A New Spectrum Occupancy Model for 802.11 WLAN Traffic

Anurag Gupta, Satyam Agarwal, *Student Member, IEEE*, and Swades De, *Senior Member, IEEE*

***Abstract*—We report characterization of busy and idle periods in 802.11 WLAN via experiments in software-defined radio. A new channel occupancy model, called Gaussian mixture (GM) model, is proposed and shown to fit the empirical data significantly better than the other available models. Further simulation results demonstrate that, utilizing GM model highly-improves secondary user goodput and energy efficiency.**

***Index Terms*—802.11 WLAN channel availability model, Gaussian mixture (GM) model, software defined radio, cognitive radio**

I. Introduction

Cognitive radio (CR) is an intelligent communication paradigm where the secondary users (SUs) dynamically exploit the spectral and temporal voids in the licensed spectrum. Proper modeling of licensed/primary user (PU) activity is essential for CR user to maximize spectrum utilization while ensuring quality of service (QoS) to the PUs.

In this letter, based on extensive measurement experiments we propose a new PU (WiFi user) activity model for wireless local area networks (WLAN). The proposed model can be used for improved opportunistic access of WLAN spectrum. Our study is motivated by the emergence of LTE-U (Long Term Evolution - Unlicensed), where an LTE-U user (SU) can operate in an unlicensed band such as that of 802.11g WLAN.

Various medium access control (MAC) methods to exploit voids of the PU channels mostly consider exponentially distributed PU activity (ON/OFF) durations (e.g., [1]–[3]). A few experimental studies have reported that the ON/OFF durations are not truly exponentially distributed [4]–[6].

To model busy/idle periods in 802.11b WLAN, traffic sources taken were UDP [4], [5], [7], Skype voice [7], and HTTP [4]. In [5], the idle period was modeled as generalized Pareto distributed, while [4] modeled it as hyper-Erlang distributed. These studies were done in interference-controlled settings. The spectrum usage models in [6] did not consider WLAN traffic. Though general environment was considered in [8] to model idle period, it was on 802.15.4 standard with QoS objectives different from those in 802.11. Also, these empirical models were not evaluated on any CR-MAC protocols.

Different from the prior art, we conduct measurement experiments on 802.11 WLAN traffic, in both interference-controlled and general environments. Based on our observations, we propose a new WLAN channel occupancy model, called Gaussian mixture (GM) model. For performance comparison with the other considered models (exponential [3], generalized Pareto [5], and hyper-Erlang [4]), statistical goodness-of-fit test is conducted using the empirical data. Besides, performance of the competitive models are studied by employing them in two recent CR-MAC protocols [1], [2]. Novelty and significance of the work are as follows:

1) The GM model is proposed via extensive measurements in interference-controlled as well as general environment.
2) Goodness-of-fit of the GM model is significantly better than the available WLAN channel occupancy models.
3) Compared to the other occupancy models, GM model yields notably-improved CR-MAC performance.

This work has been supported in parts by the ITRA Media Lab Asia project under Grant no. ITRA/15(63)/Mobile/MBSSCRN/01 and the Department of Science and Technology under Grant no. SB/S3/EECE/0248/2014.

The authors are with the Department of Electrical Engineering and Bharti School of Telecommunication, Indian Institute of Technology Delhi, New Delhi 110016, India (e-mail: anurag.gupta137@gmail.com; satyam6099@gmail.com; swadesd@ee.iitd.ac.in)

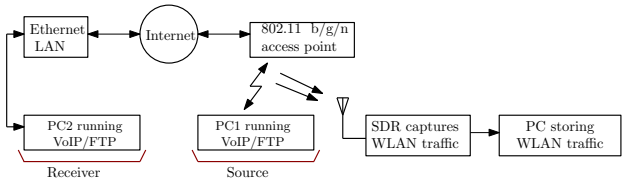

Fig. 1: Experimental setup in interference-controlled environment.

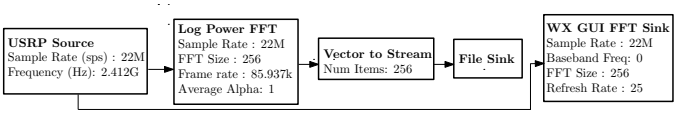

Fig. 2: Algorithm flowgraph for power sample measurement in SDR.

II. WiFi user (PU) activity characterization

A. Experimental setup

In the interference-controlled environment, we consider two application types: Skype video over IP (VoIP) and file transfer protocol (FTP), over 802.11g WLAN in absence of any other traffic. One computer (PC1) was connected to the Internet via 802.11g WLAN access point (AP) operating at 2.412 GHz, while the other (PC2) was connected via Ethernet LAN (cf. Fig. 1). A VoIP Skype call was setup between PC1 and PC2 for a duration of 5 minutes. A similar setup was used for FTP. The spectrum occupancy was observed by recording the received power samples using Amitec software defined radio (SDR) (amitec.co) at 22 MSps rate and 256 point FFT. Fig. 2 presents flowgraph of the algorithm implemented in the SDR. The recorded power samples were processed in MATLAB.

In general environment, multiple WiFi-enabled devices operate concurrently over a WLAN channel. The devices connect to a public 802.11/b/g/n AP. Specifically, we measured the channel usage in a public place. Power samples were collected similarly as in the interference-controlled environment.

The experiments were repeated 3 times in each environment.

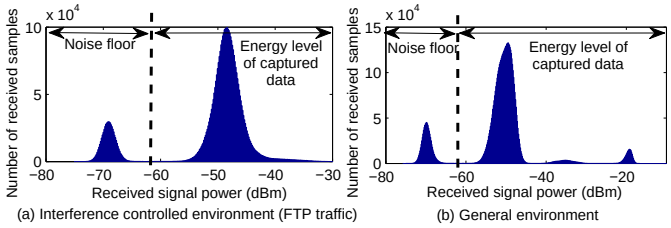

Fig. 3: Power sample histograms for WLAN traffic.

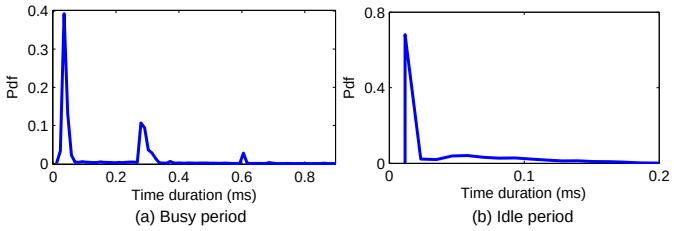

Fig. 4: Busy/idle period pdf of WLAN traffic in general environment.

Algorithm 1: EM algorithm for K-GM model.

1. Initialize μ_j, σ_j, w_j and evaluate $\mathcal{L}_K$ from (2)

do

2. **E step**: Evaluate $p(j|x_i)$ for each x_i in $\mathbf{x}$ as:

$$p(j|x_i) = \frac{w_j \mathcal{N}(x_i|\mu_j, \sigma_j)}{\sum_{j=1}^{K} w_j \mathcal{N}(x_i|\mu_j, \sigma_j)}$$

3. **M step**: Maximize $\mathcal{L}_K$ by equating its partial derivative with respect to μ_j, σ_j, and w_j to 0 for new estimate of the parameters. Updated parameters are:

$$w_j^{new} = \frac{\sum_{i=1}^{N} p(j|x_i)}{N}, \quad \mu_j^{new} = \frac{\sum_{i=1}^{N} p(j|x_i) x_i}{\sum_{i=1}^{N} p(j|x_i)},$$

$$\text{and} \quad \sigma_j^{new} = \frac{\sum_{i=1}^{N} p(j|x_i)(x_i - \mu_j^{new})^2}{\sum_{i=1}^{N} p(j|x_i)}$$

4. Compute the new log-likelihood $\mathcal{L}_K^{new}$

while $|\mathcal{L}_K^{new} - \mathcal{L}_K^{old}| > \epsilon$;

B. WLAN channel (PU) activity modeling

Energy detection method [9] is used to identify the channel state from the power samples. Channel is marked idle/*OFF* (respectively, busy/*ON*) if the energy of the captured samples is below (respectively, above) a threshold. Histogram of the captured samples are plotted in Fig. 3. The first peak indicates the 'noise floor' and the others represent the signal power. In general environment, multiple peaks after the 'noise floor' are due to the transmissions from multiple devices. To minimize false alarm probability, threshold is set 3 dB higher than the saddle point after the peak noise power.

Fig. 4 shows probability density function (pdf) of the samples for busy/idle periods in general environment. The pdfs in interference-controlled environment look similar. Each pdf is noted to consist of multiple Gaussian components. This led to our intuition of GM for channel occupancy modeling. GM model [10] is a pdf with weighted sum of multiple Gaussian components. K-component GM (K-GM) is expressed as:

$$f_K(x) = \sum_{j=1}^{K} w_j \mathcal{N}(x|\mu_j, \sigma_j), \text{with } w_j \geq 0 \text{ and } \sum_{j=1}^{K} w_j = 1. \quad (1)$$

$\mathcal{N}(x|\mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$ is the jth component, and w_j is the mixing weight. Consider that N i.i.d. sample points $\mathbf{x}=\{x_1, x_2, \cdots, x_N\}$ are observed. The goal is to obtain the optimal parameters (K and $\mu_j, \sigma_j, w_j, \forall j \in \{1, \cdots, K\}$) such that likelihood of the samples matching the proposed model is maximized. The log-likelihood function is defined as:

$$\mathcal{L}(\mathbf{x}, f_K) \triangleq \mathcal{L}_K = \sum_{i=1}^{N} \ln \left\{ \sum_{j=1}^{K} w_j \mathcal{N}(x_i|\mu_j, \sigma_j) \right\}. \quad (2)$$

C. Model parameter estimation

An iterative **E**xpectation **M**aximization (EM) algorithm [11] is used to find the optimal parameters (μ_j, σ_j, w_j) (cf. Algorithm 1). Each iteration has 2 steps: E and M. In the E step, posterior probability $p(j|x_i)$ of each GM component j is evaluated for each data point x_i using the current parameters. In the M step, $p(j|x_i)$'s from E step are used to obtain new estimates of the parameters.

EM algorithm guarantees convergence and its complexity is $O(KN^2)$. However, the algorithm is sensitive to the initial parameter values and it can converge to a local maxima.

To achieve global optimality, a greedy learning heuristic was proposed in [12]. For faster convergence, we propose the following modifications: initialization using observations from the available data set, and exploitation search for obtaining candidate Gaussian component.

In the modified algorithm (Algorithm 2), initialization is carried out by identifying the significant peaks in the empirical pdf of the data (e.g., 2 significant peaks in Fig. 4(a)), each peak forming a Gaussian component. Starting with this initial number, the algorithm is iterated. At each iteration, a new optimal Gaussian component is added to the existing GM model and the EM algorithm is applied to maximize $\mathcal{L}_K$.

Kolmogorov-Smirnov (KS) test (a goodness-of-fit measure) [13] is used as a stopping criterion for fitting a model. KS distance is the maximum distance between empirical cumulative distribution function (CDF), $\mathcal{G}(x)$ and the estimated CDF, $F(x;K)$ for K-GM model, which is defined as $\mathcal{D}(K) = \max_x |\mathcal{G}(x) - F(x;K)|$. The algorithm is stopped when the change in KS distance between two successive GM models is below a threshold ϵ.

To obtain a new optimal component in step 4 of Algorithm 2, $(m+1)k$ candidate components are generated from the current k-GM mixture. Data set $\mathbf{x}$ is divided in k disjoint sets $Z_n = \{x_i \in \mathbf{x} : n = \text{argmax}_j \ p(j|x_i)\}$. $m+1$ candidate components are constructed for each set Z_n. Two search techniques, exploration and exploitation, are used. In exploration search, m candidates are generated by randomly picking two data points x_{nl} and x_{nr} from Z_n at a time. Z_n is partitioned in two disjoint subsets Z_{nl} and Z_{nr} based on the closeness of the data points in Z_n to x_{nl} and x_{nr}. The

Algorithm 2: Modified greedy learning algorithm.

1. Identify k significant peaks in the data pdf
2. Obtain μ_j, σ_j, w_j $\forall j \in \{1, \cdots, k\}$ for k-GM as:
(i) $\mu_j \leftarrow$ time of occurrence of the jth peak in the pdf
(ii) $\sigma_j \leftarrow \frac{\mathbf{x}_\sigma}{N}$, where $\mathbf{x}_\sigma$ is the sample variance
(iii) $w_j = \frac{y_j}{\sum_{i=1}^{k} y_i}$ where y_j is the amplitude of the jth Gaussian peak identified in step 1
3. Apply EM algorithm without initialization to compute optimal μ_j, σ_j, w_j in k-GM; obtain $\mathcal{D}(k)$
do
4. Obtain a new optimal component $\mathcal{N}(x|\mu^*, \sigma^*)$ with mixing weight β^* such that:

$$\{\mu^*, \sigma^*, \beta^*\} = \underset{\{\mu,\sigma,\beta\}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln[(1-\beta) f_k(x_i) + \beta \mathcal{N}(x_i|\mu, \sigma)]$$

5. Add the optimal component to the k-GM. Set: $f_{k+1}(x_i) \leftarrow (1-\beta^*) f_k(x_i) + \beta^* \mathcal{N}(x_i|\mu^*, \sigma^*) \forall x_i \in \mathbf{x}$
6. Apply EM algorithm to update $f_{k+1}(x)$
7. Set $k \leftarrow k+1$ and compute $\mathcal{D}(k)$
while $|\mathcal{D}(k) - \mathcal{D}(k-1)| > \epsilon$;

mean and variance of sets Z_{nl} and Z_{nr} give the parameters μ and σ of the two candidate components. In a similar way, other $m-2$ components are generated. One component is generated from the exploitation search by subtracting the nth estimated Gaussian component pdf from the empirical pdf over the set Z_n. The mean and variance of the resulting difference give μ and σ of the $(m+1)$th candidate component in the set Z_n. The initial mixing weights β for the candidate components of set Z_n are set to $\frac{w_n}{2}$. In this way, we generate total $(m+1)k$ candidate components and search over those candidates with parameters $\{\mu, \sigma, \beta\}$ in step 4, to obtain the optimal component.

Complexity of the algorithm is $O(NK^2)$ [12], where K is the number of Gaussian components. Simulations indicate that our proposed modified greedy learning algorithm is on average 6% faster than the conventional greedy algorithm with $m = 6$.

D. Model fitness results

CDFs of busy/idle periods in general environment are plotted in Fig. 5. For fitting the empirical CDF with exponential and generalized Pareto distributions maximum likelihood estimation is used, while the parameters of hyper-Erlang distribution are obtained by EM approach. K-GM is noted to fit the empirical CDF more accurately than any other models. The results in interference-controlled environment are similar.

Table I shows $\mathcal{D}$ values for various traffic in both environments (average of 3 experiments). Significance level of 1% is considered. In 4-GM and 5-GM, KS test does not reject the null hypothesis, while for others it rejects the null hypothesis. Compared to GM model with $K \geq 4$, $\mathcal{D}$ for exponential, generalized Pareto, and hyper-Erlang are noted to be large - *hence showing poor fit*. GM model performance nearly saturates for $K \geq 4$. So, we choose $K = 4$ to fit the empirical data. Table II presents the GM parameters in general

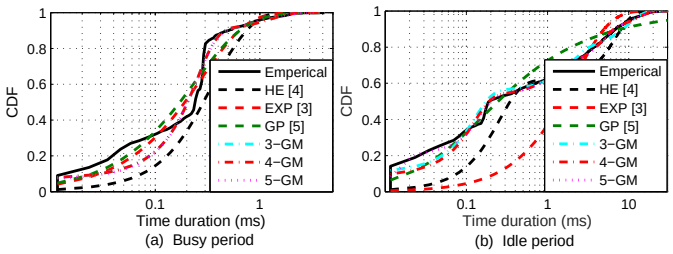

Fig. 5: CDF of busy and idle periods in general environment. Exp: exponential, GP: generalized Pareto, HE: hyper-Erlang.

TABLE I: $\mathcal{D}$ for various fitting model distributions.

	Model	Interference-controlled environment				General environment	
		VoIP Skype		FTP			
		Busy	Idle	Busy	Idle	Busy	Idle
$\mathcal{D}$	**Exp**	0.5635	0.6810	0.1800	0.5100	0.4743	0.4766
	GP	0.2593	0.1946	0.2110	0.2780	0.2811	0.2227
	HE	0.3069	0.3174	0.1930	0.3910	0.3206	0.2709
	1-GM	0.4947	0.4912	0.3890	0.5260	0.4816	0.4297
	2-GM	0.2804	0.1675	0.3890	0.4080	0.2042	0.1741
	3-GM	0.0952	0.1675	0.1650	0.2870	0.1998	0.1546
	4-GM	0.0952	0.1372	0.1650	0.2250	0.1417	0.1333
	5-GM	0.0952	0.1356	0.1460	0.2250	0.1263	0.1109

TABLE II: GM model parameters in general environment.

Components		1	2	3	4
Busy	μ_j **(ms)**	0.18	0.19	0.74	1.5
	σ_j **(ms)**	$1.57\text{x}10^{-5}$	$1.57\text{x}10^{-5}$	$5.21\text{x}10^{-5}$	$4.17\text{x}10^{-4}$
	w_j	0.45	0.44	0.06	0.05
Idle	μ_j **(ms)**	0.085	0.93	4	10.8
	σ_j **(ms)**	$5.41\text{x}10^{-6}$	$3.21\text{x}10^{-4}$	$3.76\text{x}10^{-3}$	$2.92\text{x}10^{-2}$
	w_j	0.51	0.18	0.21	0.10

environment. *The other PU activity parameters, such as, channel occupancy ratio, PU's channel access probability, can be predicted from the discussed PU occupancy distribution.*

III. Performance analysis of GM model

To study the GM model we consider two MAC protocols [1], [2] that require to estimate PU traffic for their operation.

A. Description of MAC protocols and performance measures

In [1], the proposed eDSA V.2 protocol was to maximize the unused PU channel utilization, where the PU busy/idle periods were considered exponentially distributed. The SU vacates the channel for optimal length v each time it senses the channel busy, while it transmits a data packet of optimal length l on sensing the channel idle. CR performance is measured in terms of *SU goodput* G (data transmitted per unit time) and *energy efficiency* E (data transmitted per unit energy consumption) while maintaining the fraction of overlapping time R_c (fraction of time the SU transmission interferes with PU transmission) below a threshold r_0. For the GM model, these metrics are found as:

$$R_c(l,v) \approx \frac{l}{2(\mathbb{E}[I]+\mathbb{E}[B])}; G(l,v) \approx \frac{\int_0^\infty (1-\tilde{F}_I(x))dx - l}{\mathbb{E}[I]+\mathbb{E}[B]};$$

$$E(l,v) \approx \frac{\int_0^\infty (1-\tilde{F}_I(x))dx - l}{\int_0^\infty \left((\phi_t + \frac{\phi_s}{l})(1-\tilde{F}_I(x)) + (\phi_i + \frac{\phi_s}{v})(1-\tilde{F}_B(x))\right)dx}.$$

Expected SU transmission time in a busy-idle cycle is $\int_0^\infty (1-\tilde{F}_I(x))dx$, of which $\int_0^\infty (1-\tilde{F}_I(x))dx - l$ is successful. Expected time the SU spends in sensing and idling in a busy period is $\int_0^\infty (1-\tilde{F}_B(x))dx$. ϕ_t and ϕ_i are respectively the data transmission and idling power consumption, while ϕ_s is the energy consumption per channel sensing. $\mathbb{E}[I]$ and $\mathbb{E}[B]$ are respectively the expected idle and busy periods, and $\tilde{F}_I(x)$ and $\tilde{F}_B(x)$ are respectively the CDFs of residual idle and busy period, given as:

$$\tilde{F}_I(x) = p(X \leq x + \frac{v}{2} | X > \frac{v}{2}) = \frac{F_I(x+\frac{v}{2}) - F_I(\frac{v}{2})}{1 - F_I(\frac{v}{2})},$$

$$\tilde{F}_B(x) = p(X \leq x + \frac{l}{2} | X > \frac{l}{2}) = \frac{F_B(x+\frac{l}{2}) - F_B(\frac{l}{2})}{1 - F_B(\frac{l}{2})}.$$

$F_I(x) = 1 - \sum_{i=1}^{K} w_i Q\left(\frac{x-\mu_i}{\sigma_i}\right)$ is the CDF of idle period obtained from K-GM model, where $Q(\cdot)$ is the Q-function. $F_B(x)$ is the CDF of busy period, obtained similarly.

The optimization problems (G^*; E^*) are formulated as:

$$G^* = \max_{l,v} G(l,v); E^* = \max_{l,v} E(l,v), \text{ s.t. } R_c(l,v) \leq r_0$$

Obtaining closed-form expressions for G^* and E^* is difficult. Hence, we use derivative-free numerical optimization technique (Nelder-Mead method) to compute G^* and E^*.

In RIBS [2], for a new frame transmission, an SU generates exponentially distributed random back-off times to decide sensing instants. When the channel is sensed idle, it transmits the data based on its estimated maximum duration y_{max} so that PU interference probability is below a threshold η. Denoting the residual idle time CDF as F_{RI}, y_{max} is obtained as:

$$F_{RI}(y) = \int_0^y \frac{1-F_I(z)}{\mathbb{E}[I]} dz; \; y_{max} = \max\{y : F_{RI}(y) \leq \eta\}. \quad (3)$$

To obtain y_{max} for K-GM model, we equate $F_{RI}(y_{max}) = \eta$ (reduced to equality constraint). We then have:

$$\sum_{i=1}^{K} \frac{w_i(y_{max}-\mu_i)}{\sigma_i} Q\left(\frac{y_{max}-\mu_i}{\sigma_i}\right) - \sum_{i=1}^{K} \frac{w_i}{\sqrt{2\pi}} e^{-\frac{(y_{max}-\mu_i)^2}{2\sigma_i^2}} + \sum_{i=1}^{K} w_i \left[\frac{\mu_i}{\sigma_i} Q\left(\frac{-\mu_i}{\sigma_i}\right) + \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu_i^2}{2\sigma_i^2}}\right] = \eta \mathbb{E}[I]. \quad (4)$$

Having no closed-form solution for y_{max} in (4), numerical method is used. Since $F_{RI}(y)$ is an increasing function in y, we use bisection method to obtain y_{max}. RIBS performance measures are *SU goodput* and *energy efficiency*.

B. Performance results and remarks

Key simulation parameters are [14]: data rate 6 Mbps, slot size 11.64 μs; energy/slot for sensing 14 μJ, transmission 24.3 μJ, and idling 5.9 μJ.

Figs. 6(a) and 6(b) show, G^* and E^* of the eDSA V.2 in [1] with the GM model clearly performing better compared to exponential model. Specifically, at $r_0 = 0.1$ GM model yields respectively 17.3% and 6.8% higher G^* and E^*. At $r_0 = 0.2$, these gains are 10.9% and 6.5%. *We remark that, the GM model performs better because it captures the WLAN activity much accurately as compared to the exponential model.*

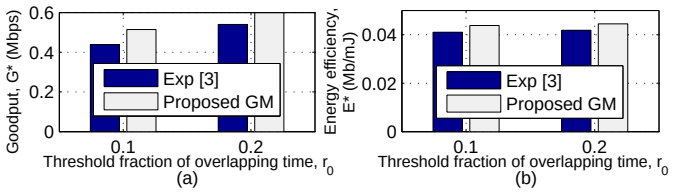

Fig. 6: Performance of eDSA V.2 in [1] with the two PU channel models. (a) SU goodput; (b) energy efficiency. Exp: Exponential.

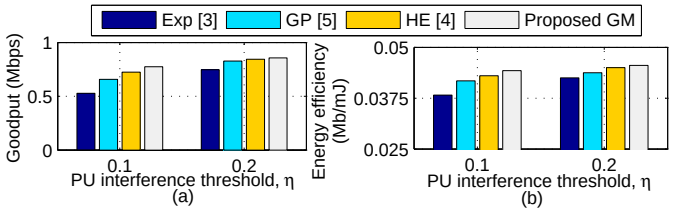

Fig. 7: RIBS [2] performance: (a) SU goodput; (b) energy efficiency.

Figs. 7(a) and 7(b) show SU goodput and energy efficiency in RIBS [2]. GM model offers respectively 30.8%, 10.7%, and 4.2% higher goodput on average, compared to exponential, generalized Pareto, and hyper-Erlang. For energy efficiency, the respective gains are 11.4%, 5.1%, and 2.1%.

These results further corroborate the observations from Fig. 5 that, the K-GM with $K \geq 4$ models the WLAN channel much better, aiding the CR performance.

REFERENCES

[1] S. Agarwal and S. De, "eDSA: Energy-efficient dynamic spectrum access protocols for cognitive radio networks," *IEEE Trans. Mobile Comput.*, 2016. [Online]. Available: 10.1109/TMC.2016.2535405

[2] M. Sharma and A. Sahoo, "Stochastic model based opportunistic channel access in dynamic spectrum access networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 7, pp. 1625–1639, Jul. 2014.

[3] Q. Jiang *et al.*, "Energy-efficient adaptive rate control for streaming media transmission over cognitive radio," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4682–4693, Dec. 2015.

[4] S. Geirhofer *et al.*, "Cognitive radios for dynamic spectrum access - Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Commun.*, vol. 45, no. 5, pp. 66–72, May 2007.

[5] ——, "A measurement-based model for dynamic spectrum access in WLAN channels," in *Proc. IEEE MILCOM*, Washington, DC, Oct. 2006.

[6] M. Lopez-Benitez and F. Casadevall, "Time-dimension models of spectrum usage for the analysis, design, and simulation of cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2091–2104, Jun. 2013.

[7] S. Geirhofer *et al.*, "Dynamic spectrum access in WLAN channels: Empirical model and its stochastic analysis," in *Proc. Wksp. TAPAS*, Boston, MA, Aug. 2006.

[8] L. Stabellini, "Quantifying and modeling spectrum opportunities in a real wireless environment," in *Proc. IEEE WCNC*, Sydney, Australia, Apr. 2010.

[9] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, Apr. 1967.

[10] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Springer, 2009, pp. 659–663.

[11] Z. Chen *et al.*, *Correlative Learning, Appendix E: Expectation-Maximization Algorithm*. John Wiley & Sons, Inc., 2007.

[12] J. J. Verbeek *et al.*, "Efficient greedy learning of Gaussian mixture models," *Neural Computation*, vol. 15, pp. 469–485, 2003.

[13] R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit Techniques*. CRC Press, 1986.

[14] S. Maleki *et al.*, "Energy-efficient distributed spectrum sensing for cognitive sensor networks," *IEEE Sensors J.*, vol. 11, no. 3, pp. 565–573, Mar. 2011.