows. In Section II, we summarize the main cognitive algorithms that have been applied to RRM for wireless networks; to the best of the authors' knowledge, they refer to scenarios where a centralized node is in charge of managing radio resources. This is why, in Section III, we focus on a scenario where the intelligence is delivered to and thus distributed among multiple nodes; we introduce and rationalize the concept of docition and its application to wireless networks as a paradigm to improve the learning capabilities of distributed networks. In Section IV, we apply the introduced paradigms to the IEEE 802.22 standard for Wireless Regional Area Networks (WRANs) and demonstrate their superiority. Finally, we conclude by highlighting pertinent open research issues in Section V.

II. PERTINENT COGNITIVE ALGORITHMS

A high-level operational cycle of cognitive radios introduced in [6] is depicted in Figure 1 (upper-left) which, in the context of RRM, relies on the following elements:

- **Acquisition.** The acquisition unit provides quintessential information of the surrounding environment, such as spectrum occupancy or interference temperature. This data can be obtained by means of numerous methods, such as sensing performed by the node itself and/or in conjunction with spatially adjacent cooperating nodes; docitive information from neighboring nodes or databases; etc.
- **Intelligent Decision.** The core of a cognitive radio is without doubt the intelligent decision engine, which typically learns from past experiences gathered from e.g. dynamics of interference or statistics of spectral occupancy. Based on some intelligent algorithms, it then draws decisions on choice of band and resource block, transmission power, etc.
- **Action.** With the decision taken, an important aspect of the cognitive radio is to ensure that the intelligent decisions are being carried out, which is typically handled by a suitably reconfigurable software defined radio (SDR), some policy enforcement protocols, among others.

We will henceforth concentrate on the intelligent decision engine which, according to a verbal definition coined by Haykin [2], needs to draw intelligent decisions by learning the dynamics of the environment, taking into account past and present system states, learning from its own policies, and generally using long term benefit estimations. In mathematical terms, this can be formulated as a Markov decision process (MDP).

More precisely, an MDP is a discrete time stochastic control process which provides a mathematical framework for modeling centralized decision-making in situations where outcomes are partly random and partly under the control of a decision maker. It is defined based on a finite state set $\mathcal{S} = \{s(1), \ldots, s(q)\}$. At each time step t, the controller observes the system's current state (e.g. occupancy of the spectrum, signal to noise and interference ratio (SINR), etc.) and selects an action (e.g. transmit, not transmit, transmission parameters, etc.), which is executed by being applied as input to the system. Let us assume that $s(i)$ is the observed state, and that the action is selected from a finite set of admissible actions $\mathcal{A} = \{a(1), \ldots, a(m)\}$. When the central node executes action $a(k) \in \mathcal{A}$, the system's state at the next step changes from $s(i)$ to $s(j)$, with a state transition probability $P_{i,j}$. We further assume that the application of action $a(k)$ in state $s(i)$ incurs an immediate cost $c(i,k)$. The objective of the MDP is to find a policy that minimizes the cost of each state [4].

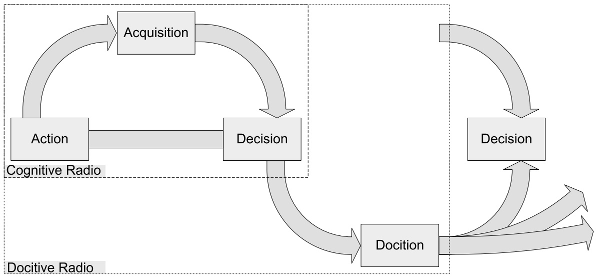

Fig. 1. Canonical cognitive cycle and its extension through docition.

In the machine learning literature, two ways have been identified to solve MDPs. The first one is an analytical model-based approach, which relies on the knowledge of the state transition probability function between two states after executing a certain action. The second one, in turn, does not rely on this previous knowledge making it a model-free approach; it is based on reinforcement learning (RL). RL is a family of learning approaches which is mainly concerned with the development of algorithms that automatically learn the properties of the environment and adapt their behavior to them by means os trial and error. This allows it to gather experience on the run, and to be able to adapt to the temporal dynamics of the system. Numerous embodiments of RL exist (Q-learning, temporal difference, SARSA, etc.) [7] where we concentrate on Q-Learning, which in its very nature is suited for dynamical wireless systems [8].

Q-Learning works by estimating a Q-function, the Q-values $Q(s,a)$ of which, for each state-action pair, are stored in a Q-table. The value $Q(s,a)$ is defined to be the expected discounted sum of future cost obtained by taking action a from state s and following an optimal policy thereafter. Once these values have been learned, the optimal action from any state is the one with the highest Q-value. After being initialized to arbitrary numbers, Q-values are estimated on the run, on the basis of experience as follows [9]: 1) from the current state s, select an action a, which will cause a receipt of an immediate cost and arrival at a next state $s^{'}$; 2) update $Q(s,a)$ based upon this experience; and 3) go to step 1). More details are given in Algorithm 1.

Algorithm 1 Q-Learning.

```
Initialize:
for each $s \in \mathcal{S}$, $a \in \mathcal{A}$ do
    initialize the Q-value representation mechanism $Q(s,a)$
end for
evaluate the starting state $s(i) \in \mathcal{S}$
Learning:
loop
    generate a random number $r$ between 0 and 1
    if $(r < \varepsilon)$ then
        select action randomly
    else
        select the action $a(k) \in \mathcal{A}$ characterized by the minimum Q-value
    end if
    execute $a(k)$
    receive an immediate cost $c(i,k)$
    observe the next state $s^{'} = s(j)$
    update the table entry as follows:
        $Q(s,a) \leftarrow Q(s,a) + \alpha[c + \gamma \min_a Q(s^{'},a) - Q(s,a)]$
    $s(i) = s(j)$
end loop
```

The prime advantages of this approach are that Q-Learning learns from past and current experiences, draws intelligent decisions, and it can be demonstrated that in case of complete information it is able to find an optimal policy for making decisions in a dynamical system. Most importantly, however, it is also able to find sufficiently well performing policies under partial knowledge, which is often the case in many real world problems. The resulting formal model is called a Partially Observable Markov Decision Process (POMDP) [10], [11].

A slightly extended taxonomy of the one briefly discussed in this section is depicted in Figure 2.

III. PRIMER ON DOCITIVE NETWORKS

As illustrated in Figure 1, the canonical cognitive cycle is advantageously extended by the following element:

- **Docition.** It is realized by means of an entity which facilitates knowledge dissemination and propagation with the non-trivial aim to facilitate learning.

It is inspired by the so-far-successful problem based learning (PBL) concept used at schools in our society. In PBL, teachers are encouraged to be coaches and not information givers with the aim to have pupils work as a team using critical thinking to synthesize and apply knowledge; apprehend through dialogue, questioning, reciprocal teaching, and mentoring.

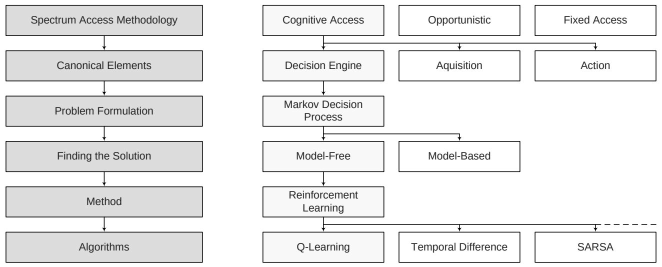

Fig. 2. Taxonomy of pertinent cognitive approaches.

Translated back to the wireless setting, this implies a distributed approach where nodes share potentially differing amounts of intelligence acquired on the run. This, in turn, is expected to sharpen and speed up the learning process. Any achieved gains, however, need to be gauged against the overhead incurred due to the exchange of docitive information. The range of application scenarios is vast, including infrastructure-less cognitive radio networks, novel cellular systems such as femtocells, etc. The distributed learning and teaching paradigm applied to these novel networking architectures, however, raises unprecedented questions, where we first concentrate on learning and subsequently on teaching issues.

As for distributed learning approaches, the characteristics of the decentralized scenarios are as follows: (1) the intelligent decisions are made by multiple intelligent and uncoordinated nodes; (2) the nodes partially observes the overall scenario; and (3) their inputs to the intelligent decision process are different from node to node since they come from spatially distributed sources of information. Multiple nodes have thus to distributively learn an optimal policy to achieve a common objective. Known as "multi-agent learning" problem, it can be solved by means of distributed RL approaches such as distributed Q-learning. However, in this field many problems still remain open – even for machine learning experts. The main challenge is how to ensure that individual decisions of the nodes result in jointly optimal decisions for the group, considering that the standard convergence proof for Q-learning does not hold in this case as the transition model depends on the unknown policy of the other learning nodes.

In principle, it is possible to treat the distributed cognitive radio network as a centralized one, where each node has complete information about the other nodes and learns the optimal joint policy using standard RL techniques. However, both the state and action spaces scale exponentially with the number of nodes, rendering this approach infeasible for most problems. Alternatively, we can let each node learn its policy independently of the other nodes, but then the transition model depends on the policy of the other learning nodes, which may result in oscillatory behaviors and in slow speed of convergence to prior set targets [12]. This introduces game-theoretic issues to the learning process, which are not yet fully understood [13].

As for teaching approaches, some early contributions in literature [14] suggest that the performances of a decentralized learning system can be improved by using cooperation among learners in a variety of ways. A node e.g. can take advantage of the exchange of information and expert knowledge from other nodes [14], the so-called docitive nodes. Depending on the degree of docition among nodes, we consider in this paper the following cases:

- **No Docition.** Nodes do not cooperate, ignore the actions and rewards of the other nodes in the system and learn their strategies independently. The standard convergence proof for Q-learning does not hold in this case. In particular, each agent's adaptation to the environment can change the environment itself in a way that may make the other nodes' adaptations invalid. Despite that, this method has been applied successfully in multiple cases [11], [12].

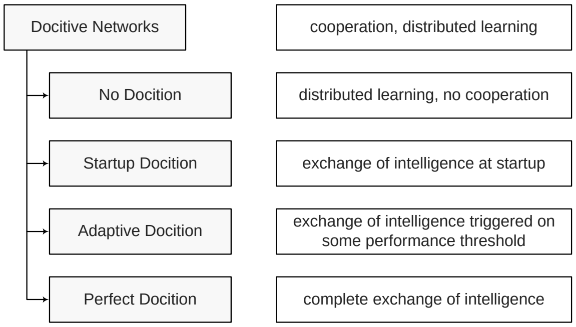

Fig. 3. Taxonomy of docitive algorithms with different degrees of docition.

- **Startup Docition.** Docitive radios teach their policies to any newcomers joining the network. In this case, again, each node learns independently; however, when a new node joins the network, instead of learning from scratch how to act in the surrounding environment, it learns the policies already acquired by more expert neighbors. Gains are expected due to a high correlation in the environments of adjacent expert and newcomer nodes.
- **Adaptive Docition.** Docitive radios here share policies, based on performances. The nodes cooperate by exchanging information about the performances of their learning processes, e.g. the variance of the oscillation with respect to the target, the speed of convergence, etc. Based on this information, each node may learn from expert neighbors who are performing better, i.e. are more intelligent.
- **Perfect Docition.** The multi-user system can be regarded as an intelligent system in which each joint action is represented as a single action. The optimal Q-values for the joint actions can be learned using standard centralized Q-learning. In order to apply this approach, a central controller should model the MDP and communicate to each node its individual actions. Alternatively, all nodes should model the complete MDP separately and select their individual actions; in this case, no communication is needed between the nodes but they all have to observe the joint action and all individual rewards. Although this approach leads to the optimal solution, it is infeasible for problems with many nodes since the joint action space, which is exponential in the number of nodes, becomes intractable. This is why it will not be analyzed in the next section.

The degree of cooperation, and thus the overhead, augments with an increasing degree of docition. The optimum operating point hence depends on the system architecture, performance requirements, etc. A summary of the taxonomy introduced in this section is shown in Figure 3.

IV. DOCITIVE RADIOS FOR IEEE 802.22 COEXISTENCE

We evaluate the proposed docitive framework in the context of IEEE 802.22 standard specifications for WRANs systems. The primary system is characterized by a DTV broadcasting station (hereafter primary BS), working at 615 MHz, located in the center of a circular cell, and several DTV receivers (hereafter primary receivers) randomly located in the DTV's coverage area. The primary BS transmits at P_{DTV}=1 MW (90 dBm) effective radiated power (ERP) with an antenna height of 500 m.

The secondary WRAN cell, with radius r_{SS}, operates in the same channel as the primary system, and consists of a point-to-multipoint wireless air interface whereby a secondary BS manages the medium access of the associated M secondary users (SUs). Since the IEEE 802.22 standard relies on orthogonal frequency division multiple access (OFDMA)/ time division multiple access (TDMA), we consider the simplifying hypothesis that the WRAN BS assigns at any time all the available OFDMA subchannels in the 6 MHz channel bandwidth to one SU which results in only one SU transmitting at any time from one WRAN cell. The secondary BSs antenna height is 75 m. The secondary users are located randomly around the secondary BS, which is in charge of allocating a transmission power level to them. The $l = 20$ available power levels for SUs range from -80 dBm and 29.84 dBm ERP. With respect to the propagation models among nodes we consider the international telecommunication union recommendation (ITU-R) P.1546-1, where the lognormal shadowing parameter is fixed at 5.54 dB. We consider that the WRAN systems switch on randomly due to varying traffic requirements in the given geographical area. We study the effect of interference generated by the SU uplink transmissions onto the primary DTV system.

Considering that the secondary system is unaware of the position of the passive primary receivers, the SUs have to operate far enough from the primary BS in order not to cause harmful interference with the primary system. In IEEE 802.22, a protection region is defined around the primary BS as a geographical limit where the primary receivers must not receive harmful interference. To investigate the gains due to docition, we consider self-organized SU power control approach which maintains the interference at the protection contour below a given threshold. This is to guarantee that the aggregated interference at the primary receivers is even lower.

The system scenario is depicted in Figure 4, where the secondary system consists of multiple tiers of secondary cells located around the boundary of the keep-out region. In the following, we will refer to the tangent point where the $n-$th secondary BS coverage area and the protection contour intersect as the control point X_n. For these points, interference needs to be controlled. In our system, the multiple nodes with distributed learning and docitive capabilities are the secondary BSs. We identify the system state, action, associated cost and the next state, to apply docition in the context of decentralized Q-Learning to this scenario:

- **State.** As the system uses a decentralized Q-learning algorithm, the state is defined based on the local views of each WRAN system. The system state of node n at time *t* is defined as:

  $$s_t^n = \{I_t^n, d_t^n\} \tag{1}$$

  where $n \in \{1, 2, \ldots, N\}$ is the WRAN cell index. $I_t^n \in I$ represents a binary indicator to specify whether the secondary system is generating an aggregated interference above or below the threshold of the primary receivers.

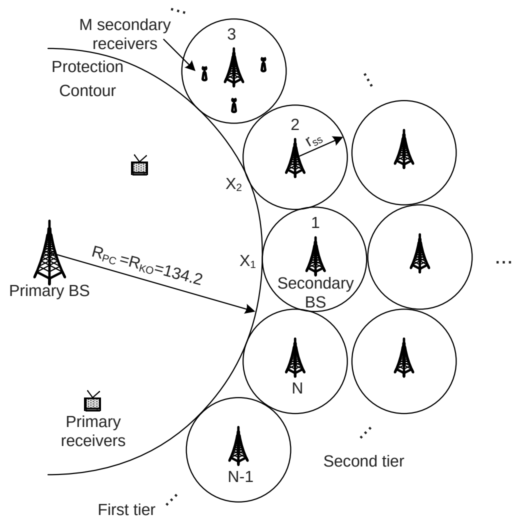

Fig. 4. Docitive scenario composed of a DTV cell and WRAN cells.

This measure is based on the instantaneous SINR value computed or estimated at the control point of the $n-$th WRAN cell. $d_t^n \in D$, where $D = \{1, 2, \ldots, d\}$, indicates an approximate distance between the secondary user and the protection contour. This information is supposed to be reported by the SUs to the secondary BS together with the spectrum sensing information during the collaborative spectrum sensing procedure [15].

- **Actions.** The set of possible actions is the set P of power levels that the secondary BS can assign to the $i-$th SU.
- **Cost.** The cost c_t^n assesses the immediate return incurred due to the assignment of action a at state s. The considered cost function is:

$$c = (SINR_t^n - SINR_{Th})^2 \tag{2}$$

  where $SINR_t^n$ is the instantaneous SINR in the control point of WRAN cell n and $SINR_{Th}$ is the threshold of 23 dB defined by the Federal Communication Commission (FCC) for DTV receivers. Q-learning aims to minimize this cost, so that the SINR at the control points is $SINR_{Th}$, which guarantees that interference at the primary receivers is below the threshold.
- **Next State.** The state transition from one state to another is determined by the power allocation of the secondary user.

The starting point for the docitive investigations are the results presented in [11]. Here, the authors analyze a cognitive approach based on independent learners and demonstrate its superiority with respect to other opportunistic approaches. The reason is that the cognitive approach is capable of capturing and adapting to the dynamics of the environment, depending on multiple factors such as shadowing, variable number of WRAN systems that randomly switch on and off, etc. We now compare the performance of (1) no docition, i.e. independent distributed learning; (2) startup docition, i.e. radios sharing their policies to newcomers; and (3) adaptive docition, i.e. radios sharing policies based on performance criteria.

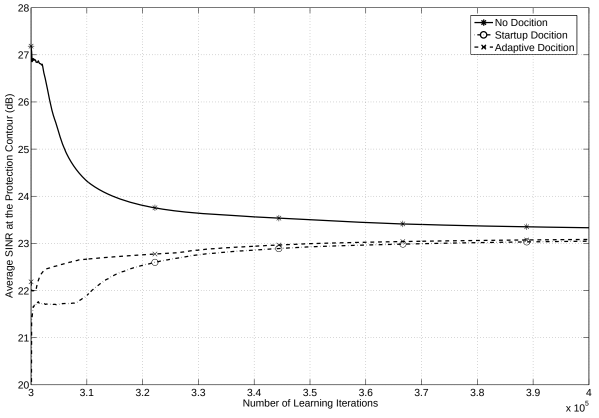

Fig. 5. Convergence of the average SINR at the control point w.r.t. the specified target of 23 dB.

The algorithm of startup docition is implemented in such a way that when a WRAN system switches on, it is able to acquire the whole Q-table of its closest neighbor. The rationale behind this algorithm is that each node does not have to learn its policy on its own, but rather can start learning from its neighbor's policy. The proximity of the two nodes guarantees that the system states of the two nodes are typically correlated and so their decision policies.

On the other hand, the adaptive docition is implemented periodically, so that each secondary system can acquire significant rows of the Q-table of the neighbor that, at any moment, is experiencing best results. The performance of the learning process of each learning node is analyzed by comparing the variance of the SINR at the corresponding control point with respect to the target of $SINR_{Th}$=23 dB. The communication overhead of the proposed schemes consists in the transmission of the Q-table at start-up and in the periodic transmissions of its rows. Based on [11], a Q-table consists of 480 bytes and each row of 80 bytes; such a small overhead is natively supported by typical inter-basestation control planes. For instance, IEEE 802.22 supports inter-basestation communication [16] by means of the Co-Existence Beacon Protocol (CBP).

Figure 5 shows the convergence curves of one of the secondary WRAN systems for the three docitive approaches. It can be observed that the docitive paradigms clearly speed up the learning process with respect to the case of independent learners. In particular, if we tolerate a margin of 1 dB in the algorithms' convergence, adaptive docition needs 900 learning iterations to achieve convergence, startup docition 10,900 and no docition 14,700 iterations. The achieved gain in speed of convergence is thus an improvement by an order of magnitude when adaptive docition is used.

As for the performance in terms of precision, i.e. oscillations around the target SINR, Figure 6 depicts the complementary cumulative distribution function (CCDF) of the variance of the SINR at the control point with respect to the set target of $SINR_{Th}$=23 dB. It can be observed that due to the distribution of intelligence among interactive learners the paradigm of docition stabilizes the oscillations by reducing the variance of

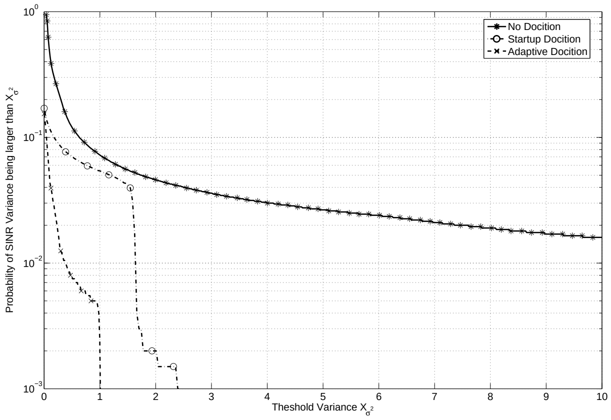

Fig. 6. CCDF of the variance of the SINR at the control point with respect to the specified target.

the SINR with respect to the specified target. More precisely, at an example target outage of 1 %, we observe that adaptive docition is about twice as good as startup docition and by several orders of magnitude better than no docition.

Finally, we have utilized the entropy as one possible measure to quantify the "intelligence" of a cognitive algorithm where a cleverer algorithm increases the order at the reference point and hence decreases the entropy measured there. For the three cases, we respectively obtained an entropy of 3.67 (no docition), 3.01 (startup docition) and 2.73 (adaptive docition). An increase in intelligence i.e. a gradient of more than 34% is observed between different degrees of intelligence.

V. CONCLUDING REMARKS

Docitive radios and networks emphasize on the teaching mechanisms and capabilities of truly cognitive networks. It is understood to be a general framework encompassing prior and emerging mechanisms in this domain. Whilst the exchange of end-results among cooperatively sensing nodes has been explored in the wireless communications community and the joint learning via exchange of states has been of very recent interest in the machine learning community, no viable framework is available to date which quantifies the gains of a docitive system operating in a wireless setting.

The aim of this paper was therefore to introduce a working taxonomy for docitive systems and position them rigorously w.r.t. opportunistic and cognitive systems. We have demonstrated that benefits by order(s) of magnitude can be achieved if nodes teach other nodes so that the general cognitive learning process is improved. Notably, applied to the example of a DTV system coexisting with IEEE 802.22 WRAN networks, we have shown that docition applied at startup as well as continuously on the run yields significant gains in terms of convergence speed and precision.

Due to its completely distributed nature and only a provision of a low-bandwidth cooperating interface, the docitive approach may thus be of greatest interest to developing cognitive radio standards. Examples of such standards are IEEE P1900, 802.22, 802.16h, 802.11y; ETSI TC RRS; ITU-R WP 1B, 5A; etc, which increasingly rely on distributed and cognitive deployments. Also femto networks and emerging 3GPP systems may benefit from docition, as the X2 interface can be readily used to exchange low-bandwidth docitive information.

Numerous interesting problems emerge across various communities in the context of docitive radios. For instance, from an information theoretical point of view, one of the core problems is how to quantify the degree of intelligence of a cognitive algorithm. With this information at hand, intelligence gradients can be established where docition should primarily happen along the strongest gradient. This would also allow one to quantify the tradeoff between providing docitive information versus the cost to deliver it via the wireless interface. Some other pertinent questions encompassing also the physical and medium access control layers are how much information should be taught, can it be encoded such that learning radios with differing degrees of intelligence can profit from a single multicast transmission, how much feedback is needed, how often should be taught, etc?

We believe that we just touched upon the tip of an iceberg as preliminary investigations have shown that docitive networks are a true facilitator for utmost efficient management and utilization of scarce spectral resources and thus an enabler for emerging as well as unprecedented wireless applications.

REFERENCES

[1] FCC, "Et docket no 03-222 notice of proposed rule making and order, december 2003." 2003.
[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Select. Areas Commun*, vol. 23, pp. 201–220, Feb. 2005.
[3] IEEEP1900 home page. [Online]. Available: http://www.ieeep1900.org
[4] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.
[5] D. Fudenberg and L. D., *The Theory of Learning in Games*. MIT Press, 1998.
[6] J. I. Mitola and G. Q. J. Maguire, "Cognitive radio: making software radios more personal," *IEEE [see also IEEE Wireless Communications] Personal Communications*, vol. 6, no. 4, pp. 13–18, Aug. 1999.
[7] M. E. Harmon and S. S. Harmon, "Reinforcement learning: A tutorial," 2000.
[8] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1676 –1687, Sept 1999.
[9] J. Watkins and P. Dayan, "Technical note: Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.
[10] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," pp. 362–370, 1995.
[11] A. Galindo-Serrano and L. Giupponi, "Decentralized Q-learning for aggregated interference control in completely and parially observable cognitive radio networks," in *In Proc. of IEEE Consumer Communications and Networking Conference, IEEE CCNC 2010, (BEST PAPER AWARD)*, Las Vegas, USA, 9-12 Jan. 2010.
[12] ——, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. on Vehicular Technology*, to appear.
[13] P. Hoen and K. Tuyls, "Analyzing multi-agent reinforcement learning using evolutionary dynamics," in *Proc. of the 15th European Conference on Machine Learning (ECML)*.
[14] M. Tan, *Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents*. M. N. Huhns and M. P. Singh, editors, Readings in Agents, pages 487494. Morgan Kaufmann, San Francisco, CA, USA., 1993.
[15] C. Cordeiro, K. Challapali, and D. Birru, "IEEE 802.22: An introduction to the first wireless standard based on cognitive radios," *Journal of Communications*, vol. 1, no. 1, April 2006.
[16] C. Stevenson, C. Cordeiro, E. Sofer, and G. Chouinard, "Functional requirements for the 802.22 WRAN standard," *IEEE doc.: IEEE 802.22-05/0007r46*, Sept. 2005.

BIOGRAPHIES

Lorenza Giupponi (lorenza.giupponi@cttc.es) received the Telecommunications Engineering degree at University of Rome ”La Sapienza” in July 2002 and the PhD at the Dept. of Signal Theory and Communications (TSC) of the Technical University of Catalonia (UPC) in 2007. She joined the Radio Communications Group of UPC in 2003 with a grant of the Spanish Ministry of Education. During 2006 and 2007 she was assistant professor at UPC. In September 2007 she joined the CTTC as Director of Institutional Relations and Research Associate in the area of Access Technologies.

Ana Maria Galindo (ana.maria.galindo@cttc.es) received her telecommunication and electronic engineering degree from Instituto Superior Politcnico Jose Antonio Echeverria (ISPJAE), La Habana, 2007, with a graduation thesis concerning to design WIMAX Networks. Currently, she is pursuing a Ph.D degree in Telecommunication Engineering at the Signal Theory and Communication department (TSC) of the Universitat Politecnica de Catalunya (UPC), Barcelona. She has been awarded with CTTCs PhD fellowship since November 2007 and is working in the Access Technologies Area.

Pol Blasco (pol.blasco@cttc.es) graduated in Telecommunications Engineering in 2009 in the Technical University of Catalonia (UPC, Barcelona). In 2006, in the framework of Erasmus students exchange program, he joined Technical University of Darmstadt (TUD, Germany) to complete his Msc. He graduated in Electrical and Information Theory Engineering in 2008 by the Technical University of Darmstadt (TUD, Germany). In 2008 he was awarded a Faro-Global scholarship by the Spanish Government to carry out the Msc Thesis in ESOC (European Space Operations Center, Darmstadt in Germany), were he performed research on autonomous receiver architecture, bitrate estimation techniques and modulation classification algorithms. From 2007 to 2008 he carried out undergraduate research in neuroscience in collaboration with both Technical University of Darmstadt (TUD, Germany) and the Max-Planck-Institut fr Hirnforschung (Frankfurt, Germany). In 2009 he worked in the Cortical Network and Events Lab from IDIBAPS (Institut d’Investigacions Biomdiques August Pi i Sunyer, Barcelona) performing research in neuroscience. In 2009 he joined the CTTC in order to pursue a PhD in the area of radio communications and started Msc MERIT studies in the Signal Theroy Department (TSC) of the Technical University of Catalonia (UPC, Barcelona). His current research interest cover the system and analysis neuroscience, spectrum estimates techniques, cognitive radio, satellite communications, image and signal processing.

Mischa Dohler (mischa.dohler@cttc.es) is now leading the Intelligent Energy [IQe] group at CTTC in Barcelona, with focus on Smart Grids and Green Radios. He is working on wireless sensor, machine-to-machine, femto, cooperative, cognitive and docitive networks. Prior to this, from June 2005 to February 2008, he has been Senior Research Expert in the R&D division of France Telecom. From September 2003 to June 2005, he has been lecturer at King’s College London, Centre for Telecommunications Research. At that time, he has also been London Technology Network Business Fellow for King’s College London, as well as Student Representative of the IEEE UKRI Section and member of the Student Activity Committee of IEEE Region 8 (Europe, Africa, Middle-East and Russia). He obtained his PhD in Telecommunications from King’s College London, UK, in 2003, his Diploma in Electrical Engineering from Dresden University of Technology, Germany, in 2000, and his MSc degree in Telecommunications from King’s College London, UK, in 1999. Prior to Telecommunications, he studied Physics in Moscow. He has won various competitions in Mathematics and Physics, and participated in the 3rd round of the International Physics Olympics for Germany. In the framework of the Mobile VCE, he has pioneered research on distributed cooperative space-time encoded communication systems, dating back to December 1999. He has published more than 120 technical journal and conference papers at a citation h-index of 21 and citation g-index of 43, holds several patents, co-edited and contributed to several books, has given numerous international short-courses, and participated in standardization activities. He has been TPC member and co-chair of various conferences, such as technical chair of IEEE PIMRC 2008 held in Cannes, France. He is and has been editor for numerous IEEE and non-IEEE journals and is Senior Member of the IEEE.