User to User QoE routing system

Hai Anh Tran and Abdelhamid Mellouk

Image, Signal and Intelligent Systems Lab-LiSSi Lab
University of Paris-Est Creteil Val de Marne (UPEC)
122 rue Paul Armangot 94400, Vitry sur Seine
{hai-anh.tran,mellouk}@u-pec.fr

Abstract. Recently, wealthy network services such as Internet protocol television (IPTV) and Voice over IP (VoIP) are expected to become more pervasive over the Next Generation Network (NGN). In order to serve this purpose, the quality of these services should be evaluated subjectively by users. This is referred to as the quality of experience (QoE). The most important tendency of actual network services is maintaining the best QoE with network functions such as admission control, resource management, routing, traffic control, etc. Among of them, we focus here on routing mechanism. We propose in this paper a protocol integrating QoE measurement in routing paradigm to construct an adaptive and evolutionary system. Our approach is based on Reinforcement Learning concept. More concretely, we have used a least squares reinforcement learning technique called Least Squares Policy Iteration. Experimental results showed a significant performance gain over traditional routing protocols.

Keywords: Quality of Service (QoS), Quality of Experience (QoE), Network Services, Inductive Routing System.

1 Introduction

The theory of quality of experience (QoE) has become commonly used to represent user perception. For evaluating network service, one has to measure, monitor, quantify and analyze the QoE in order to characterize, evaluate and manage services offered over this network. For users, also for operators and Internet service providers, the end-to-end quality is one of the major factors to be achieved. Actually the new term of QoE has been introduced to make end-to-end (e2e) QoS more clearly captures the experience of the users, which attempts to objectively measure the service delivered, QoE also takes in account the needs and the desires of the subscribers when using network services. Furthermore, the main challenge lies in how to take the dynamic changes in the resources of communication networks into account to provide e2e QoS for individual flows. Developing an efficient and accurate QoS mechanism is the focus of many challenges; mostly it is the complexity and stability. In reality, e2e QoS with more than two non correlated criteria is NP-complete [1]. Both technologies and needs continue to develop, so complexity and cost become limiting factors in the future evolution

of networks.
In order to solve a part of this problem, one has integrated QoE in network systems to adapt decision control in real-time. In fact, QoE does not replace QoS, but improves it. As an important factor of the end-to-end user perception, the QoE is an key metric for the design of systems and engineering processes. In addition, with QoE paradigm, we can reach a better solution to real-adaptive control network components.
Many people think that QoE is a part of QoS but it is a wrong perception. In reality, QoE covers the QoS concept. In order to demonstrate this theoretical point, we observe a quality chain of user-to-user services in Fig. 1. The quality perceived by end-user when using end-devices is called QoD (Quality of Design). The quality of network service including core network and access network is determined by QoS access and QoS backbone. QoE is satisfied only in the case that QoD and QoS are satisfied. So we can see that the user-to-user QoE is a conjunction of QoD and QoS.
In fact, there are many of network functions through which we can implement

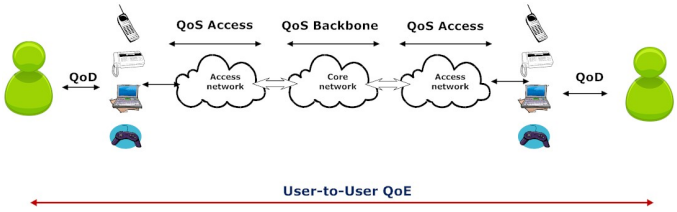

Fig. 1. Quality chain in an e2e service

QoE concept to improve system quality such as admission control, resource management, routing, traffic control, etc. Among of them, we focus on routing mechanism. Our purpose is to construct an adaptive routing method that can retrieve environment information and adapt to the environment changes. This adaptive routing mechanism maintains the required QoE of end-users. It is very necessary with network systems that have great dynamics (i.e. unreliable communication) and multiple user profiles where the required QoE levels are different. Most of the actual adaptive routing mechanisms that focus on optimizing simultaneously more than two non-correlated QoS metrics (i.e. e2e delay, jitter, response time, bandwidth) meet NP-complete problem we have mentioned above.
That is why in this paper we want to present a novel approach that integrates QoE measurement in a routing paradigm to construct an adaptive and evolutionary system.
Our approach is based on Reinforcement Learning (RL). More concretely, we based our protocol on Least Squares Policy Iteration (LSPI) [2], a RL technique that combines least squares function approximation with policy iteration.
The paper is structured as follows: Section II surveys briefly related works. In section III, some preliminaries of reinforcement learning for routing paradigm

are expressed. In section IV we focus on our approach based on reinforcement learning. The experimental results are shown in section V. Paper is ended with conclusion and some future works in section VI.

2 Related work

Routing mechanism is key to the success of large-scale, distributed communication and heterogeneous networks. The goal of every algorithm is to direct traffic from source to destinations maximizing network performance while minimizing costs. We cannot apply a static routing mechanism because network environment changes all time. So most existing routing techniques are designed to be capable to take into account the dynamics of the network. In other word, they are called adaptive routing protocol.

[3] presents a dynamic routing algorithm for best-effort traffic in Integrated-Services networks. Their purpose is to improve the throughput of high-bandwidth traffic in a network where flows exhibiting different QoS requirements are present. In this algorithm, one of two approaches is captured: either performing hop-based path computation, or evaluating the available bandwidth on a specific path. The problem is that the authors just focus on QoS requirements. Furthermore, this approach lacks learning process for an adaptive method.

The idea of applying RL to routing in networks was firstly introduced by Boyan and Littman [4]. Authors described the Q routing algorithm for packet routing. Reinforcement learning module is integrated into each node in network. Therefore, each router uses only local communication to maintain accurate statistics in order to minimize delivery times. However, this proposal is a simple reinforcement learning approach. It just focus on optimizing one basis network metric (delivery times).

AntNet, an adaptive, distributed, mobile-agents-based algorithm is introduced in [5]. In this algorithm, each artificial ant builds a path from its source to destination node. Collecting explicit information about the time length of the path components and implicit information about the load status of the network is realized in the same time of building the path. This information is then back-propagated by another ant moving in the opposite direction and is used to modify the routing tables of visited nodes.

In [6], authors proposed an application of gradient ascent algorithm for reinforcement learning to a complex domain of packet routing in network communication. This approach updates the local policies while avoiding the necessity for centralized control or global knowledge of the networks structure. The only global information required by the learning algorithm is the network utility expressed as a reward signal distributed once in an epoch and dependent on the average routing time.

In [7], KOQRA, a QoS based routing algorithm based on a multi-path routing approach combined with the Q-routing algorithm, is presented. The global learning algorithm finds K best paths in terms of cumulative link cost and optimizes the average delivery time on these paths. The technique used to estimate the

end-to-end delay is based on the reinforcement learning approach to take into account dynamic changes in networks.
We can see that all of these approaches above do not take into account the perception and satisfaction of end-users. In other word, they ignored QoE concept. In order to solve this lack, other proposals are presented:
In [8], authors propose to use an overlay network that is able to optimize the e2e QoE by routing around failures in the IP network. Network components are located both in the core and at the edge of the network. However, this system has no adaptive mechanism that can retrieve QoE feedback of end-users and adapt to the system actions.
[9] presents a new adaptive mechanism to maximize the overall video quality at the client. This proposal has taken into account the perception of end-users, in other words, the QoE. Overlay path selection is dynamically done based on available bandwidth estimation, while the QoE is subjectively measured using Pseudo-Subjective Quality Assessment (PSQA) tool. After receiving a client demand, the video server chooses an initial strategy and an initial scheme to start the video streaming. Then, client uses PSQA to evaluate the QoE of the received video in real time and send this feedback to server. After examining this feedback, the video server will decide to keep or to change its strategy. With this adaptive mechanism, all is done dynamically and in real time. This approach has well considered end-users perception. However this adaptive mechanism seems to be quite simple procedure. One should add learning process to construct much more user's profile.
In [10], authors proposed a routing scheme that adaptively learns an optimal routing strategy for sensor networks. This approach is based on LSPI, the same way on which we base our approach. However, in [10], the feedback information is gathered on each intermediate node along the routing path, that is not applicable for integrating the QoE feedback that we just obtain until data flux reach the destination. In our approach, we have changed LSPI technique such that we only require the feedback from the end node of the routing path.

3 Preliminaries

3.1 Reinforcement Learning model

In a reinforcement learning (RL) model [11] [12], an agent is connected to its environment via perception and action (Fig. 2). Whenever the agent receives as input the indication of current environment state, the agent then chooses an action, a, to generate as output. The action changes the state of the environment. The new value is communicated to the agent through a reinforcement signal, r. Formally, the model consists of:

- a discrete set of environment states: $\mathcal{S}$;
- a discrete set of agent actions: $\mathcal{A}$;
- a set of scalar reinforcement signals: $\mathcal{R}$;

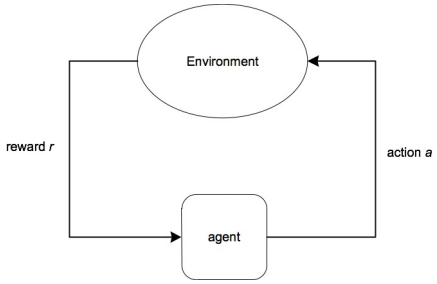

Fig. 2. Reinforcement Learning model

So the agent's goal is to find a policy π, mapping states to actions, that maximizes some long-run measure of reinforcement. In RL paradigm, one uses two types of value functions to estimate how good is for the agent to be in a given state:

- State-Value function $V^{\pi}(s)$: expected return when starting in s and following π thereafter.
$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} \tag{1}$$
- Action-Value function $Q^{\pi}(s, a)$: expected return starting from s, taking the action a, and thereafter following policy π
$$Q^{\pi}(s, a) = E_{\pi}\{R_t | s_t = s, a_t = a\} \tag{2}$$

where: t is any time step and R is the return.
Regarding policy iteration notion, the reason for computing the value function for a policy is to find better policies. Once a policy, π, has been improved using V^{π} to yield a better policy, π', we can then compute $V^{\pi'}$ and improve it again to yield an even better π''. We can thus obtain a sequence of monotonically improving policies and value functions:

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} ... \xrightarrow{I} \pi^* \xrightarrow{E} V^* \tag{3}$$

where E denote a policy evaluation and I denotes a policy improvement.
The goal is to find an optimal policy that maps states to actions such that the cumulative reward is maximized, which is equivalent to determining the following Q function (Bellman equation):

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) max_{a'} Q^*(s', a') \tag{4}$$

In our approach, we don't apply directly this basic Q function. In fact, this basic Q-Learning value function requires a large number of tries before being able to converge to the optimal solution, furthermore it is sensitive to the parameter setting. So we must approximate it using a technique called Function Approximation that is presented in the next section.

3.2 Function Approximation

The table representing the estimate of value functions with one entry for each state or for each state-action pair is limited to tasks with small numbers of states and actions. Time and data needed to accurately fill them is a problem too. In routing system to which we applied reinforcement learning, most states encountered will never have been experienced exactly before. Generalizing from previously experienced states to ones that have never been seen is the only way to learn anything at all on these tasks.
We require a generalization method called *function approximation* [12] because it takes examples from a desired function (e.g., a value function) and attempts to generalize from them to construct an approximation of the entire function. Function approximation is a combination between reinforcement learning methods and existing generalization methods. One of the most important special cases in function approximation is *Linear method* that is the method we have applied to our approach.

4 Proposed approach

Our idea to take into account end-to-end QoE consists to develop adaptive mechanisms that can retrieve the information from their environment (QoE) and adapt to initiate actions. These actions should be executed in response to unforeseen or unwanted event as an unsatisfactory QoS, a negative feedback or malfunctioning network elements.
The purpose of this approach consists of:

- Integrate parameters from the terminal application (or user) in terms of QoE metrics in the decision loop system.
- Taking into account the feedback of QoS and QoE metrics as input for autonomy in the adaptive model.
- Developing a viable model of a knowledge plane enabling scaling on equipment located in different parts of the network.
- Taking into account a propagation model of dysfunction and stable algorithms.

Concretely, the system integrates the QoE measurement in an evolutionary routing system in order to improve the user perception based on the choice of the best optimal QoE paths (Fig. 3). So in that way, the routing process is build according to the improvement of QoE parameters. The system is based on adaptive algorithms that simultaneously process an observation and learn to perform better.
In this section, we describe in detail our proposal in order to integrate QoE measurement in an adaptive routing system. We first present our mapping of reinforcement learning model to our routing model. Then we describe Least-Squares Policy Iteration technique on which we based our approach. Finally the learning procedure is depicted.

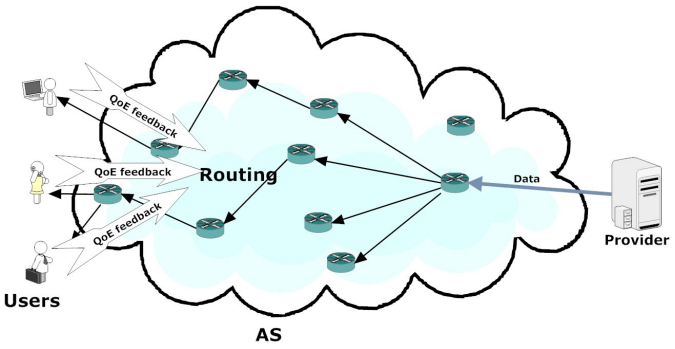

Fig. 3. Integration of QoE measurement in routing system

4.1 Reinforcement learning based routing system including QoE measurement

In order to integrate reinforcement learning notion into our routing system, we have mapped reinforcement learning model to our routing model in the context of learning routing strategy (Fig. 4). We consider each router in the system as a state. The states are arranged along the routing path. Furthermore, we consider each link emerging from a router as an action to choose. The system routing mechanism corresponds to the policy π.

After data reach end-users, QoE evaluation is realized to give a QoE feedback to

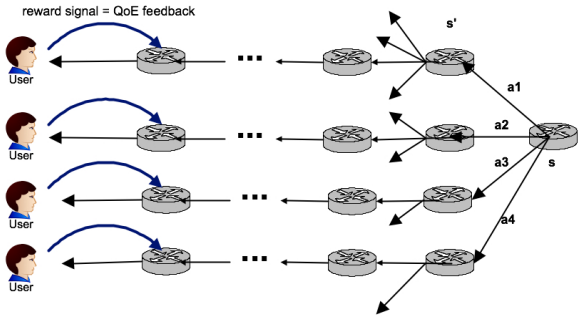

Fig. 4. Routing system based on reinforcement learning

the system. We consider this feedback as environment reward and our purpose is to improve the policy π using this QoE feedback. Concretely, the policy π is chosen so that it is equal to $argmax$ of action value function Q in policy π:

$$\pi_{t+1}(s_t) = argmax Q^{\pi_t}(s_t, a) \tag{5}$$

4.2 Least-Squares Policy Iteration

Least-Squares Policy Iteration (LSPI) [2] is a recently introduced reinforcement learning method. Our choice is based on the fact that this technique learns the weights of the linear functions, thus can update the Q-values based on the most updated information regarding the features. It does not need carefully tuning initial parameters (e.g., learning rate). Furthermore, LSPI converges faster with less samples than basic Q-learning. In this technique, instead to calculate directly action-value function Q (Equation 4), this latter is approximated with a parametric function approximation. In other words, the value function is approximated as a linear weighted combination:

$$\hat{Q}^{\pi}(s,a,\omega) = \sum_{i=1}^{k} \phi_i(s,a)\omega_i = \phi(s,a)^T\omega \tag{6}$$

where $\phi(s,a)$ is the basis features vector and ω is weight vector in the linear equation. The k basis functions represent characteristics of each state-action pair.
We have to update the weight vector ω to improve system policy.
Equation 4 and 6 can be transformed to $\Phi\omega \approx R + \gamma P^{\pi}\Phi\omega$, where Φ represent the basis features for all state-action pairs. This equation is reformulated as:

$$\Phi^T(\Phi - \gamma P^{\pi}\Phi)\omega^{\pi} = \Phi^T R \tag{7}$$

Basing on equation 7, the weight ω of the linear functions in equation 6 is extracted:

$$\omega = (\Phi^T(\Phi - \gamma P^{\pi}\Phi))^{-1} \times \Phi^T R \tag{8}$$

4.3 Learning procedure

For a router s and forwarding action a, s' is the corresponding neighbor router with $P(s'|s,a) = 1$. Our learning procedure is realized as follows: when a packet is forwarded from node s to s' by action a, which has been chosen by current Q-values $\phi(s,a)^T\omega$, a record $< s,a,s',\phi(s,a) >$ is inserted to the packet. The gathering process (Fig. 5) is realized until the packet arrives at the destination (end-users). Thus with this way, one can trace the information of the whole

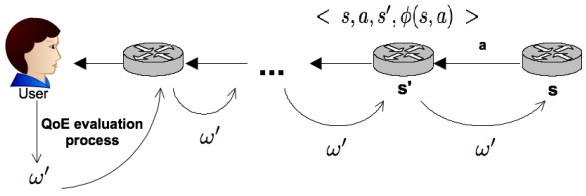

Fig. 5. Learning procedure

routing path. At the end-users, a QoE evaluation process is realized to give a QoE score that is the value of R vector in equation 8. Furthermore, with the gathered information, the new value of ω is determined using equation 8. Then this new weight value ω' is sent back to the system along the routing path in order to improve policy procedure in each router on the routing path. With the new weights ω', policy improvement is realized in each router on the routing path by selecting the action a with the highest Q-value:

$$\pi(s|\omega') = argmax_a \phi(s,a)^T \omega' \tag{9}$$

The next section presents some experiment results in comparing our approach with other routing protocols.

5 Simulation results

We have used Opnet simulator version 14.0 to implement our approach in an autonomous system. Regarding network topology, we have implemented an irregular network with 3 separated areas including 38 routers, each area is connected to each other by one link, all links are similar (Fig. 6). For QoE evaluation

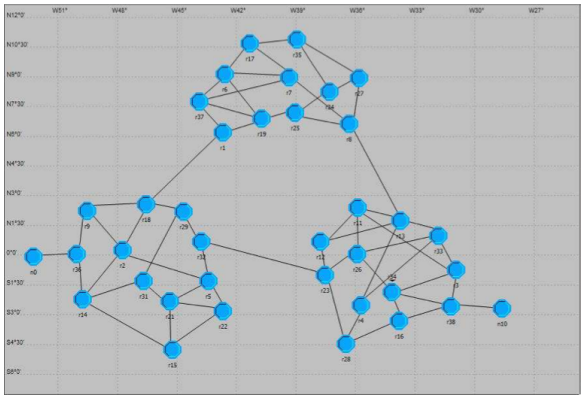

Fig. 6. Simulated network

method, we have used a reference table represents human estimated MOS [1] function [13] of loss packet rate based on real experiment.
To validate our results, we compare our approach with two kinds of algorithm:

[1] Mean Opinion Score (MOS) gives a numerical indication of the perceived quality of the media received after being transmitted. MOS is expressed in one number, from 1 to 5, 1 being the worst and 5 the best. The MOS is generated by averaging the results of a set of standard, subjective tests where a number of users rate the service quality

– Those based on Shortest Path First (SPF) technique where routing in this family used the only best path based on delay constraint.
– The Routing Information Protocol (RIP), a dynamic routing protocol used in local and wide area networks. It uses the distance-vector routing algorithm. It employs the hop count as a routing metric. RIP prevents routing loops by implementing a limit on the number of hops allowed in a path from the source to a destination.

In order to have QoE evaluation results, we have implemented a measurement model that is depicted in Fig. 7. The video server provides video streaming to the

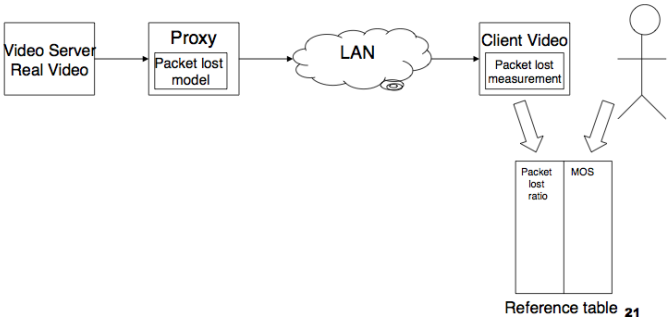

Fig. 7. QoE evaluation model

client through a proxy into which we have integrated a packet lost model. In the destination, in order to construct the reference table, the Client Video realizes packet lost measurement and the end-users give a MOS score. Thus we have an equivalent MOS score for each packet lost ratio. Our simulation results are based on this reference table to give MOS scores equivalent to packet lost ratio. Fig. 8 illustrates the result of average MOS score of three protocols: SPF, RIP and our approach. With the three protocols we have implemented in the simulation, we can notice that each is based on a different criteria. The SPF focuses on minimizing the routing path. On the other side, RIP focuses on the hop-count of routing packages. As our protocol, we rely on user perception. As shown in Fig. 8, we can see that our approach gives better results than two other algorithms in the same delay. So with our approach, despite network environment changes, we can maintain a better QoE without any other e2e delay.

6 Conclusion

In this paper, we present a novel approach for routing systems in considering the user-to-user QoE, which has integrated QoE measurement to routing paradigm for an adaptive and evolutionary system. The approach is based on

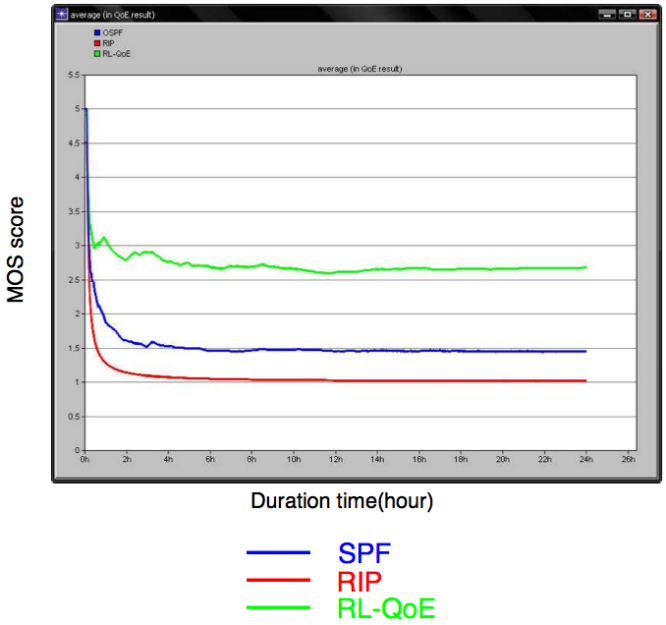

Fig. 8. Simulation results

a least squares reinforcement learning technique. The simulation results we obtained with OPNET simulations demonstrates that our proposed approach yields significant QoE evaluation improvements over traditional approaches using the same e2e delay.
Some future works includes applying our protocol to real testbed to verify its feasibility, using standard QoE evaluation methods and implementing protocol on IPTV flows.

References

1. Z. Wang and J. Crowcroft, "Quality of service routing for supporting multimedia applications," *IEEE Journal on selected areas in communications*, vol. 14, no. 7, pp. 1228–1234, 1996.
2. M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *Journal of Machine Learning Research*, vol. 4, p. 1149, 2003.
3. C. Casetti, G. Favalessa, M. Mellia, and M. Munafo, "An adaptive routing algorithm for best-effort traffic in integrated-services networks," *Teletraffic science and engineering*, pp. 1281–1290, 1999.
4. J. A. Boyan and M. L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," *Advances in Neural Information Processing Systems*, p. 671, 1994.
5. G. D. Caro and M. Dorigo, "Antnet: A mobile agents approach to adaptive routing," *The Hawaii International Conference On System Sciences*, vol. 31, pp. 74–85, 1998.

6. L. Peshkin and V. Savova, "Reinforcement learning for adaptive routing," *Intnl Joint Conf on Neural Networks*, 2002.
7. A. Mellouk, S. Hoceini, and S. Zeadally, "Design and performance analysis of an inductive qos routing algorithm," *Computer Communications*, vol. 32, no. 1371-1376, 2009.
8. B. D. Vleeschauwer, F. D. Turck, B. Dhoedt, P. Demeester, M. Wijnants, and W. Lamotte, "End-to-end qoe optimization through overlay network deployment," *International Conference on Information Networking*, 2008.
9. G. Majd, V. Cesar, and K. Adlen, "An adaptive mechanism for multipath video streaming over video distribution network (vdn)," *First International Conference on Advances in Multimedia*, 2009.
10. P. Wang and T. Wang, "Adaptive routing for sensor networks using reinforcement learning," *IEEE International Conference on Computer and Information Technology*, 2006.
11. L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of AI Research*, vol. 4, pp. 237–285, 1996.
12. R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE transactions on neural networks*, vol. 9, 1998.
13. I.-T. R. P.801, "Mean opinion score (mos) terminology," *International Telecommunication Union, Geneva*, 2006.