White Space Prediction for Low-power Wireless Networks: A Data-Driven Approach

Indika S. A. Dhanapala[1,*], Ramona Marfievici[1], Sameera Palipana[1], Piyush Agrawal[2], Dirk Pesch[1]
[1]Nimbus Centre for Embedded Systems Research, Cork Institute of Technology, Cork, Ireland
[2]United Technologies Research Centre, Cork, Ireland
[*]Contact Author: I.S.A.Dhanapala@mycit.ie

***Abstract*—In the 2.4 GHz unlicensed spectrum, the coexistence of WiFi, Bluetooth and IEEE 802.15.4 devices generates increased channel contention. Notably, low-power wireless networks experience packet loss and delays due to interference. To improve the performance of low-power wireless networks under interference, we propose a data driven proactive approach based on interference modeling for white space prediction. We leverage statistical analysis of real-world traces from two indoor environments characterized by varying channel conditions to identify interference patterns. We characterize interference in terms of Inter-Arrival Time (IAT) and number of interfering signals and use a Gaussian Mixture Model (GMM) to accurately estimate the interference distribution as observed by the low-power wireless nodes. Then, we use a Hidden Markov Model (HMM) for white space prediction. Our validation w.r.t. real-world traces from two environments show that our GMM model can estimate interference with an accuracy higher than 94.7%. Moreover, the white space prediction evaluation shows an average accuracy of 97.7% and 89.5% across the two environments.**

***Keywords*-Cross Technology Interference, low-power wireless communication, wireless sensor networks, interference modeling, white space, predictive models**

I. Introduction

Wireless communication systems operating in unlicensed radio spectrum, such as the 2.4 GHz ISM band, suffer from Cross Technology Interference (CTI), which is the overlapping of transmissions from different systems in time and frequency. The interference occurs due to the broadcast nature of wireless transmissions of co-located devices of different technologies such as IEEE802.11 (WiFi), IEEE802.15.1 (Bluetooth) or IEEE802.15.4 and who cannot coordinate their transmissions. CTI creates packet losses, increases channel contention which increases the delay, and ultimately under-utilizes the scarce frequency spectrum [1], [2].

These problems are exacerbated for the IEEE802.15.4 based low-power wireless networks, our focus in this paper. In the presence of the interference, low-power wireless nodes need to adapt to changing interference patterns and adjust their transmission schedules in order to avoid interfering transmissions and maximize the reliability of their communication. To achieve this, nodes need to acquire a detailed understanding of the surrounding interference through interference power measurements. Using these measurements, nodes can then parameterize *interference estimation* and *white space prediction* models in order to schedule their channel access or tune their communication protocols accordingly.

Recent solutions [3] provide white space prediction using Pareto models, relying on the assumption that interfering data traffic exhibits heavy-tailed distributions. However, our interference traces contradict this assumption. We accounted for this aspect in our previous approach [4], and used a combination of a 2nd order Markov Modulated Poisson Process (MMPP(2)) model for interference estimation and a Hidden Markov Model (HMM) for white space prediction. Despite the encouraging results, our approach was validated in a limited set of settings.

In this paper, we exploit a larger set of real-world data traces to create models for both estimating the interference and predicting white spaces. We characterize and analyze the traces using the mean interference Inter Arrival Time (IAT) and the number of interference signals in a slot of fixed duration. The analysis revealed: i) interference traces of arbitrary distribution, and ii) the presence of peak and off-peak patterns. The first motivated us to evaluate the potential of a Gaussian Mixture Model (GMM) for modeling the interference, while the second led to the necessity of using two interleaved models to account for the observed patterns. Then the estimated interference generated with the GMM is used as input for a HMM to predict white spaces (i.e., when the communication channel is interference FREE), so as to allow low-power wireless nodes to better schedule their transmissions.

The accuracy of our GMM-based interference estimation model is evaluated w.r.t. the ground truth traces. Our results show that the accuracy we obtain with our approach, over 94.7% in all tested cases, is significantly superior to the state-of-the-art approaches. The accuracy of the white space prediction is 97.7% and 89.5% in the two tested environments. Moreover, when the white space prediction is used by an application to schedule its transmissions, the Packet Loss Ratio (PLR) is 2.3% under moderate interference and 10.5% under heavy interference.

The rest of the paper is organised as follows. We concisely summarize the characteristics of the collected interference traces in Section II. Our approach is described in Section III and evaluated in Section IV. We discuss limitations of our models and explore possible future work in Section V. We end the paper by surveying the related work in Section VI, followed by brief concluding remarks in Section VII.

II. TRACES

The central pillar of our paper are the traces acquired in real-world environments. Therefore, the first contribution of this paper is the analysis of a large set of interference traces. The location of the experiments was chosen to cover different interference conditions. The design of the data traces collection was informed by our interest in understanding the interference and its short- and long-term, channel and location variations.

A. Measuring the Interference

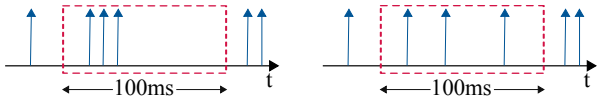
Location. Our study areas are two typical indoor environments: OFFICE and HOME, covering different conditions of interference. The first is an office building, while the other is a student dormitory exhibiting more bursty traffic.

Hardware/software platforms. The interference measurements have been acquired by TMote Sky nodes, equipped with the ChipCon2420 radio chip compliant with IEEE802.15.4. Each node was connected to the USB port of a PC. To enable fast interference detection, we used the Clear Channel Assessment (CCA) and Start of Frame Delimiter (SFD) pins of the CC2420 transceiver, leveraging the experience from [3]. In this case, when a signal above the CCA threshold (i.e., -77 dBm, the default value used for the CC2420 transceiver) is detected, the CCA pin goes low indicating a busy channel, while the SFD pin high indicates the start of an incoming IEEE802.15.4 packet. We captured the hardware interrupts of the two pins, with both CCA and SFD low indicating the presence of an interference signal. When this occurs, a packet is sent from the node to the PC, it is timestamped and stored for future processing and analysis.

Data collection execution. The main findings reported in this paper were gathered in three experimental campaigns. The FIRST and the SECOND were performed in OFFICE. During the FIRST we deployed three nodes at the same location between two WiFi APs (Access Points) detecting the interference on IEEE802.15.4 channels 13, 18 and 23. For the SECOND, we interleaved the three nodes and the APs, and used only channel 18. These choices allowed us to explore interference traces from IEEE802.15.4 channels overlapping with different WiFi channels and different characteristics of the WiFi traffic. The collection of traces was executed for 24 hours, during a working day of the week, from 1:00PM or 4:00PM. For the THIRD campaign, a single node was used to collect traces on channel 18 from OFFICE and HOME. As the goal was to assess the interference for long-term, we ran the campaign for two weeks, during September 12–26, 2017.

B. Interference Characterization

Methodology. We recall that our goal is to predict white spaces for the low-power wireless networks in the presence of interference. For this, we divided the time axis into slots of 100 ms duration. We empirically determined that higher values of the slot length induce high accuracy in prediction but reduce the throughput of the application. Therefore, 100 ms turned out to be a good trade-off for both. Note that we define a white space as the length in time in which an IEEE802.15.4 packet and its ACK can be transmitted without preemption. Since we use time-slots, the lower and upper boundaries of the length of a white space are 8.512 ms and 100 ms respectively. We characterized the traces in terms of mean IAT and number of arrival signals per slot statistics. Although mean IAT is the most directly informative statistical property of the interference trace, if used alone to characterize the traffic within a slot, leads to an increase in the false discovery rate of bursts of interference signals (BUSY periods).



(a) Burst arrivals. (b) Scattered arrivals.

Fig. 1: False discovery of BUSY periods.

The intuition behind this behavior is shown in Fig. 1. For example, in a 100 ms slot, three interference signals can arrive in a burst (e.g. small mean IAT of 3 ms) or scattered (e.g. large mean IAT of 20 ms). We define an IAT threshold, TH_{IAT} = 8.512 ms, the maximum time on air for a 127 bytes IEEE802.15.4 packet and its ACK transmitted at a rate of 250 Kbps, in order to decide the state of the channel, BUSY or FREE. In our example, during the bursty arrivals, the channel is incorrectly identified as BUSY, while during the scattered arrivals the channel is correctly identified as FREE. During a 100 ms slot, a node can send 11 packets, considering the 8.512 ms time on air. If, in each 8.512 ms sub-slot an interference signal arrives, the slot is identified as BUSY. In this respect, the number of signal arrivals is key in reducing the false discovery rate of the BUSY periods. Therefore, a threshold for the *count* of signal arrivals per slot, TH_{count}, along with the TH_{IAT}, can be used to decide the state of the channel as follows:

$$Channel = \begin{cases} \text{BUSY}, & \text{if } IAT \leq TH_{IAT} \text{ and } count \geq TH_{\text{count}} \\ \text{FREE}, & \text{otherwise} \end{cases}$$

In our example from Fig. 1, the combined use of the two thresholds, correctly identifies the slot as FREE for both cases. Once the interference traffic trace is characterized in terms of mean IAT and number of interference signal arrivals per slot, we compute the two-dimensional probability distribution of the trace. Next, the hourly traffic patterns are identified by comparing the distribution with a one-hour peak-traffic distribution extracted from the trace. By comparing the trace with peak-hour traffic, we are able to classify peak traffic hours and off-peak traffic hours. To this end, Normalized Cross-Likelihood Ratio (NCLR) [5] was used with values of NCLR close to zero indicating highly similar distributions. Thus, peak

TABLE I: NCLR comparison of the interference traces from FIRST and SECOND.

FIRST				SECOND			
Channel	**13**	**18**	**23**	**Location**	**1**	**2**	**3**
13	0	0.78	0.22	**1**	0	0.12	0.88
18	0.78	0	1	**2**	0.12	0	1
23	0.22	1	0	**3**	0.88	1	0

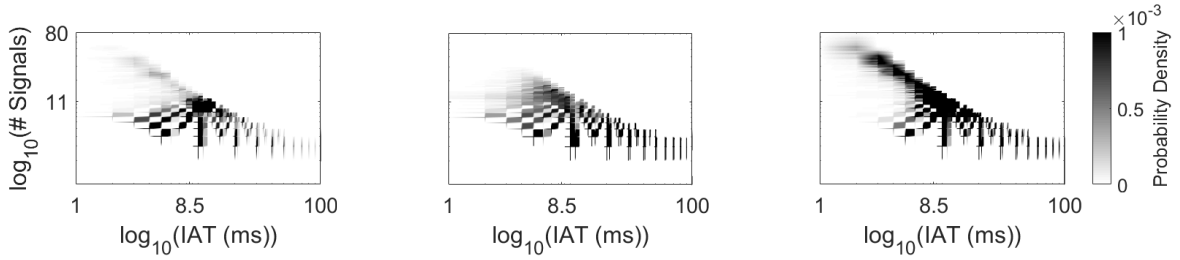


Fig. 2: Probability distribution function of traces from FIRST on channel 13 (left), 18 (center) and 23 (right).

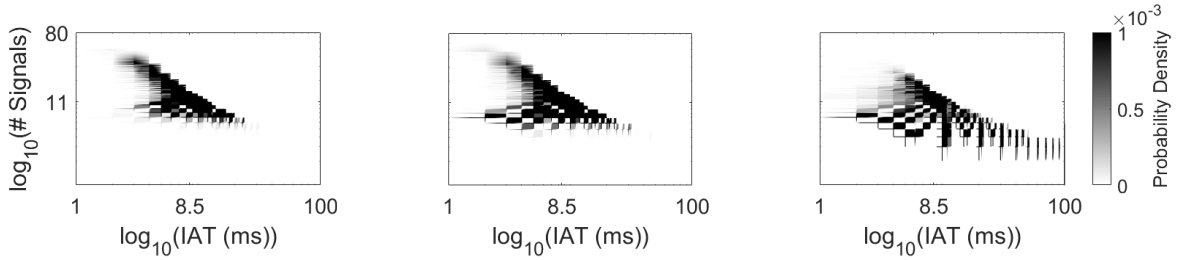
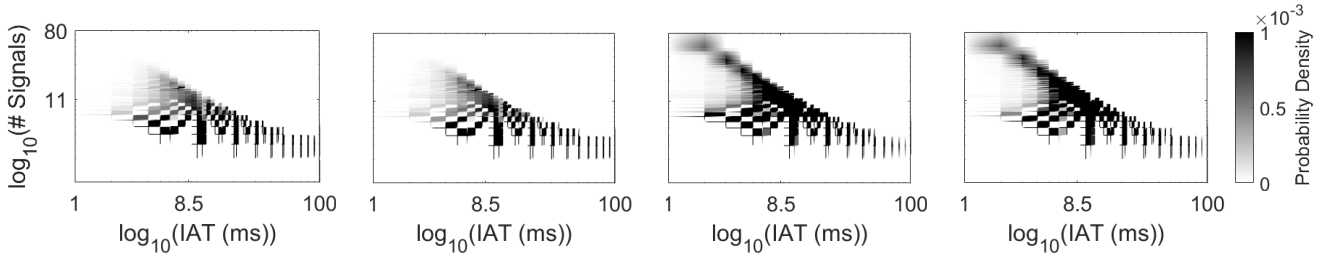


Fig. 3: Probability distribution function of traces from SECOND at location 1 (left), 2 (center), and 3 (right).



(a) First week OFFICE. (b) Second week OFFICE. (c) First week HOME. (d) Second week HOME.

Fig. 4: Probability distribution function of traces during THIRD from OFFICE and HOME.

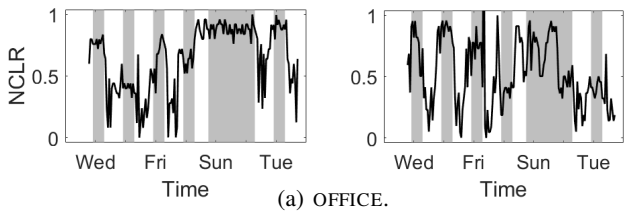
and off-peak traffic hours can be identified from the trace, which in turn are useful for training the model. A threshold value of 0.5 for NCLR is used to distinguish among the two.

Next, we characterize the interference traces in two ways. The first is to compute the probability density function (PDF) of the interference traces w.r.t. their mean IAT and number of signal arrivals, as shown in Fig. 2, 3, and 4. The other is to compute the NCLR for traces from the same campaign, from different channels (FIRST), locations (SECOND) or weeks (THIRD).

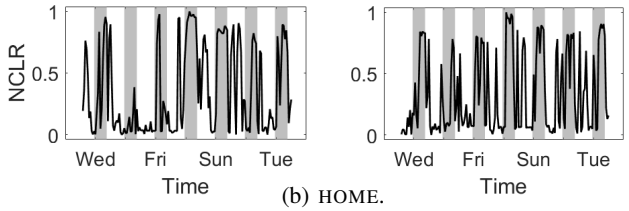
FIRST. Through the PDF lens, see Fig. 2, it appears like the interference on channel 13 and 23 is similar. This is further confirmed by the low value of $NCLR = 0.22$ in Table I. On the other hand, the interference on channel 18 is different. We conjecture that this is a combined effect of the IEEE802.15.4 channels overlapping with different WiFi channels and being interfered by different APs. In our OFFICE building environment the APs channel allocation is dynamic.

SECOND. The location of the nodes induces different trends in their PDFs, in this case location 1 and 2 show similar behavior, as shown in Fig. 3. This can also be seen in Table I with $NCLR = 0.12$, and explained by different interference characteristics induced by the position of the nodes in the proximity of two different APs (i.e., location 1 and 2 are close to AP1, while 3 is close to AP2).

THIRD. Fig. 4 shows the results from the THIRD campaign. A few trends are clearly identifiable. First, the quantity of the traffic increases as one progresses from OFFICE to HOME. The trend is more marked during the second week. Second, the traffic in HOME is more bursty than OFFICE, the PDFs show high probabilities in the bursty zone, mostly due to the video streaming done by students in HOME (i.e., student dormitory). We now turn our attention to variations induced by the interleave of night and day. Fig. 5 shows the NCLR obtained from the comparison of 1-hour peak trace with each 1-hour interference trace for both environments. In OFFICE the 1-hour peak trace represents the most busy traffic period during the day, while for HOME during the night. In the OFFICE, Fig. 5a, we easily identified patterns in the interference distribution over time of day and week-ends. The regions with high NCLR, match the outside of office hours (7:00-22:00) time and the week-ends (19:00 Saturday-7:00 Monday) when there is no activity in the OFFICE building, therefore less interference. Moreover, an increase in the interference, with NCLR decreasing close to zero, can be observed during the busiest office hours, 10:00-11:00 and 13:00-15:00. Interestingly, Thursday night of the first week and the week-end days of the second week, show an increase in the interference, that we ascribe to a set of experiments run in the OFFICE building. In HOME, the



(a) OFFICE.



(b) HOME.

Fig. 5: Traffic patterns in OFFICE and HOME for two weeks.

variations over time appear to be somewhat dependent on night (i.e., off-peak between 23:00 and 9:00) and day variations, but are not as clearly marked as in the OFFICE. Also, in HOME the range of variations between FREE and BUSY periods is more dramatic, while the BUSY periods are smoother (i.e., longer bursty interference periods) than in OFFICE. These are the effects of more users and devices (WiFi/Bluetooth/microwave ovens) in the students HOME than OFFICE, plus no strict access time policies.

In a nutshell, the observations from our experimental campaigns show that the environment in which the low-power wireless nodes are immersed, the location where the nodes are placed, and the channel used, have an impact on how the interference is perceived. Moreover, these observations directly inform modeling decisions, suggesting that at least two models accounting for the peak and off-peak interference patterns should be adopted.

III. Modeling Approach

Next, we build on the above analysis to exploit the set of traces to create two models: i) for estimating the interference, and for ii) predicting white spaces for low-power wireless nodes in the presence of interference.

A. Interference Estimation

Model. The fundamental motivation for our modeling approach for estimating the interference is that the observed traces display an arbitrary distribution and GMM models can produce smooth estimations of arbitrarily shaped distributions [6]. To this end, we use a GMM, whose defining parameters are the number of components ($\mathcal{M}$) and three matrices: mixture component weights ($\boldsymbol{W}$), component means ($\boldsymbol{\mu}$) and covariances ($\boldsymbol{\Sigma}$). The former is a stochastic matrix which determines the weight at which each Gaussian component should model data, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ define the mean and the covariance of each component. In our approach, we use two GMM models, for peak and off-peak periods.

The choice of the number of components ($\mathcal{M}$) affects the estimation accuracy. Moreover, each component ($\mathcal{M}$) has ($\mathcal{Q}$) dimensions given by the number of features used to characterize the distributions.

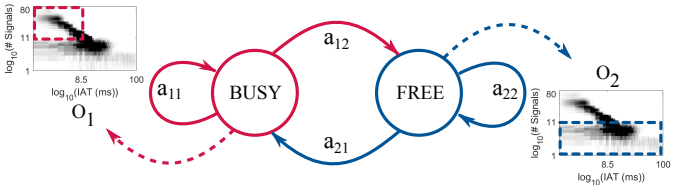


Fig. 6: Hidden Markov model.

Parameters. In our case, the components are two-dimensional ($\mathcal{Q} = 2$), the interference traces are characterized by the mean IAT and the number of signal arrivals per slot. The number of components is estimated empirically, by comparing the GMM model estimates w.r.t. the ground truth trace using Area Under Curve (AUC) as metric. In Section IV we show how this is done for our approach.

Training. Once the parameters are computed and set, the model can be trained. The matrices (i.e., $\boldsymbol{W}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$), of the GMM model were estimated using the expectation maximization (EM) algorithm. A diagonal covariance matrix $\boldsymbol{\Sigma}$, the most used in the literature, was adopted, requiring less samples for training, and approximating full covariance using a linear combination of diagonal covariances.

B. White Space Prediction

As discussed in Section I, the contribution we put forth here is an approach for predicting transmission opportunities for low-power wireless in the presence of interference. Next, we describe how we exploit the output of the GMM model, the estimated interference, for white space prediction. To this end, we use a HMM model.

We adopt the notation from [4] to indicate the complete parameter set of the HMM model: 1) hidden (unobserved) states $\boldsymbol{S} = \{\text{FREE}, \text{BUSY}\}$, correspond to the two different regimes of the wireless channel; 2) initial state probabilities $\boldsymbol{\pi}$; 3) observations $\boldsymbol{O} = \{o_1, o_2\}$, correspond to the two features used to characterize the interference, mean IAT and number of signal arrivals per slot; 4) state transition probability matrix $\boldsymbol{A}$, models the evolution of the wireless channel as transitions among the set of unobserved states; 5) observation probability matrix $\boldsymbol{B}$.

Fig. 6 shows a graphical representation of the HMM model, with transition probabilities overlaid on the arrows showing the state transitions (i.e., a_{11}, a_{12}, a_{21}, and a_{22}), and the emission distributions for each state represented by the modeled interference distributions corresponding to the BUSY and FREE states of the channel. The model parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ are initialized using uniformly distributed probability matrices while $\boldsymbol{\pi}$ is initialized for the data set under consideration, and all are recomputed using the *Baum-Welch* algorithm [7]. In addition, the training data used for the HMM is labeled as FREE or BUSY with the help of the two thresholds, TH_{IAT} and TH_{count}, introduced in Section II-B. In our approach, we use two HMM models, for peak and off-peak periods.

IV. Performance Evaluation

We validate our approach for white space prediction by: i) conducting a statistical comparison between the interference

data traces (training set) we collected from the two indoor environments and estimated traces from the GMM model, a state-of-the-art Pareto model and from our previous proposed MMPP(2) model; ii) comparing our prediction from HMM model with a 0.5- and 1-persistent random access method and our previous approach [4]. In our evaluation we use collected traces from all campaigns.

A. Metrics

To establish the number of components for the GMM model, we computed the AUC, searching for the optimal operating point for the model while varying the number of components, and minimizing the *False Positive Rate* (FPR), and maximizing the *True Positive Rate* (TPR). To assess the performance of our approach, we considered two metrics: *accuracy* and FPR. The former is a measure of the predictability of the model, while the latter provides an assessment of the packet loss of the low-power wireless network when the prediction mechanism is being used. All metrics are derived from the elements of the confusion matrix in Table II, as follows: $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$, and $\mathrm{accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$.

B. Model Parameter Selection

GMM. The number of components was empirically identified, varying it from three to ten and computing the AUC. The results indicated that seven components ($\mathcal{M} = 7$) are enough for a satisfactory accuracy of 99.9% of the estimated interference.

HMM. The training duration of HMM determines the accurate estimation of the model parameters and, consequently, the white space prediction. A 1-hour training set provides statistical relevance for the channel behavior. The training trace for peak and off-peak models are obtained by computing the mean NCLR for the traces and picking the 1-hour trace closest to this mean. The matrices that characterize the model, $\boldsymbol{A}$ and $\boldsymbol{B}$ are initialized with uniformly distributed probabilities. Moreover, the matrix $\boldsymbol{\pi}$ in the peak and off-peak models is initialized using the proportion of FREE, BUSY slots from the total number of FREE and BUSY slots.

C. Interference Estimation Validation

Our evaluation of the interference modeling is divided in two parts. First, we assess the performance of the GMM model with different interference characteristics. For this, traces from all campaigns were used (i.e., channel 23 in FIRST, location 3 in SECOND, and first week from THIRD). Second, we compare to the state-of-the-art, a Pareto model [3], and with our previous approach based on an MMPP(2) model [4]. Traces from the first week of the THIRD campaign were used.

We quantitatively evaluate the *accuracy* and FPR of the estimated interference trace w.r.t. the ground truth trace. The output of the GMM model is a trace characterized in terms of mean IAT and number of signal arrivals per slot. To perform our comparison, mean IAT and number of signal arrivals of both traces (estimated and ground truth) were translated into a channel state, BUSY and FREE, using the TH_{IAT} and TH_{count} thresholds during each time slot. From this, the confusion matrix of the two channel state sequence is derived along with the metrics. The results are shown in Fig. 7. One can see that in OFFICE, during the 24 hours of the FIRST and SECOND campaign, Fig. 7a and Fig. 7b, the accuracy of the interference estimation is high, except from 10AM to 3PM in location 3 when the accuracy decreases as low as 82.8% and FPR increases up to 43.4%. We argue that this behavior is induced by the increase in the number of signal arrivals during those hours, as shown in Fig. 8. During the first week of the THIRD campaign, Fig. 7c and Fig. 7d, the accuracy of the estimation is high in both environments, over 98%. Moreover, it is evident that the GMM model can better estimate the behavior of the interference in OFFICE than HOME. This can be explained with arguments similar to those for the other campaigns. In HOME, as depicted in Fig. 9, the interference is more bursty. In OFFICE, the estimation accuracy is stable at 100%, except during the burstiest traffic hours, 9AM on Monday and 11AM on Friday, when the accuracy drops at 99.61% and 99.56% respectively in the two weeks. On the contrary, HOME exhibits much more frequent variations in accuracy and FPR, but the accuracy does not decrease below 98%.

We now show that our GMM modeling approach for estimating the interference provides more accurate estimates than the state-of-the-art approaches, Pareto and MMPP(2). Table III shows the results w.r.t. accuracy and FPR in predicting the actual interference for both environments in THIRD. For each different period in the life of the interference trace, we chose a two hours test trace. The three periods in Table III correspond to peak (*day*) and off-peak (*night*, *week-end*), the two off-peaks exhibiting different characteristics (i.e., different *NCLR* values).

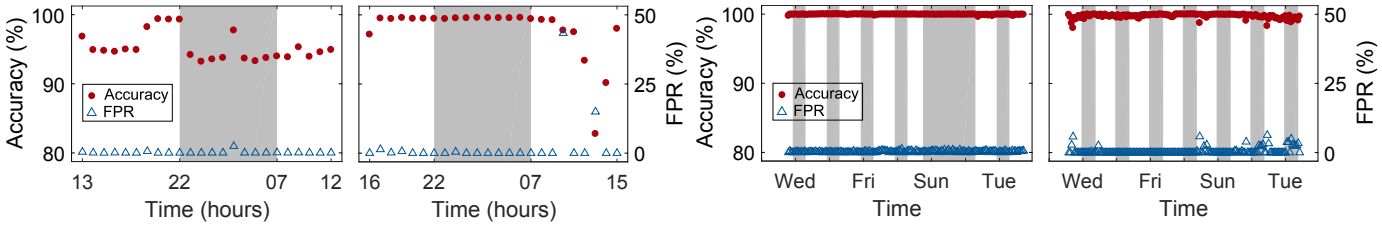
The Pareto-based approach relies on the self-similarity property of the traffic, meaning characteristics of the traffic are preserved irrespective of scaling in time. Therefore, to ensure a fair comparison with Pareto, we had to resort to at most a two-hour test trace in which the traffic exhibits self-similarity. The approach assesses the state of the channel upon the arrival of an application packet, therefore, the white space prediction probability is conditioned by this state.

Moreover, the performance of the MMPP(2) model depends on the training duration x and the modeling duration factor k. Here, we calibrated the MMPP(2) model for maximizing the AUC value, and used k = 1, x = 240, 180, 300 s in OFFICE and x = 420, 240, 540 s in HOME, for day, night and week-end.

Table III shows the results w.r.t. the ground truth in OFFICE and HOME. We note that the GMM approach achieves the best results, highest *accuracy* and lowest FPR, compared to the alternatives across all combinations of environments, channels, locations and time intervals. Moreover, GMM is slightly worse in HOME than OFFICE, due to the more bursty interference.

TABLE II: Confusion matrix.

		Predicted	
Actual	TN	FP	BUSY
	FN	TP	FREE
	BUSY	FREE	



(a) Channel 23 in FIRST. (b) Location 3 in SECOND. (c) First week OFFICE. (d) First week HOME.

Fig. 7: GMM model accuracy and FPR across campaigns.

Nevertheless, the *accuracy* does not decrease below 99.42% and the FPR is lower than 1.16%. On the other hand, through the lens of both metrics, Pareto performs better in HOME than OFFICE. Although Pareto's accuracy does not go over 32.41%, its FPR is low 5.99% during the busiest traffic periods in HOME, arguably due to its distribution, i.e., Pareto models high bursty traffic. Notably, MMPP(2) is better than Pareto in correctly identifying the two states of the channel, BUSY and FREE, translated into higher accuracy. The FREE state is better identified, translated into high FPR. High accuracy and high FPR in MMPP(2) can be attributed to high amount of free slots than BUSY slots in the trace and MMPP(2) mainly estimating the FREE slots.

D. White Space Prediction Evaluation

Next, we show that our GMM-HMM modeling approach provides accurate predictions of white spaces and compare it with two p-persistent channel access methods for $p = \{0.5, 1\}$ (i.e., transmission attempt is with probabilities 0.5 or 1 when the channel is sensed as idle), and with our previous approach based on the combination of HMM model with MMPP(2) [4].

Our performance metrics are *accuracy*, for the white space prediction model, and *Packet Loss Ratio* (*PLR*), for an application that uses the model to take decisions on when to transmit. The white space prediction works as follows: when a packet is ready to be transmitted, the GMM-HMM model produces a sequence of BUSY and FREE channel states, from which the closest 100 ms FREE slot is selected. Then, we compare the state of the chosen slot against the corresponding slot of the ground truth. A packet is marked as lost if the predicted state is FREE while the ground truth shows BUSY. Note that the GMM-HMM model predicts the closest FREE slot before the next scheduled packet transmission, and if the model is unable to predict a FREE slot, the current packet is considered as lost.

We report the HMM performance across the entire set of traces, and investigate the impact on the reliability of a staple data collection application ensuring variability in the transmission interval, 1 s and 60 s, representative for high and low data rate applications. For selecting the MMPP(2) parameters and the training duration for the MMPP(2)-HMM, we used the same procedure as in Section IV-C. Moreover, we selected a uniform distribution for 0.5-persistent method random numbers generation. Table IV shows the results w.r.t. the ground truth in both environments. First, across both environments and independent of the time of day/week, the *accuracy* places our approach ahead of the other methods and slightly (i.e., 3.06%) below the 1-persistent method in HOME during weekdays. In OFFICE, we are better than 1-persistent as we accurately predict both FREE and BUSY states, increasing *accuracy* and decreasing PLR. Looking at the Table IV, one can notice that *accuracy* and PLR, for both GMM-HMM and 1-persistent, sum up to 100%, because half of the confusion matrix (i.e., TNs and FNs) is always zero.

TABLE III: GMM vs. alternative solutions in THIRD.

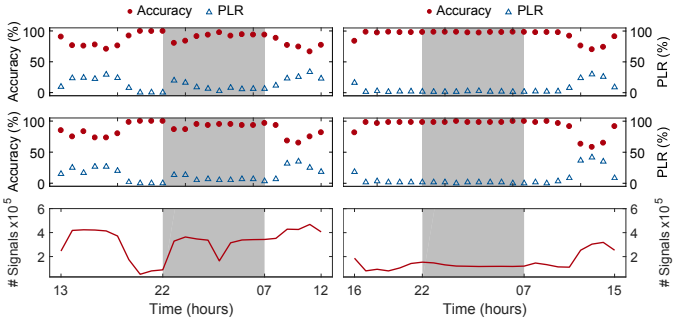
Period		Metric	OFFICE			HOME		
			GMM	MMPP	Pareto	GMM	MMPP	Pareto
Weekday	Day	Accuracy	99.82	86.86	16.71	99.42	68.94	32.41
		FPR	0.08	98.19	6.51	0.04	97.27	5.99
	Night	Accuracy	99.98	87.95	18.26	99.72	86.54	19.21
		FPR	0	99.34	8.28	1.16	99.53	8.15
Weekend		Accuracy	99.96	97.17	8.59	99.81	86.18	18.64
		FPR	0	98.86	8.53	0.39	99.32	7.21

Secondly, in OFFICE, our approach performs worst during the day, with an *accuracy* of 97.71%, due to the high traffic (i.e., low mean IAT, high number of signal arrivals). The accuracy decreases more as we progress from OFFICE to HOME, going as low as 89.44%. This can be explained with arguments similar to those for the GMM performance, the HOME exhibits more bursty interference and the model can not account for the short-term channel variations. Thirdly, through the lens of *PLR*, the environment induces different trends: in OFFICE, our method performs slightly better than all the others (i.e., maximum PLR is 2.29%), while in HOME it is the worst with *PLR* values as high as 10.56%. Our previous approach based on MMPP(2) model has low accuracy being unable to accurately capture the characteristics of the interference. Although 1-persistent random access method has a high *accuracy* in both environments, it fails in HOME, always predicting the channel as FREE while it is BUSY.

We now turn our attention to hourly variations of the GMM-HMM model *accuracy* and application PLR across all traces from our three campaigns, shown in Fig. 8 and Fig. 9. In the OFFICE environment, channel 23, location 3 and week 1, the off-peak GMM-HMM model has high accuracy ($\geq 97.1\%$) while the application has low PLR ($\leq 2.9\%$). This is also valid for the peak GMM-HMM model, except Thursday (minimum accuracy 56.7%) and Monday (minimum accuracy 66.7%) in week 1, and on channel 23 (minimum accuracy 65%) and location 3 (minimum accuracy 58.3%). When we look at HOME, the average accuracy, $88.3 \pm 16.8\%$, is lower than that of OFFICE, $98.3 \pm 1.4\%$. The off-peak GMM-HMM model in HOME performs in a similar way as that in OFFICE,

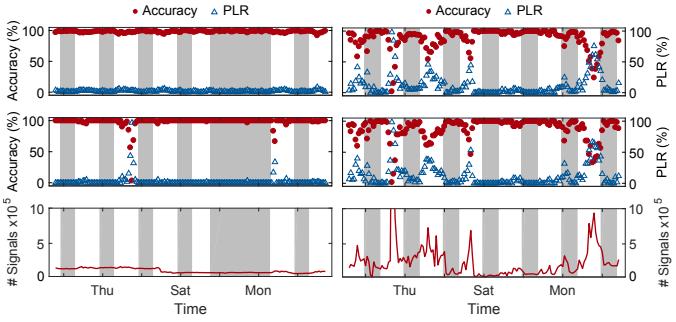
TABLE IV: Comparison of GMM-HMM prediction performance to alternative solutions.

Period		Metric	OFFICE				HOME			
			GMM-HMM	MMPP	0.5-p	1-p	GMM-HMM	MMPP	0.5-p	1-p
Weekday	Day	Accuracy	97.71	74.31	50.69	96.94	90.76	63.54	50.69	93.33
		PLR	2.29	3.19	3.01	3.06	9.24	6.67	6.58	6.68
	Night	Accuracy	99.72	65.07	50.83	99.58	89.44	75.00	49.17	92.50
		PLR	0.28	0.53	0.27	0.42	10.56	7.74	8.90	7.50
Weekend		Accuracy	100.0	87.92	50.76	99.51	92.29	67.36	51.32	89.51
		PLR	0	0.55	0.41	0.49	7.71	10.13	9.32	10.49



(a) Channel 23. (b) Location 3.

Fig. 8: GMM-HMM model performance prediction and the PLR of a 1 s (top) and 60 s (middle) data rate applications in FIRST and SECOND, hourly variations of interference (bottom).



(a) OFFICE. (b) HOME.

Fig. 9: GMM-HMM model performance prediction and the PLR of a 1 s (top) and 60 s (middle) data rate applications in THIRD, hourly variations of interference (bottom).

minimum accuracy 56.7%. However, the performance of the peak model in HOME is worse than that of OFFICE. These observations can be explained with the following arguments: i) the amount of hourly interference signal arrivals has an impact on the accuracy of the prediction, e.g. across all campaigns, the model exhibits high accuracy when the number of interference signals is $\leq 2 \times 10^5$; ii) variations of the interference signal arrivals has an impact on the accuracy of the prediction, e.g. sudden and dramatic variations induce a decrease in the accuracy; iii) sub-optimality of the HMM model, which is good in characterizing the distributions and channel state transitions but has difficulties in capturing the variations of the interference in time. The impact of the amount and the variations of the interference signals is more marked in HOME than in OFFICE. This translates into lower accuracy and higher PLR in HOME. Additionally, during the periods with the most severe interference, the number of available FREE slots is limited, going as low as 20% (i.e., limited transmission opportunities).

V. Discussion

The GMM-HMM model is a purely data-driven approach. In its current implementation, our choice of slot length of 100 ms is motivated by the trade-off between the accuracy of the model and the throughput requirements of the application that uses the model to take transmission decisions. However, we acknowledge that the throughput is only affected when the model predicts BUSY while the actual channel state is FREE. When training, the duration of our data trace was 1-hour. First, for durations shorter than 1 hour, the collected interference traces did not exhibit enough samples to properly capture both the FREE and BUSY states of the channel. Second, with few training samples it is difficult to retrain the full parameter set of both GMM and HMM models, and obtain a generic model for the testing data trace.

In environments with dramatic and rapid variations in the number of interference signals arrivals, two models (i.e., peak and off-peak) are not enough to capture this behavior. Therefore, to overcome this, the approach requires a continuous assessment of the channel condition.

Finally, a practical use of the GMM-HMM approach would be its integration with a MAC protocol where transmission decisions are taken by leveraging the white space prediction. Moreover, our GMM interference estimation model can be used for emulating radio interference in testbeds.

VI. Related Work

Detecting and classifying interference. Several works aim to measure, understand the impact of interference on low-power wireless networks, and classify interfering sources [8]–[13]. Musaloiu and Terzis [13], use $RSSI$ based features to quantify the interference on all IEEE802.15.4 channels to select the least interfered one. Noda et al. [8] compute the ratio of channel idle and busy time for assessing channel quality in the presence of interference. SpeckSense [9] classifies $RSSI$ bursts to characterize the channel as periodic, bursty or a combination of both. SoNIC [11] uses information from corrupted packets for interference source classification. These works succeed in detecting and identifying interference but it is not clear how these techniques are useful for autonomous interference mitigation due to interferers diversity. TIIM [10] makes a step further and extracts features from corrupted

packets to quantify the interference conditions instead of identifying the interferer. Thus the interference condition can be mapped to a specific mitigation technique. Nonetheless, an implementation of these mitigation techniques is not provided. CrossZig [12], the follow up work, contains an implementation of an adaptive packet recovery and FEC coding to address the problem. ART [14] proposes a probabilistic mechanism to adaptively use CSMA according to real-time interference level assessment done using packet delivery ratio (PDR) and fine-tuning the trade-off between throughput and PDR. All these solutions, however, are reactive, depending on the prevailing channel conditions, and do not aim to predict the white spaces through modeling, which is instead our goal in this paper.

Modeling interference. Creating lightweight models of interference is not a trivial task. Several researchers have proposed models for channel occupancy [3], [15]–[18] and for emulating interference caused by WiFi and Bluetooth [19]. A two-state semi-Markov model for channel occupancy is defined in [15], and exploited by each node to identify the less interfered channel and to switch accordingly. In comparison, we do not limit interference caused only by WiFi, but identify the white spaces for a specific channel through modeling interference in time domain. For modeling WiFi interference, Geirhofer et al. [17] propose a semi-Markov model and its continuous-time Markov chain, while Laganà et al. [18] enhance this model with a local view component. This model considers the limited detection range of sensor nodes and uses likelihood maximization and neural networks for estimating model's parameters. Boano et al. [16], [20] define a two-state semi-Markov model for channel occupancy and noise measurements are used to measure the duration of the FREE and BUSY instants, and compute their CDFs. Based on the longest BUSY period, MAC protocols' parameters are derived to meet the application requirements. JamLab [19] models and regenerates WiFi/Bluetooth/microwave interference patterns using sensor nodes, considering both saturated (always BUSY) and unsaturated traffic scenarios. A Markov chain model is used for saturated traffic and a probability mass function of empirical data for the non-saturated one. In contrast, our goal is not to emulate interference traffic but to estimate it, and for this we use a GMM to capture the ambient interference conditions.

The work from [3] is closely related to ours, focusing on a model-based prediction of the length of the immediate white space when a ZigBee frame is ready to be transmitted in the presence of WiFi interference. Depending on the length of the white space, the MAC frame is split in order to minimize collision probability. Nevertheless, continuous sampling of the operating channel is required as the model's parameters are calibrated whenever there is a frame to be transmitted. Moreover, their prediction is short-term in contrast to ours which is long-term and provides more information about when to transmit.

VII. Conclusions

In this work, we demonstrated that the combination of accurate interference estimation offered by a GMM model and reliable prediction of the wireless channel state enabled by an HMM model yields unprecedented accuracy in predicting transmission opportunities for low-power wireless networks. We validated our interference estimation approach against real-traces and state-of-the-art Pareto and MMPP(2) models, showing superior accuracy in all cases. The prediction mechanism was evaluated against 0.5- and 1-persistent random access methods and a MMPP(2)-HMM model. Results show that we can estimate interference with more than 94.7% accuracy in all scenarios, while an application using our GMM-HMM white space prediction for taking transmission decisions has less than 2.3% PLR under moderate interference and 10.5% PLR under heavy interference.

Acknowledgments

This work has been funded by the Irish Research Council in collaboration with United Technologies Research Centre, Cork, Ireland.

References

[1] C.-j. M. Liang *et al.*, "Surviving Wi-Fi Interference in Low Power ZigBee Networks," in *Proc. of SenSys*, 2010.
[2] A. Hithnawi *et al.*, "Understanding the Impact of Cross Technology Interference on IEEE 802.15.4," in *Proc. of WiNTECH*, 2014.
[3] J. Huang *et al.*, "Beyond Co-existence: Exploiting WiFi White Space for Zigbee Performance Assurance," in *Proc. of ICNP*, 2010.
[4] I. S. A. Dhanapala *et al.*, "Modeling WiFi Traffic for White Space Prediction in Wireless Sensor Networks," in *Proc. of LCN*, 2017.
[5] V.-B. Le, O. Mella, and D. Fohr, "Speaker Diarization using Normalized Cross Likelihood Ratio," in *Proc. of INTERSPEECH*, 2007.
[6] D. Reynolds, "Gaussian Mixture Models," *Ency. of Biometrics*, 2015.
[7] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, vol. 3, no. 1, Jan. 1986.
[8] C. Noda *et al.*, "Quantifying the Channel Quality for Interference-aware Wireless Sensor Networks," *ACM SIGBED Review*, vol. 8, no. 4, 2011.
[9] V. Iyer *et al.*, "Detecting and Avoiding Multiple Sources of Interference in the 2.4 GHz Spectrum," in *Proc. of EWSN*, 2015.
[10] A. Hithnawi *et al.*, "TIIM: Technology-Independent Interference Mitigation for Low-power Wireless Networks," in *Proc. of IPSN*, 2015.
[11] F. Hermans *et al.*, "SoNIC: Classifying Interference in 802.15.4 Sensor Networks," in *Proc. of IPSN*, 2013.
[12] A. Hithnawi *et al.*, "CrossZig: Combating Cross-Technology Interference in Low-power Wireless Networks," in *Proc. of IPSN*, 2016.
[13] R. Musaloiu-E and A. Terzis, "Minimising the Effect of WiFi Interference in 802.15.4 Wireless Sensor Networks," *Int. J. Sen. Netw.*, vol. 3, no. 1, Dec. 2008.
[14] F. Li *et al.*, "ART: Adaptive fRequency-Temporal Co-Existing of ZigBee and WiFi," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, Mar. 2017.
[15] L. Stabellini and J. Zander, "Energy Efficient Detection of Intermittent Interference in Wireless Sensor Networks," *Int. J. Sen. Netw.*, vol. 8, no. 1, Jul. 2010.
[16] C. A. Boano *et al.*, "JAG: Reliable and Predictable Wireless Agreement under External Radio Interference," in *Proc. of RTSS*, 2012.
[17] S. Geirhofer *et al.*, "Cognitive Medium Access: Constraining Interference Based on Experimental Models," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, Mar. 2008.
[18] M. Lagan *et al.*, "Modeling and Estimation of Partially Observed WLAN Activity for Cognitive WSNs," in *Proc. of WCNC*, 2012.
[19] C. A. Boano *et al.*, "JamLab: Augmenting Sensornet Testbeds with Realistic and Controlled Interference Generation," in *Proc. of IPSN*, 2011.
[20] ——, "Making Sensornet MAC Protocols Robust Against Interference," in *Proc. of EWSN*, 2010.