

# Base Station Localization in Search of Empty Spectrum Spaces in Cognitive Radio Networks

Yuan Zhang  
Computer Science and Technology Dept.  
Jilin University, China  
zhangyuan2u@gmail.com

Lichun Bao, Max Welling and Shih-Hsien Yang  
Computer Science Department  
University of California, Irvine  
{lbao,welling,shihhsy}@ics.uci.edu

**Abstract**—The radio spectrum in wireless communication systems is being allocated up quickly with ever increasing demands in wireless industries. Cognitive radio is a way of opportunistically sharing the scarce spectrum among primary and secondary users of the spectrum. The key challenge in deploying cognitive radio networks is to find out the spectrum holes in the primary wireless systems in order to allow the secondary users to operate. In this paper, we present novel localization algorithms based on grid-search and EM (expectation maximization) methods under the GMM (Gaussian mixture model) to find out the positions of the base stations and identify the spatial spectrum holes for cognitive radio deployments. Under this approach, we model the problem as localizing multiple unknown radio sources using a mobile measurement station, which is different from the problem modeling of many previous localization solutions. We evaluate our localization algorithms using both simulations and experiments.

**Keywords:** localization, expectation maximization (EM), Gaussian Mixture Model (GMM), path loss model.

## I. INTRODUCTION

RF (radio frequency) spectrum is increasingly expensive real estate for wireless communication systems. A well-known fact about the licensed RF spectrum is that the RF spectrum is uneven, sometimes wasteful, utilized from both spatial and temporal perspectives. Spatially, the spectrum is over-populated by mobile wireless systems in dense urban areas, while under-utilized in remote or suburban regions. Temporally, the spectrum is extraordinarily busy in certain time periods, while under-utilized otherwise, which is highly dependent on the purposes of the individual wireless systems and the behavioral model of mobile users. Examples include wireless systems primarily for voice communications, which are busy during day-time, but less used during night-time, and wireless broadcast systems for televisions, which behave conversely.

On the other hand, recent advancements in wireless ad hoc and sensor networks demand ever increasing amounts of RF resources for communication purposes. Practical wireless networks such as WLANs (wireless LANs), WMNs (wireless mesh networks), peer-to-peer MANETs, WSNs (wireless sensor networks) based on WiFi, WiMAX, ZigBee, Bluetooth are over-crowding the limited ISM (Industrial, Scientific, and Medical) bands. Therefore, the Federal Communication Commission (FCC) is searching for new oppor-

tunistic ways of utilizing the commercial and government-regulated radio spectrum using cognitive radio (CR) systems in order to improve the spectrum utilization efficiency [2].

The *key challenge* in cognitive radio is to find “spectrum holes” that are under-utilized in spatial and temporal domains so that the secondary users may opportunistically share the spectrum along with the legitimate primary spectrum holders [10], [18]. In this paper, we explore a base-station localization algorithm in order to search for spatial spectrum holes in the GSM system.

GSM (Global System for Mobile communications) is the most widely deployed systems in the world. GSM systems operate in 900 MHz, 1800 MHz or 2100 MHz bands, each consisting of two bands for uplink (mobile to base station) and downlink (base station to mobile) connections, respectively, for frequency division duplex communications.

In GSM 900 MHz systems, the bandwidth of both uplink and downlink bands is 25 MHz, which is a little wider than the DSSS-based WiFi bandwidth of 22MHz. Therefore, both GSM uplink and downlink bands are candidates for WiFi systems to operate on as secondary spectrum users. However, GSM mobile handsets are roaming clients, and could be potentially next to the WiFi wireless stations. Therefore, it is impossible for the WiFi system to use the GSM downlink band because WiFi systems could build high levels of noise to the close-by GSM mobiles.

Fortunately, it is feasible for the WiFi systems to use the GSM uplink band because the WiFi systems can be deployed far away from the GSM base stations so that the interferences from the WiFi systems are negligible to the GSM systems. This works because 1) the WiFi systems are to be deployed far away from the GSM base stations, 2) the transmission power in WiFi systems is limited by FCC so that a usual WLAN only spans up to a hundred meters, 3) wireless signals of WiFi system are sent using spread spectrum technologies, thus dissipating the energy over a much wider band than GSM signals. Note that the WiFi systems operating in the GSM uplink band do not affect the operations of GSM mobiles because the GSM mobiles only transmit in the uplink band.

In order to find the suitable locations for WiFi system deployments in the GSM uplink band, the key problem is to find out the locations of the GSM base stations.

In this paper, we propose a localization scheme that requires only a single mobile station with the minimum system knowledge about the GSM, and design a new probabilistic localization algorithm, called the LGMM (Localization estimation based Gaussian Mixture Model) algorithm. By the “minimum knowledge” assumption, we mean that only the received signal strength (RSS) values are collected by the mobile station in order to localize the GSM base stations, thus independent of any higher level knowledge, such as the number of base stations or their identifiers. The signal strength values are easy to collect using inverse FFT (fast Fourier transform) over the interested radio spectrum bands, and such an assumption results in more general localization algorithms in finding spatial spectrum holes over arbitrary radio bands in cognitive radio networks.

Basically, LGMM works on a series of RSS values of the GSM base stations, collected by a roaming mobile station along a certain path inside the interested areas. The mobile station is equipped with a GPS device, therefore able to tag each RSS value with the corresponding GPS location. By modeling the received GSM signals as Gaussian random variables coming from an unknown number of sources, LGMM calculates the number and locations of the GSM base stations using a two-step procedure. First, LGMM estimates the rough locations of the base stations using a grid-search method and the maximum likelihood estimation, which is called Grid-LGMM. Secondly, LGMM improves the location estimation accuracy using an EM (expectation maximization) method to refine the results of Grid-LGMM, which is called EM-LGMM.

In contrast to other localization works, LGMM formulates a new paradigm of localization problems, which require passive participation by targets, and need only one signal measurement station to collect RSS information. In addition, comparing with direct use of the EM method, LGMM avoids local minima by first finding the coarse estimates of the base station locations with Grid-LGMM. The localization results from both simulations and a real testbed experiment have shown the validity and accuracy of the proposed approach.

The rest of the paper is organized as follows. In Section II, we present a brief review of the related work. Section III describes our proposed probabilistic localization approach in detail. Section IV evaluates our proposed approach using both simulation and a real testbed experiment. Section V concludes the paper.

## II. RELATED WORK

The positions of wireless nodes can be used for many interesting applications in wireless mobile networks, such as location tracking, mobility management or location dependent service delivery. The simplest solution for network localization is to equip every wireless node with a Global Positioning System (GPS) device. However, when GPS information is unavailable due to either cost considerations

or the GPS signal reception limitations, a variety of solutions has been explored, requiring either none or very few reference coordinates in order to localize wireless stations in a wireless network. Current localization approaches can be categorized into two classes: range-free and range-based.

Range-free approaches use topological information to infer nodal locations, therefore saving any special hardware costs, and trading off the accuracy and scalability of the location estimates [11], [17]. He *et al.* [9] proposed APIT range-free localization scheme, which requires a heterogeneous network of sensing devices, some of which work as anchors with known location information. Then, the APIT divides the network area into triangular regions to gradually narrow down the locations of nodes. In [14], [3], a localization algorithm based on hop count, called DV-HOP, was presented, assuming that the wireless network consists of sensing nodes and location-known anchors. In order to localize the nodes, the anchors flood their location information along with incremental hop-distance information to the corresponding anchors. Nodes calculate their relative locations based on the received anchor locations, the hop-count and the average-distance per hop.

Range-based approaches are more widely studied in the literature, which derive the position of the unknown nodes using range estimates from location-known anchors, such as RSS, angle of arrival (AOA), time of arrival (TOA), and/or time difference of arrival (TDOA). In general, accurate range measurements require special hardware, sometimes expensive. Therefore less accurate but easily available RSS-based measurements are used extensively in many localization algorithms under probabilistic models.

RADAR [4] is a range-based indoor localization system based on RSS. It measures RSS at all positions in the entire building and records the RSS into a database during the calibration phase. Then, the system determines the location and orientation of a node by finding the best match of a set of RSS measurements in the database. NMDS-MLE [20] applies a maximum likelihood estimation (MLE) method to compute the distances between neighboring nodes based on their RSS. In [8], Niculescu *et al.* proposed APS (Ad hoc Positioning System) using AOA. The method derives the positions of all nodes in an ad hoc network with a small number of anchors. Caballero *et al.* addressed on the WSN localization problem in outdoor environments by using a mobile robot equipped with a GPS device to exchange its location information with wireless sensors [6].

Location estimation methods based on signal propagation models have received wide attention. Peng *et al.* presented a distributed, RSS-based probabilistic approach for outdoor wireless sensor network localization, accounting for inaccurate range measurements [15]. For the purpose of estimating multiple transmitter locations, Nelson *et al.* proposed quasi expectation maximization (EM) algorithm under a log-normal shadowing signal propagation model [13]. For

real time target tracking, Ding *et al.* presented a Gaussian mixture model based approach to capture the spatial characteristics of the target signal in a sensor network, and proposed a mean-shift continuous optimization method for target localization [7].

### III. PROBABILISTIC LOCALIZATION USING GAUSSIAN MIXTURE MODEL

#### A. Network Assumptions

We assume that in a GSM network, the GSM base station uses a well-known channel to periodically broadcast beacon messages. In order to collect the received signal strength values of these base stations, we deploy a GPS-equipped mobile station along a certain path in the area of interest. This way, whenever the mobile station receives the GSM beacon signal, it records the corresponding received signal strength as well as its current GPS location. Such signal strength plus location information are later used together for estimating two pieces of information, 1) the number of base stations, 2) the locations of the base stations.

In order to infer the locations of the wireless signal sources, we assume the models of the wireless signal path loss and the wireless channel using the log-distance path loss model, and the white Gaussian noise channel, respectively. In addition, because we do not know the number of wireless signal sources, we use a Gaussian mixture model (GMM) to capture the probabilistic nature of the signal sources.

#### B. Path Loss Model

The path loss model is useful in defining the relationship between RSS (received signal strength) values and the distance between transmitter and receiver. In this paper, we use a log-distance path loss model [16], as shown by Eq. (1).

$$r = t - l_0 - 10\gamma \log\left(\frac{d}{d_0}\right); \quad d \geq d_0 \quad (1)$$

in which,  $t$  and  $r$  is the transmit and receive signal power in dBm, respectively,  $d$  is the distance between the transmitter and receiver,  $d_0$  is the reference distance (e.g. 1 meter in our experiment),  $l_0$  means the path loss in dBm at  $d_0$ ,  $\gamma$  refers to the path loss exponent which depends on the channel characteristics and environments.

#### C. Gaussian Channel Model

We assume all interfering signals as additive white Gaussian noise background, and model the received wireless signal using the Gaussian distribution after applying the path loss model [16].

The Gaussian distribution is usually described by its probability density function, as shown by Eq. (2).

$$g(r; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(r - \mu)^2}{2\sigma^2}\right] \quad (2)$$

in which,  $r$  is the RSS value at mobile station modeled as a random variable,  $\mu$  is the mean value of the RSS, and  $\sigma$  is the standard deviation of the RSS.

#### D. Gaussian Mixture Model

In our problem definition, the mobile measurement station receives signals from multiple base stations. Therefore, each RSS measurement may come from any of the multiple base stations probabilistically. In order to capture such a fact, we have to assume that any RSS measurement is a mixture of probabilities that it comes from all base stations, with certain weight functions. That is, each RSS measurement is captured by a composite Gaussian mixture model (GMM), which was also adopted by several other solutions [5].

Therefore, given any RSS measurement  $r$  conditioned on the fact that the locations of the base stations are fixed, the probability of the signal coming from the mixture of multiple Gaussian sources is modeled as:

$$p(r) = \sum_{j=1}^M w_j g(r; \mu_j, \sigma_j) = \sum_{j=1}^M \frac{w_j}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(r - \mu_j)^2}{2\sigma_j^2}\right] \quad (3)$$

Eq. (3) describes the probability of RSS  $r$  in the mixture of  $M$  Gaussian components used in our approach, where  $g(\cdot)$  is the Gaussian probability density function,  $M$  is the number of base stations,  $\mu_j$  and  $\sigma_j$  are the mean and standard deviation of the RSS measurement  $r$  derived from the path loss model using Eq. (1) with the current location of the  $j$ -th base station,  $w_j$  is the weight of the  $j$ -th Gaussian component, satisfying  $\sum_{j=1}^M w_j = 1$ ,  $w_j \geq 0$ .

In practice, we collect a large amount of such RSS measurements, and assume that the RSS measurements are mutually independent. Therefore, given any RSS measurement sequence  $R = \{r_1, r_2, \dots, r_n\}$  conditioned on the fact that the locations of the base stations are fixed, the probability of the RSS measurement sequence  $R$  is

$$\begin{aligned} p(R) &= \prod_{i=1}^n \sum_{j=1}^M w_{ij} g(r_i; \mu_{ij}, \sigma_{ij}) \\ &= \prod_{i=1}^n \sum_{j=1}^M \frac{w_{ij}}{\sqrt{2\pi}\sigma_{ij}} \exp\left[-\frac{(r_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right] \end{aligned} \quad (4)$$

in which  $n$  is the number of RSS measurements. It is interesting to see that each RSS measurement  $r_i$  introduces one Gaussian mixture model with  $M$  sources. Under this model, our goal is to find out the true locations of  $M$  base stations such that the probability  $p(R)$  of the GMM is maximized.

#### E. Maximum Likelihood Estimation

Eq. (4) is the basis for searching the best possible locations of the multiple base stations, which can maximize the probability  $p(R)$ . For convenience, we equivalently derive the logarithm of  $p(R)$ , and calculate the log-likelihood function of Eq. (4), as given by Eq. (5).

$$\log(p(R)) = \sum_{i=1}^n \log \sum_{j=1}^M \frac{w_{ij} \cdot \exp\left[-\frac{1}{2}\left(\frac{r_i - \mu_{ij}}{\sigma_{ij}}\right)^2\right]}{\sqrt{2\pi}\sigma_{ij}} \quad (5)$$

We specify the algorithms to calculate each of these parameters. For convenience, we use symbol  $i$  to indicate the  $i$ -th RSS measurement of the series  $R = \{r_1, r_2, \dots, r_n\}$ , and  $j$  to indicate the  $j$ -th Gaussian component of the  $M$  Gaussian mixture.

In both Eqs. (4) and (5), there are three key parameters:  $w_{ij}$ ,  $\mu_{ij}$ , and  $\sigma_{ij}$ , which are the  $j$ -th Gaussian component weight, the expected value and the standard deviation of each RSS measurement  $r_i$  in  $j$ 's Gaussian component, respectively.

The weight value  $w_{ij}$  of each signal measurement  $r_i$  depends on the Cartesian distance  $d_{ij}$  between the base station  $j$  and the mobile station using their respective coordinates when collecting the  $i$ -th RSS measurement, which can be calculated by

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Here  $(x_j, y_j)$  is the current coordinate of base station  $j$ ,  $(x_i, y_i)$  is the coordinate of the mobile station when collecting the  $i$ -th RSS measurement. The formula of computing the weight value is

$$w_{ij} = \frac{e^{-d_{ij}}}{\sum_{k=1}^M e^{-d_{ik}}}$$

which is normalized over  $M$  base stations. The heuristic in weight calculation for each Gaussian component is that we place more trust in receiving RSS measurement from closer base stations than from farther ones, thus enforcing a myopic policy to filter the RSS data. That is, the closer the current distance  $d_{ij}$  between the base station and the mobile station, the more weight the corresponding Gaussian component has. Another alternative heuristic is to place equal weights on all Gaussian components. However, such policy may fail because distance information does affect the RSS data collection.

Parameter  $\mu_{ij}$  is the expected value of the  $i$ -th RSS measurement, which can be computed by Eq. (1) using the distance between each base station  $j$  and the mobile station at each RSS measurement point  $i$ . That is,

$$\mu_{ij} = t - l_0 - 10\gamma \log(\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}) \quad (6)$$

Parameter  $\sigma_{ij}$  is the standard deviation of the Gaussian model. For convenience, we set  $\sigma_{ij} = b \cdot \mu_{ij}$ , where  $b < 1$  is a constant, and we set  $b = 0.5$  in LGMM.

#### F. Bayesian Information Criterion (BIC)

The selection of the number of components from the data is an important computational issue for the maximum likelihood estimation using GMM. In fact, the more Gaussian components we add into the maximum likelihood estimation, the better the likelihood of the GMM. Therefore, we need to introduce a penalty for introducing too many Gaussian components. For this purpose, the Bayesian information criterion (BIC) is commonly used to select model parameters [19], [7].

Given  $k$  as the number of free parameters to be estimated and  $R$  as the data, Eq. (7) defines the BIC.

$$BIC = 2 \max \log p(R|k) - k \log(n) \quad (7)$$

in which  $\max \log p(R|k)$  is the maximum likelihood of the data  $R$  given the number of parameters  $k$ , and  $n$  is the number of data samples.

In our problem description, the parameters to be estimated are the two-dimension coordinates of  $M$  base stations. Thus, the number of parameters  $k = 2M$ . To use BIC for model selection, we simply choose the model that leads to the maximum BIC.

#### G. Base Station Location Estimation Using Grids

So far, we have prepared all the mathematical background for estimating the locations of the base stations using the RSS measurements collected by the mobile station.

In order to derive the location of the base stations, we take a two-step approach called the LGMM algorithm.

- 1) Iterative searching of the base stations' locations using a grid (the Grid-LGMM algorithm).
- 2) Iterative refinements of the base stations' locations using the EM-method (the EM-LGMM algorithm).

In this section, we specify the first part of the algorithms to iteratively search for the locations of the base stations along a grid structure on the search area.

In addition, because we do not know the number of base stations in the area, we devise a gradual approach to incrementally add base stations to the problem set until the number of base stations yields the BIC condition on the collected RSS measurements.

First, with the collection of the RSS measurements and their corresponding location information, we can derive the possible base station location area by finding the boundaries. Simply put, the boundaries of the fixed area is defined by a rectangle with  $(x_{min} - R_{tx}, y_{min} - R_{tx})$  and  $(x_{max} + R_{tx}, y_{max} + R_{tx})$  as the lower-left and upper-right corners' coordinates, and  $x_{min}$ ,  $y_{min}$  and  $x_{max}$ ,  $y_{max}$  are the minimum and maximum  $x$  and  $y$  coordinates of the data samples, respectively.  $R_{tx}$  is the communication radius of the mobile station.

Given the area definition, we draw a grid structure on the area. The edge length of each lattice in the grid structure depends on the accuracy and computational complexity of our base station localization algorithm. For instance, because GSM cell sizes ranges from a couple of hundred meters to dozens of kilometers, the size of each lattice can be chosen at around 20 meters in order to roughly estimate the base station locations. Note that the location estimates derived from grid search algorithm will be further refined using the EM-method in our next step.

Using the grid structure, the Grid-LGMM algorithm works as follows. It starts with one base station, and adds one more base station in the next round. In each step when

Table I  
NOTATIONS IN ALGORITHM 1

Notation	Meaning
$n$	The number of the RSS measurements.
$R$	An array of the RSS measurements, each with the corresponding coordinates.
$M$	The estimated number of the base stations.
$\text{BSLoc}$	An array of the estimated locations of the base stations.

a new base station is added to the grid, Grid-LGMM first tries to fit the new base station to all possible grid points, and finds its best possible location according to the  $BIC$  value. Then Grid-LGMM iteratively readjusts the previously added base stations to see if the new base station would find better locations for the existing base stations according to the  $BIC$ , until no moves can be made to any of the base stations. After each step, Grid-LGMM registers the  $BIC$  value, calculated by Eq. (7). After certain number of steps, Grid-LGMM could find the best number of base stations that maximizes the registered  $BIC$  value, as well as their locations.

We specify in details of the Grid-LGMM algorithm in Algorithm 1, of which the notation is presented in Table I.

In Algorithm 1, lines 1-3 initialize the base station location array  $\text{BSLoc}$  and the target value  $BIC$  with unlikely and minimum values, respectively. Lines 4-29 searches for the best number and the locations of base stations that produce the maximum BIC, satisfying the BIC condition. For that purpose, we search for the value of  $M$  starting from 1 on line 4. At each step, we place the location storage of the newly added base station at the beginning of the array  $\text{BSLoc}$ , and start grid-searching for the best locations of the base stations one by one on lines 11-20. If grid-search yields better target value  $BIC$ , then the grid-search continues on lines 14-18. Note that on line 17, by setting  $j = 0$ , we restart adjusting the base station location from the first element of  $\text{BSLoc}$  using the C-style **for** loop end operation  $j = j + 1$ .

However, when the target value  $BIC$  starts to drop, we know that the best estimate of the selected model has been achieved, and on lines 21-28 we stop the grid-search, and return the previously found best values for  $M$  and  $\text{BSLoc}$ .

Note that the function  $\text{findBIC}(R, j)$  is unspecified here, but has been described in Eq. (7). It is a simple operation by moving base station  $j$  to all the grid points and find the point that provides the best  $BIC$  estimate under current base station number  $M$ .

#### H. Refining Location Estimation Using the EM-Method

Grid-LGMM finds the coarse locations of the base stations placed on the grid points. We further refine the locations by edging in to the true locations using the EM-method (expectation maximization), which is called EM-LGMM.

---

#### Algorithm 1: Grid-LGMM

---

```

Input:  $R$ 
Output:  $M, \text{BSLoc}$ 
1 // Initialize.
2 foreach  $j = \{1, 2, \dots, \infty\}$  do  $\text{BSLoc}[j] = \{0,0\}$ ;
3  $BIC = -\infty$ ;
4 for ( $M=1$ ;  $M < \infty$ ;  $M = M + 1$ ) do
5   // Save the old BIC value.
6    $\text{oldBIC} = BIC$ ;
7   // Make space for the new base station location.
8   for ( $j = M - 1$ ;  $j > 0$ ;  $j = j - 1$ ) do
9      $\text{BSLoc}[j + 1] = \text{BSLoc}[j]$ ;
10  end
11  for  $j = 1$ ;  $j \leq M$ ;  $j = j + 1$  do
12    // Grid search  $j$ 's location.
13     $BIC = \text{findBIC}(R, j)$ ;
14    if ( $\text{oldBIC} < BIC$ ) then
15       $\text{oldBIC} = BIC$ ;
16      // Re-grid-search all base stations'
17      // locations.
18      if  $j \equiv M$  then  $j = 0$ ;
19    end
20    else break; // No more adjustments.
21  end
22  if  $\text{oldBIC} > BIC$  then
23    // BIC condition exceeded.
24     $M = M - 1$ ;
25    for ( $j = 1$ ;  $j \leq M$ ;  $j = j + 1$ ) do
26       $\text{BSLoc}[j] = \text{BSLoc}[j + 1]$ ;
27    end
28    return  $M, \text{BSLoc}$ ;
29 end

```

---

We specify EM-LGMM using the formulas only, which are needed in the iterative EM-method.

For simplify our presentations, we represent the location coordinates  $\{x_j, y_j\}$  of a base station  $j \in \{1, 2, \dots, M\}$  with  $\theta_j$ , i.e.

$$\theta_j = (x_j, y_j), \text{ and } \theta = \langle \theta_1, \theta_2, \dots, \theta_M \rangle,$$

and replace the log-likelihood estimation  $\log p(R|\theta)$  with  $L$ , i.e.

$$L = \log p(R|\theta).$$

In EM-method, we gradually improve the location estimates  $\theta^t$ ,  $t = 1, 2, \dots$ , by iterating through the E-step and the M-step, starting with an initial condition

$$\theta^1 = ((x_1, y_1), \dots, (x_j, y_j), \dots, (x_M, y_M))^T,$$

in which  $T$  indicates the transpose operation.

**E-step:** Herein, we introduce a new notation to denote posterior probability  $\beta_j^t(r_i)$  that the  $j$ -th component generated  $r_i$  using the estimation of the parameters from the M-step, which is shown in Eq. (8).

$$\beta_j^t(r_i) = \frac{w_{ij} \cdot g(r_i | x_j^{t-1}, y_j^{t-1})}{\sum_{j=1}^M w_{ij} \cdot g(r_i | x_j^{t-1}, y_j^{t-1})} \quad (8)$$

**M-step:** Compute a new set of parameters using  $\beta_j^t(r_i)$  and gradient descend method as shown in Eqs. (9) and (11).

$$x_j^t = x_j^{t-1} + \eta \frac{\partial L}{\partial x_j^{t-1}} \quad (9)$$

in which symbol  $\eta$  represents the step size in the numeric calculations, and  $\frac{\partial L}{\partial x_j^{t-1}}$  is calculated by Eq (10). For clarity, we have removed the step indicator  $t - 1$  in the more complicated parts.

$$\begin{aligned} \frac{\partial L}{\partial x_j^{t-1}} &= \frac{1}{3} \left( \frac{\partial L}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial x_j} + \frac{\partial L}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial x_j} + \frac{\partial L}{\partial w_{ij}} \frac{\partial w_{ij}}{\partial x_j} \right) \\ &= \frac{1}{3} \sum_{i=1}^N \beta_j(r_i) \left[ \frac{-10\gamma(r_i - \mu_{ij})(x_j - x_i)}{\sigma_{ij}^2 [(x_j - x_i)^2 + (y_j - y_i)^2]} \right. \\ &\quad \left. + \frac{-10\gamma b(x_j - x_i)}{(x_j - x_i)^2 + (y_j - y_i)^2} \left( \frac{(r_i - \mu_{ij})^2}{\sigma_{ij}^3} - \frac{1}{\sigma_{ij}} \right) \right. \\ &\quad \left. + \frac{a(x_j - x_i)e^{-d_{ij}}}{w_{ij} \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}} \right] \end{aligned} \quad (10)$$

$y_j^t$  is estimated similarly. For convenience, we provided the calculation formulas as follows.

$$y_j^t = y_j^{t-1} + \eta \frac{\partial L}{\partial y_j^{t-1}} \quad (11)$$

$$\begin{aligned} \frac{\partial L}{\partial y_j^{t-1}} &= \frac{1}{3} \left( \frac{\partial L}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial y_j} + \frac{\partial L}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial y_j} + \frac{\partial L}{\partial w_{ij}} \frac{\partial w_{ij}}{\partial y_j} \right) \\ &= \frac{1}{3} \sum_{i=1}^N \beta_j(r_i) \left[ \frac{-10\gamma(r_i - \mu_{ij})(y_j - y_i)}{\sigma_{ij}^2 [(x_j - x_i)^2 + (y_j - y_i)^2]} \right. \\ &\quad \left. + \frac{-10\gamma b(y_j - y_i)}{(x_j - x_i)^2 + (y_j - y_i)^2} \left( \frac{(r_i - \mu_{ij})^2}{\sigma_{ij}^3} - \frac{1}{\sigma_{ij}} \right) \right. \\ &\quad \left. + \frac{a(y_j - y_i)e^{-d_{ij}}}{w_{ij} \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}} \right] \end{aligned} \quad (12)$$

According to the EM-method, the log-likelihood function  $L$  monotonically increases after each iteration, and the EM-method is guaranteed to converge to a local maximum of the log-likelihood function [5], [12]. Thus, the iterations will stop when the increment of  $L$  drops below a certain threshold  $\varepsilon$  in Eq. (13), in which  $\varepsilon$  is a small number. In our experiment, we set  $\varepsilon = 0.5\%$ .

$$\left\| \frac{L(R|\theta^t) - L(R|\theta^{t-1})}{L(R|\theta^t)} \right\| \leq \varepsilon \quad (13)$$

#### IV. EVALUATIONS

Our algorithms were implemented in Matlab 7.0 to estimate the number and locations of the base stations, and the data for evaluations were collected and validated through two sets of experiments, one based on simulations and the other on real testbed experiments.

In the simulations, we use NCTUns v5.0 [1] to set up a GSM network over a  $500m \times 1200m$  rectangular area with 8 base stations, in which a mobile station follows the path indicated in the figure to collect periodic beacon messages from the base stations, as shown in Fig. 4. The distance between every two of the base stations is more than 200 meters. The communication radius of the mobile station is 500 meters. The transmission power of base station is 5dbm.

Because it is a simulated environment, the collected signals are actually not random variables. However, we forfeit such knowledge, and treat them as random variables so as to maintain the correctness of our algorithms.

In the real testbed experiments, we collected the AP (access point) signals in a WLAN system, instead of the original GSM signals. This is because GSM signal collection requires specialized spectrum analyzing equipments operating under the licensed frequency, thus more difficult. However, because the mathematical tools presented in our paper similarly applied to WLAN system signals, we could still evaluate the correctness and effectiveness of our algorithms in such environments.

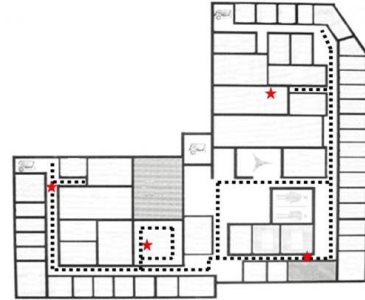


Figure 1. Testbed experiment scenario.

In our testbed, we deployed four real APs, with the same transmission power (22dBm) on our office floor as represented by stars in Fig. 1. A ThinkPad R61i notebook with Intel(R) Wireless WiFi Link 4956AG is used to collect the RSS values by walking along the dotted line. The communication radius is 30 meters.

The performance of the algorithms is measured by the number of base stations and the average distance error of the location estimations relative the actual base station locations.

#### A. Simulation Results

1) *Computation process:* In order to apply Grid-LGMM, we draw the grid structure of the square lattice with an edge length of 20 meters.

Fig. 2 shows the BIC in Grid-LGMM as we gradually increase the number of base stations. Using the *BIC* condition, we can easily identify the optimum number of base stations, which appears to be 8, matching the simulated number of base stations.

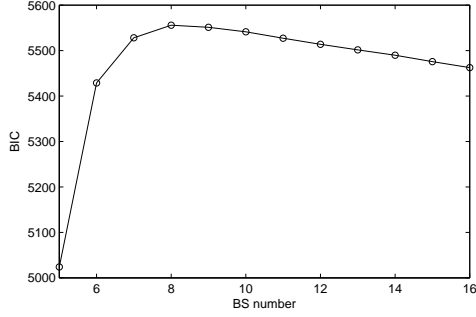


Figure 2. BIC for different number of base stations.

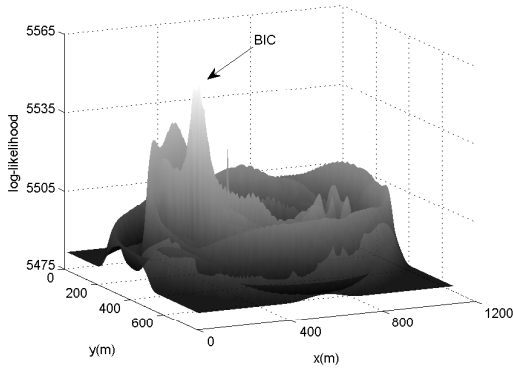


Figure 3. BIC when the number of base stations=8.

Specifically, supposing that the Grid-LGMM is done to find the optimum location of seven base stations over the collected RSS series  $R$ , Fig. 3 illustrates the log-likelihood with the BIC penalty using grid-search mechanism when a new eighth base station is added. The best location of the eighth base station is thus indicated in the figure.

2) *Performance of LGMM*: As we know, LGMM consists of two sequential steps, Grid-LGMM and EM-LGMM. Fig. 4 (a) and (b) illustrate the location estimation accuracy in the two steps. In this set of simulations, the lattice size is set at  $20m \times 20m$ .

Fig. 5 presents the location errors between the actual and the estimated locations of the base stations in the two steps of LGMM. In our calculations, the EM-LGMM algorithm converges after 286 iterations with threshold  $\varepsilon = 0.5\%$ .

In addition, by using a coarse grain location estimation globally, we avoid being trapped in local minima while searching for the global optimum. The presented algorithms can correctly and accurately find the base stations. As shown in Fig. 5, the location errors of each base stations reduce from 11.1064 meters to 3.8240 meters.

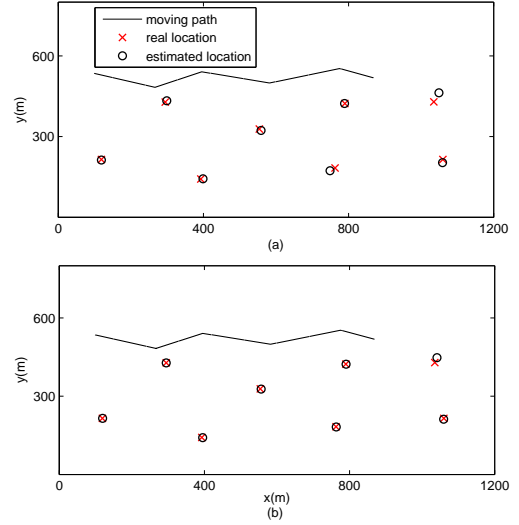


Figure 4. Location estimation errors in Grid-LGMM and EM-LGMM, respectively.

	BS number		average distance error(m)
	true	estimated	
Grid-LGMM	8	8	11.1064
EM-LGMM	8	8	3.8240

Figure 5. Numeric location estimation errors in Grid-LGMM and EM-LGMM.

## B. Testbed Experiment Results

In order to apply the Grid-LGMM algorithm in WLAN environments, we set the lattice size as  $2m \times 2m$ .

Fig. 6 illustrates the location estimation results of both Grid-LGMM and EM-LGMM algorithms. However, because real world signal propagation does not show nice statistical features like those of Gaussian models, LGMM estimated that 6 APs exist in the network, four of which match with the real APs. Fig. 6(a) shows that the Grid-LGMM algorithm finds the coarse estimates of the APs, and Fig. 6(b) shows that the EM-LGMM algorithm provides better location estimates, closer to the real AP locations.

Fig. 7 shows the numeric location estimation errors of the two steps in details. The spurious locations were presumably estimated because the collected RSS data in real environments do not strictly follow the normal distribution, and that data are more noisy due to the structural influence on the RSS data.

Overall, the results of these experimental evaluations show that our proposed approach is valid and perform fairly reasonably in real indoor environments.

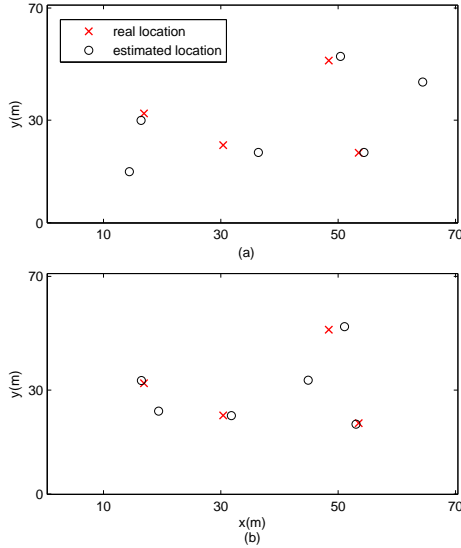


Figure 6. Estimation results: (a) Grid-LGMM; (b) EM-LGMM.

	AP number		average distance error(m)
	true	estimated	
Grid-LGMM	4	6	2.9932
EM-LGMM	4	6	1.4331

Figure 7. Estimation results of testbed experiment.

## V. CONCLUSION

We have presented LGMM (localization based on Gaussian mixture model), a probabilistic algorithm for estimating the number and locations of base stations in GSM networks using received radio signal strength (RSS) measurements, collected by a GPS-equipped mobile station. LGMM consists of two steps, of which Grid-LGMM provides coarse locations of the base stations using the maximum likelihood estimation, and EM-LGMM improves the location estimation accuracy using the EM-method. The results of simulations and the real testbed experiments show that LGMM produces satisfactory location estimates and the number of base stations in the network. Overall, LGMM distinguishes itself on two aspects: 1) it uses a single mobile station and only the RSS measurements to derive base station locations, 2) it does not assume the number of base stations, which is derived instead of given in the calculations.

## ACKNOWLEDGMENT

This work was sponsored in part by the National Science Foundation (NSF) under grant No. 0725914, and the China Scholarship Council under Grant No. B2007101731.

## REFERENCES

- [1] NCTUns 5.0 Network Simulator and Emulator <http://nsl.csie.nctu.edu.tw/nctuns.html>.
- [2] FCC. ET Docket No. 03-108. Facilitating Opportunities for Flexible and Efficient and Reliable Spectrum Use Employing Cognitive Radio Technologies, Mar. 2005.
- [3] T.A. Alhmiedat and S.H. Yang. A Survey: Localization and Tracking Mobile Targets through Wireless Sensors Network. In *The 8th PostGraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNET)*, 2007.
- [4] P. Bahl and V. Padmanabhan. RADAR: An In-Building RF-based User Location and Tracking System. In *INFOCOM*, 2000.
- [5] J.A. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Tech Report ICSI-TR-97-021, 1998.
- [6] F. Caballeroa, L. Merinob, P. Gila, I. Mazaa, and A. Ollero. A probabilistic framework for entire WSN localization using a mobile robot. *Robotics and Autonomous Systems*, 56:798–806, Oct. 2008.
- [7] M. Ding and X. Cheng. Fault Tolerant Target Tracking in Sensor Networks. In *MOBIHOC*, 2009.
- [8] D. Niculescu and B. Nath. Ad Hoc Positioning System(APS) using AOA. In *INFOCOM*, 2003.
- [9] T. He, J. A. Stankovic, C. Huang, T. Abdelzaher, and B. M. Blum. Range-Free Localization Schemes for Large Scale Sensor Networks. In *MOBICOM*, pp 81–95, 2003.
- [10] P. Kolodzy. Next generation communications: Kickoff meeting. Proc. DARPA, Oct. 17 2001.
- [11] M. Li and Y. Liu. Rendered path: range-free localization in anisotropic sensor networks with holes. In *MOBICOM*, pp 51–62, 2007.
- [12] G. McLachlan and D. Peel. *Finite Mixture Models*. New York: John Wiley and Sons, 2000.
- [13] J. K. Nelson, M. R. Gupta, J. E. Almodovar, and W. H. Mortensen. DV Based Positioning in Ad hoc Networks. *Signal Processing Letters*, 16:354–357, May. 2009.
- [14] D. Niculescu and B. Nath. DV Based Positioning in Ad hoc Networks. *Journal of Telecommunication Systems*, 2003.
- [15] R. Peng and M.L. Sichitiu. Probabilistic localization for outdoor wireless sensor networks. In *ACM SIGMOBILE Mobile Computing and Communications Review*, volume 11, pages 53–64, 2007.
- [16] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Pearson Education, 2nd edition, 2001.
- [17] R.Want, A.Hopper, V.Falco, and J.Gibbons. The Active Badge Location System. In *ACM Transactions on Information Systems*, volume 10, pages 91–102, Jan. 1992.
- [18] A. Sahai, N. Hoven, and R. Tandra. Some Fundamental Limits on Cognitive Radio. In *Proc. of Allerton Conference*, Monticello, Oct. 2004.
- [19] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [20] D. Wu, L. Bao, M. Du, and R. Li. Design and Evaluation of Localization Protocols and Algorithms in Wireless Sensor Networks Using UWB. In *IPCCC*, pp 18–25, Dec. 2008.