Q-Learning Algorithm for VoLTE Closed Loop Power Control in Indoor Small Cells

Faris B. Mismar and Brian L. Evans
Wireless Networking and Communications Group, The University of Texas at Austin, Austin, TX 78712 USA

***Abstract*—We propose a reinforcement learning (RL) based closed loop power control algorithm for the downlink of the voice over LTE (VoLTE) radio bearer for an indoor environment served by small cells. The main contributions of our paper are to 1) use RL to solve performance tuning problems in an indoor cellular network for voice bearers and 2) show that our derived lower bound loss in effective signal to interference plus noise ratio due to neighboring cell failure is sufficient for VoLTE power control purposes in practical cellular networks. In our simulation, the proposed RL-based power control algorithm significantly improves both voice retainability and mean opinion score compared to current industry standards. The improvement is due to maintaining an effective downlink signal to interference plus noise ratio against adverse network operational issues and faults.**

***Index Terms*—reinforcement learning, artificial intelligence, VoLTE, MOS, QoE, optimization, SON.**

I. Introduction

Wireless networks are vulnerable to operational issues, faults and network element failures. In addition, wireless transmission can face blockage, interference, and other impairments. While cellular data applications are made resilient against wireless impairments through modulation, coding, and retransmissions, delay-sensitive applications such as voice or low latency data transfer may not always benefit from retransmission since it increases delays and risk of data duplication. Therefore, these applications must be made resilient through other means.

The received *signal to interference plus noise ratio* (SINR) is a critical quantity to ensure resilient communications in applications such voice. In this paper, we propose a framework to automatically tune a cellular network through the employment of reinforcement learning (RL). We devise an RL-based algorithm to improve downlink SINR in an indoor environment for packetized voice using *power control* (PC) as shown in Fig. 1. The technology of focus is the fourth generation of wireless communications or *long term evolution* (4G LTE) or fifth generation of wireless communications (5G).

Downlink closed loop power control is last implemented in 3G *universal mobile telecommunications system* (UMTS) [1]. It rapidly adjusts the transmit power of a radio link of a dedicated traffic channel to match the target DL SINR. This technique is not present in 4G LTE or 5G due to the absence of dedicated traffic channels for packet data sessions. However, the introduction of *semi-persistent scheduling* (SPS) in 4G LTE has created a virtual sense of a dedicated downlink traffic channel for VoLTE on which a closed loop power control can be performed. This scheduling is at least for the length of one voice frame—which is in order of tens of LTE *transmit time intervals* (TTIs).

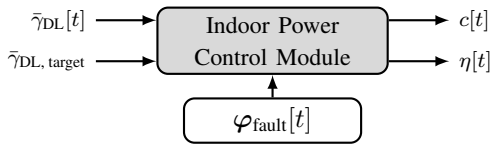

Fig. 1. Downlink power control module. $\bar{\gamma}_{\text{DL}}[t]$ is the effective signal to noise plus interference ratio (SINR) at the receiver at time t fed back to the power control module. The module maintains the downlink SINR at the receiver at $\bar{\gamma}_{\text{DL, target}}$ amid the faults captured in $\boldsymbol{\varphi}_{\text{fault}}[t]$. This is done through a power control command $c[t]$ and a repetition factor $\eta[t]$.

An improved decentralized Q-learning algorithm was used to reduce interference in the LTE femtocells environment in [2]. Various power control algorithms including open loop power control were compared with this Q-learning based approach. Combining information theory with machine learning, a deep reinforcement learning method was used in [3]. This method was a constrained optimization problem to maximize the Q-function using the Kullback-Leibler divergence and entropy constraints instead of exploration, which we used in this paper. Closed loop power control implementation for LTE with the employment of fractional path loss compensation was done in [4]. This resulted in an improved system performance. However, there was no reference to machine learning or RL in general. Q-learning based power control for indoor LTE femtocells with an outdoor macro cell was studied in [5]. The focus was on throughput and the UE reported SINR or call quality indicator was used to change the femtocell transmit power. To resolve the issue of communicating base stations, a central controller was introduced. We confine the geographical area to make it is feasible for the base stations to communicate through the backhaul. Furthermore, an assumption that downlink power control is achieved over shared data channels, which is a relaxed assumption and requires that the scheduler is aware ahead of time about the channel condition for the upcoming user to perform power control, was made. We did not make this assumption to maintain a more realistic environment since we exploit the use of SPS in packetized voice.

While the focus of our paper is on *voice over LTE*

(VoLTE), any future technology offering packetized voice can benefit from our proposed algorithm. In fact, with the highly anticipated role of *self-organizing networks* (SON) in 5G [6], similar algorithms can be a significant step towards machine learning enabled SON.

Our main contributions are as follows:

1) Use RL to solve performance tuning problems in an indoor cellular network for voice bearers.
2) Show that our derived lower bound loss in effective SINR due to neighboring cell failure is sufficient for VoLTE power control in practical cellular networks.

II. System Model

The system comprises two components:

1) A radio environment where VoLTE capable UEs are served.
2) A reinforcement learning model using Q-learning to perform closed loop power control to improve effective DL SINR measured at the receiver.

A. Radio Environment

The radio environment is an *orthogonal frequency-division multiplexing* (OFDM) indoor cellular cluster using frequency division duplex and multiple access with one tier of neighboring small cells each with a square geometry length L as shown in Fig. 2. The UEs are scattered in $\mathbb{R}^2$ according to a *homogeneous poisson point process* (PPP) [7]. This process Φ has an *intensity parameter* λ which represents the expected number of users served by the small cell per unit area. We define the point process Φ by the number of stationary users N in the service area of the small cell W sampled from a Poisson distribution with mean $\lambda W = \lambda L^2$. The i-th UE coordinates are sampled from an *independent and identically distributed* (i.i.d.) continuous uniform distribution. There are N_{UE} UEs per cell. We make them stationary to increase our channel coherence time for reinforcement learning purposes.

We choose the square geometry because in an indoor environment with a typical omnidirectional cell installed at the center of square-shaped floor plans, a square tessellation is possible. This is because as the transmitted signals are attenuated further more towards the square forming walls. In practice, square geometry has been used in indoor commercial deployments [8], and is therefore our choice.

This cellular cluster can be in a normal state or undergo some fault-generated actions. These faults $\mathcal{N}$ cause the channel impairment and are tracked in a special register. We show these faults in Table I.

We start by writing the signal model in an additive white Gaussian noise channel for our indoor system

$$y_i[t] = h_i[t]s_i[t] + n[t], \qquad i = 1, 2, \dots N_{\text{UE}} \tag{1}$$

where $y_i[t]$ is the received signal for the i-th UE, $h_i[t]$ is a single-tap channel, and $n[t]$ is a Gaussian random process sampled from $\text{Norm}(0, \sigma_n^2)$.

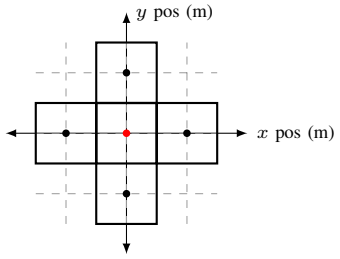

Fig. 2. Radio environment. The red point represents the serving cell. The black points in the adjacent squares are the low power nodes.

TABLE I
Network Actions $\mathcal{N}$

Action ν	Definition	Rate
0	Cluster is normal.	p_0
1	Feeder fault alarm (3 dB loss of signal).	p_1
2	Neighboring cell down.	p_2
3	VSWR out of range alarm.	p_3
4	Feeder fault alarm cleared.†	p_4
5	Neighboring cell up again.†	p_5
6	VSWR back in range.†	p_6

† These actions cannot happen if their respective alarm did not happen first.

Now, we compute the received downlink SINR for the i-th UE at TTI t ($\gamma_{\text{DL},i}[t]$) for $i = 1, 2, \dots, N_{\text{UE}}$ as:

$$\gamma_{\text{DL},i} \triangleq \frac{\mathbb{E}[|y_i|^2]}{\sigma_n^2 + \underbrace{\sum_{j:\mathbf{o}_j \in \mathcal{C}\setminus\{\mathbf{o}_0\}} k_j \mathbb{E}[|y_j|^2]}_{\text{ICI}}}. \tag{2}$$

where $\mathcal{C}$ is a set of all the cells in the cluster and $\mathbf{o}_j$ is the coordinates of the j-th base station. Without loss of generality, we assume that $\mathbf{o}_0$ is the serving cell placed at the origin, $k_j \geq 0$ is the proportion of users from the adjacent cells j whose signals are transmitted on the same PRB as the i-th UE at TTI t. Those signals therefore cause *inter-cell interference* (ICI).

The forward link budget at any TTI t is written as:

$$P_{\text{UE},i}[t] = P_{\text{TX}}[t] + G_{\text{TX}} - L_{\text{m}} - L_{\text{a},i}[t] + G_{\text{UE}} \tag{3}$$

where $P_{\text{UE},i}$ is the received power for the allocated *physical resource blocks* (PRB) transmitted at power P_{TX}, G_{TX} is the antenna gain of the transmitter, L_{m} is a miscellaneous loss (e.g., feeder loss and return loss), $L_{\text{a},i}$ is the path loss over the air interface for line of sight indoor propagation for the i-th user, and G_{UE} is the UE antenna gain.

$$\gamma_{\text{DL},i}[t] \triangleq \frac{P_{\text{UE},i}[t]}{\sigma_n^2 + \sum_{j:\mathbf{o}_j \in \mathcal{C}\setminus\{\mathbf{o}_0\}} k_j P_{\text{UE},j}[t]}. \tag{4}$$

The effective downlink received SINR for users in the serving cell $\mathbf{o}_0$ at a given TTI t, $\bar{\gamma}_{\text{DL}}[t]$ in dB is

$$\bar{\gamma}_{\text{DL}}[t] \triangleq 10 \log \left(\frac{1}{N_{\text{UE}}} \sum_{i=1}^{N_{\text{UE}}} \gamma_{\text{DL},i}[t] \right) \qquad \text{(dB)} \tag{5}$$

which is the quantity to improve (i.e., our *objective*).

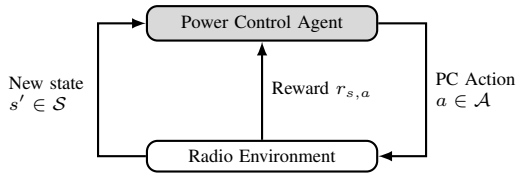

Fig. 3. Reinforcement learning elements.

B. Reinforcement Learning

We formulate the VoLTE closed loop PC problem as a reinforcement learning based problem using Tabular Q-learning as shown in Algorithm 1. The objective is to meet the target effective SINR $\bar{\gamma}_{\text{DL, target}}$ despite signal impairments, which are tracked in a register $\boldsymbol{\varphi}_{\text{fault}}[t]$ where $\boldsymbol{\varphi}_{\text{fault}} \in \mathbb{F}_2^{|\mathcal{N}|}$. The set of actions carried out by the agent is $\mathcal{A} = \{a_i\}_{i=0}^{n-1}$ and the set of states is $\mathcal{S} = \{s_i\}_{i=0}^{m-1}$. The environment grants the agent a reward $r_{s,a}$ after the algorithm takes an action $a \in \mathcal{A}$ when it is in state $s \in \mathcal{S}$ at discrete time t. These elements are shown in Fig. 3.

Following [9], we denote $Q(s,a)$ as the state-action value function, i.e., the expected discounted reward when starting in state s and selecting an action a. To derive $Q(s,a)$, we build an m-by-n table $\mathbf{Q} \in \mathbb{R}^{m \times n}$. This allows us to use the shorthand notation: $Q(s,a) \triangleq [\mathbf{Q}]_{s,a}$ and write:

$$Q(s,a) \triangleq (1-\alpha)Q(s,a) + \alpha \left[r_{s,a} + \gamma \max_{a'} Q(s',a') \right] \quad (6)$$

where $\alpha : 0 < \alpha < 1$ is the learning rate and $\gamma : 0 < \gamma < 1$ is the *discount factor* and determines the importance of the predicted future rewards. The next state is s' and the next action is a'. For our proposed closed loop PC algorithm, the the upper bound of the time complexity for Q-learning is in $\mathcal{O}(m^2)$ [10].

An *episode* is a period of time in which an interaction between the agent and the environment takes place. In our case, this period of time is τ TTIs. During an episode z : $z \in \{0, 1, \ldots, \zeta\}$, the agent makes the decision to maximize the effects of actions decided by the agent. To do so, we use the ϵ-greedy strategy for learning, where ϵ is the *exploration rate* and serves to select a random action $a \in \mathcal{A}$ with a probability $\epsilon : 0 < \epsilon < 1$, known as exploration, as opposed to selecting an action based on previous experience, which is also known as exploitation. The exploration rate decays in every episode until it reaches $\epsilon_{\min}$.

We list the actions performed by the PC agent (known as *PC commands*) in Table II. The PC agent attempts to reach $\bar{\gamma}_{\text{DL, target}}$ through a series of PC commands c in reaction to various impairments due to the network actions $\nu \in \mathcal{N}$ during any given TTI t. These actions have a finite impact on the effective downlink received SINR $\bar{\gamma}_{\text{DL}}[t]$.

After each action, we compute the effective downlink SINR gain (or loss), which is $\Delta\bar{\gamma}_{\text{DL}}[t]$, as

$$\Delta\bar{\gamma}_{\text{DL}}[t] \triangleq \bar{\gamma}_{\text{DL}}[t] - \bar{\gamma}_{\text{DL}}[t-1]. \quad (7)$$

Computation of contribution of actions $\nu = 1, 3$. When the *voltage standing wave ratio* (VSWR) changes from v_0 to v in TTI t (action $\nu = 3$), we compute the change in signal loss due to return loss, which is equal to the change in SINR, as

$$\Delta L = 10 \log \left(\left| \frac{v_0 + 1}{v_0 - 1} \right| \cdot \left| \frac{v - 1}{v + 1} \right| \right)^2. \quad (8)$$

Action $\nu = 1$ is a special case with $\Delta L = -3\,\text{dB}$.

Computation of contribution of action $\nu = 2$. When the neighbor cell k is down, the transmit power of the adjacent cells j will increase by an arbitrary quantity $0 < \varepsilon_j \leq 1$ and the number of interferers will decrease. Therefore, for simplicity of computation, we derive the lower bound of the downlink effective SINR of this action $\nu = 2$ as

$$\begin{aligned} \bar{\gamma}_{\text{DL}} &= \frac{P_{\text{UE},i}}{\sigma_n^2 + \sum_{j:j \in \mathcal{C} \setminus \{\mathbf{o}_0, \mathbf{o}_\ell\}} (1+\varepsilon_j) k_j P_{\text{UE},j}} \\ &\overset{(a)}{\geq} \frac{P_{\text{UE},i}}{\sigma_n^2 + |\mathcal{C} \setminus \{\mathbf{o}_0, \mathbf{o}_\ell\}| P_{\text{BS}}^{\max}} \\ &\overset{(b)}{=} \frac{P_{\text{UE},i}}{\sigma_n^2 + (|\mathcal{C}| - 2) P_{\text{BS}}^{\max}} \end{aligned} \quad (9)$$

where $P_{\text{BS}}^{\max}$ is the maximum transmit power of the indoor cell. (a) is since we use the maximum small cell transmit powers instead of the increased received power measured at the UE, and (b) is since the cardinality of $\mathcal{C}$ is reduced by two: the serving and neighbor cells from step (a).

Computation of contribution of actions $\nu = 4, 5, 6$. These actions are a result of their respective fault actions being cleared. Therefore, we reverse the effect of actions $1, 2$, and 3 respectively.

TABLE II
POWER CONTROL (PC) ALGORITHM ACTIONS AT TIME t

Action a	Definition
0	$c = 0$.
1	$c = -1$ repeated three times (i.e., $\eta[t] = 3$).
2	$c = -1$ repeated one time (i.e., $\eta[t] = 1$).
3	$c = +1$ repeated one time.
4	$c = +1$ repeated three times.

TABLE III
POWER CONTROL (PC) ALGORITHM STATES

State s	Definition
0	Transmit power is unchanged (i.e., $c = 0$).
1	Transmit power is increased (i.e., $c = +1$).
2	Transmit power is decreased (i.e., $c = -1$).

III. POWER CONTROL ALGORITHMS

A. Fixed Power Allocation

The *fixed power allocation* (FPA) power control method is an open-loop PC algorithm that serves as a baseline for comparison. It is a common power allocation scheme where the total transmit power is simply divided equally among all PRBs in the operating band N_{PRB} and is therefore constant but cannot exceed the maximum transmission power of the small cell:

$$P_{\text{TX}}[t] \triangleq P_{\text{BS}}^{\max} - 10 \log N_{\text{PRB}} \quad \text{(dBm)} \quad (10)$$

Algorithm 1: VoLTE Downlink Closed Loop Power Control

Input: Initially computed effective downlink SINR value ($\bar{\gamma}_{\text{DL},0}$) and desired target effective SINR value ($\bar{\gamma}_{\text{DL, target}}$).

Output: Optimal sequence of power commands required to achieve the target SINR value during a VoLTE frame z, which has a duration of τ amid network impairments captured in $\boldsymbol{\varphi}_{\text{fault}}$.

Define the power control (PC) actions $\mathcal{A}$, the set of PC states $\mathcal{S}$, the exploration rate ϵ, the decay rate d, and $\epsilon_{\min}$.
$t := 0$
$\bar{\gamma}_{\text{DL}} := \bar{\gamma}_{\text{DL},0}$
$(s, a) := (0, 0)$
$\boldsymbol{\varphi}_{\text{fault}} := [0, 0, \ldots, 0]$
repeat
  $t := t + 1$
  $\epsilon := \max(\epsilon \cdot d, \epsilon_{\min})$
  Sample $r \sim \text{Uniform}(0, 1)$
  if $r \leq \epsilon$ **then**
    Select an action $a \in \mathcal{A}$ at random.
  else
    Select an action $a \in \mathcal{A}, a = \arg\max_{a'} Q(s, a')$.
  end
  Perform action a (power control) on $P_{\text{TX}}[t]$ and obtain reward $r_{s,a}$.
  Observe next state s'.
  Update the table entry $Q(s, a)$ as in (6).
  $s := s'$
until $\bar{\gamma}_{DL} \geq \bar{\gamma}_{DL,\ target}$ or $t \geq \tau$
Proceed to the next VoLTE frame $z + 1$.

B. Closed Loop Power Control

Owed to the closed loop PC, we can write P_{TX} in dBm at any given TTI t as:

$$P_{\text{TX}}[t] = \min\big(P_{\text{BS}}^{\max}, P_{\text{TX}}[t-1] + \eta[t]c[t]\big) \qquad \text{(dBm)}$$

where $\eta[t]$ is the repetition factor of a power command c in a given TTI t. Power control cannot cause the transmit power to exceed the maximum transmit power of the serving cell. Power commands can be issued multiple times per TTI as shown in Table II. These power commands impact the algorithm state as shown in Table III.

IV. Voice Call Performance Metrics

We use two performance metrics: *call retainability* and *mean opinion score* (MOS) to compare both algorithms.

A. Call Retainability

We define call retainability for the radio environment as a function of an effective downlink SINR threshold $\bar{\gamma}_{\text{DL, min}}$ obtained during the final episode ζ:

$$\text{Retainability} \triangleq 1 - \frac{1}{\tau}\sum_{t=0}^{\tau} \mathbb{1}_{\bar{\gamma}[t] \leq \bar{\gamma}_{\text{DL, min}}}. \qquad (11)$$

B. Mean-Opinion Score

To benchmark the audio quality, we compute *mean-opinion score* (MOS) using the experimental MOS formula [11]. We obtain the packet error rate from the simulation over τ frames in the final episode ζ using the symbol probability of error of a QPSK modulation in OFDM. We choose a VoLTE data rate of 23.85 kbps and a voice *activity factor* (AF), the ratio of voice payload to silence during a voice frame, of 0.7. We refer to the source code [12] for further details. Result is in Fig. 6.

V. Simulation Results

We implement Algorithm 1 and set the machine learning parameters as in Table V. To examine the worst case scenario, we set the probabilities in Table I as: $p_0 = 5/11, p_1 = p_2 = p_3 = \ldots = p_6 = 1/11$. This way we give all faults an equally likely chance of occurrence while having the network perform reliably at least for 45% of the time. We further set the rewards as:

$$r_{s,a}[t] \triangleq \begin{cases} r_{\min}, & \bar{\gamma}_{\text{DL}}[t] = \bar{\gamma}_{\text{DL, target}} \text{ not feasible or } t \ll \tau \\ -1, & \bar{\gamma}_{\text{DL}}[t] < \bar{\gamma}_{\text{DL}}[t-1] \\ 0, & \bar{\gamma}_{\text{DL}}[t] = \bar{\gamma}_{\text{DL}}[t-1] \\ 1, & \bar{\gamma}_{\text{DL}}[t] > \bar{\gamma}_{\text{DL}}[t-1] \\ r_{\max}, & \bar{\gamma}_{\text{DL}}[t] = \bar{\gamma}_{\text{DL, target}} \text{ is met.} \end{cases} \qquad (12)$$

For the retainability, we choose $\bar{\gamma}_{\text{DL, min}} = 0\,\text{dB}$ in (11). The radio network parameters are set as in Table IV. We set $\bar{\gamma}_{\text{DL},0}$ to 4 dB and $\bar{\gamma}_{\text{DL, target}}$ to 6 dB.

The optimal Q-learning action-value function (6) is learned after ζ episodes. At this stage, the closed loop PC performs better than FPA. Fig. 4 shows the power command sequence for the final episode $z = \zeta = 707$, where the closed loop PC algorithm causes the base station to change its transmit power to meet the desired DL SINR target, unlike FPA where no power commands are sent. We show both algorithms in Fig. 5 for the final episode ζ. The closed loop PC pushes the effective DL SINR to the target through an optimal sequence of power commands. The improved retainability and experimental MOS scores due to the closed loop power control algorithm are shown in Table VI and Fig. 6 respectively. The reason why retainability and MOS have improved is understood directly from the impact of the effective DL SINR which increases the quantity of (11) and decreases the packet error rate—the main component in the experimental MOS formula. We refer to the source code [12] for further implementation details.

VI. Conclusion

We introduced downlink closed loop power control using Q-learning, which improved VoLTE performance in a realistic indoor environment compared to the open-loop fixed power allocation power control. It resulted in improvement in the quality of experience measured by the voice call retainability and MOS metrics. This was due to the robustness of maintaining the target DL SINR. The ability to maintain this target helps prevent a voice call from dropping and reduces the voice packet errors.

TABLE IV
RADIO ENVIRONMENT PARAMETERS

Parameter	Value	Parameter	Value
LTE bandwidth	20 MHz	Base station maximum power $P_{\text{LPN}}^{\max}$	33 dBm
Downlink center frequency	2.6 GHz	Base station initial power setting	13 dBm
Maximum number of UEs per serving cell N_{UE}	10	Antenna model	omnidirectional
Number of physical resource blocks N_{PRB}	100	Antenna gain G_{TX}	16 dBi
Cellular geometry	square (L = 10 m)	Antenna height	10 m
Propagation model	COST 231	User Equipment (UE) antenna gain	-1 dBi
Propagation environment	indoor	UE height	1.5 m

TABLE V
MACHINE LEARNING PARAMETERS

Parameter	Value
Number of episodes ζ	707
One episode duration τ (ms)	20
Discount factor γ	0.950
Exploration rate ϵ	1.000
Minimum exploration rate ϵ_{min}	0.010
Exploration rate decay d	0.99
Learning rate α	0.001
Number of states	3
Number of actions	5

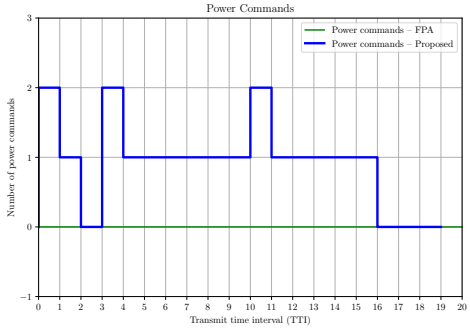

Fig. 4. Power control (PC) sequence during the final episode ζ. Unlike fixed power allocation (FPA), our proposed closed loop power control sent several PCs per transmit time interval (TTI) for the entire VoLTE frame.

TABLE VI
RETAINABILITY

	Fixed Power Allocation	Proposed
Retainability	55.00%	**78.75%**

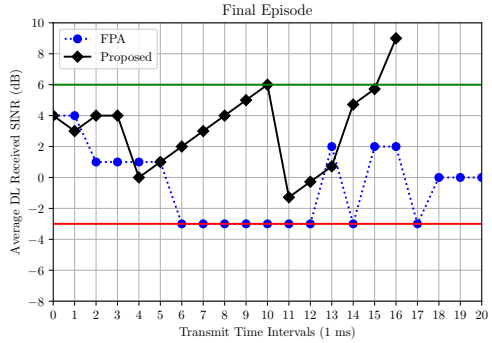

Fig. 5. Downlink signal to interference plus noise ratio (DL SINR) improvement vs. simulation time for both our proposed closed loop power control using Q-learning and fixed power allocation (FPA) for the final episode ζ. Green and red lines are $\bar{\gamma}_{\text{DL, target}}$ and $\bar{\gamma}_{\text{DL, min}}$ respectively. The proposed algorithm reaches the target while FPA does not.

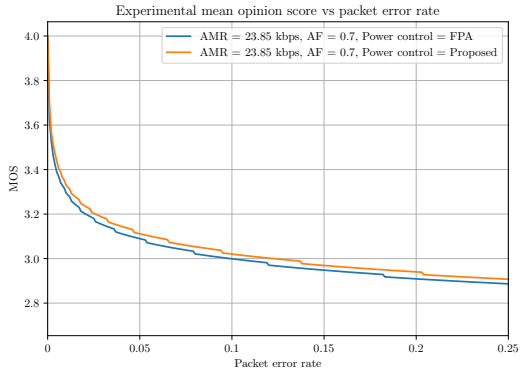

Fig. 6. Mean opinion score (MOS) based on the voice packet error rate and the experimental formula [11]. Our proposed closed loop power control algorithm has improved MOS compared to Fixed Power Allocation (FPA).

REFERENCES

[1] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3rd Generation Partnership Project (3GPP), TS 25.214, Dec. 2015. [Online]. Available: http://www.3gpp.org/dynareport/25214.htm

[2] M. Simsek, A. Czylwik, A. Galindo-Serrano, and L. Giupponi, "Improved decentralized Q-learning algorithm for interference reduction in LTE-femtocells," in *IEEE Wireless Advanced*, Jun. 2011.

[3] V. Tangkaratt, A. Abdolmaleki, and M. Sugiyama, "Deep Reinforcement Learning with Relative Entropy Stochastic Search," May 2017. [Online]. Available: https://arxiv.org/abs/1705.07606

[4] B. Muhammad and A. Mohammed, "Uplink closed loop power control for LTE system," in *Proc. Int. Conf. on Emerging Tech.*, 2010.

[5] Z. Gao, B. Wen, L. Huang, C. Chen, and Z. Su, "Q-learning-Based Power Control for LTE Enterprise Femtocell Networks," *IEEE Systems Journal*, vol. 11, no. 4, Dec. 2017.

[6] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Net.*, Nov. 2014.

[7] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I - Theory*. Now Publishers, 2009.

[8] Cisco. WLAN Radio Frequency Design Considerations. [Online]. Available: https://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/emob30dg/RFDesign.html

[9] R. S. Sutton and A. G. Barto, *Intro. to Reinforcement Learning*, 1998.

[10] S. Koenig and R. Simmons, "Complexity Analysis of Real-Time Reinforcement Learning," in *AAAI Conf. on Artif. Intelligence*, 1993.

[11] L. Yamamoto and J. Beerends, "Impact of Network Performance Parameters on the End-to-End Perceived Speech Quality," in *Proc. of Expert ATM Traffic Symposium*, 1997.

[12] F. B. Mismar. Q-Learning VoLTE Power Control Code. [Online]. Available: https://github.com/farismismar/Q-Learning-Power-Control