Abstract—Accurate positioning of a moving vehicle along a route enables various applications such as travel time estimation in transportation. Global Positioning System (GPS) based localization algorithms suffer from low availability and high energy consumption. Received signal strength indicator (RSSI) measured in the course of the normal operation of Global System for Mobile Communications (GSM) based mobile phones on the other hand consumes minimal energy in addition to the standard cell phone operation but very low accuracy. In this paper, we incorporate the fact that the motion of vehicles satisfies route constraints to improve the accuracy of the RSSI based localization by using Hidden Markov Model (HMM) where the states are the road segments on the road and the observation at each state is the RSSI vector containing the detected power levels of the pilot signals sent by the associated and neighboring cellular base stations. In contrast to prior HMM based models, we train HMM based on the statistics of the average driver behavior on the road and the probabilistic distribution of the RSSI vectors observed in each road segment. We demonstrate that such training improves the accuracy of the localization considerably and provides localization performance robust over different road segment lengths by using extensive cellular data collected in Istanbul, Turkey; Berkeley, USA and New Delhi, India.

Index Terms—localization, mobile phone, fingerprinting, route constraints

I. INTRODUCTION

Accurate positioning of a moving vehicle along a route enables various applications such as travel time and pattern estimation in transportation. Travel time estimates are used to alleviate congestion by informing the drivers of the roads with large travel times, finding better paths with smaller expected travel time in traffic-aware routing algorithms and improving traffic operations such as traffic light cycle control. Travel pattern estimation in the form of origin-destination matrices on the other hand is an essential component in enabling administrative authorities to plan and manage infrastructure improvements. Using vehicles as probes to collect these data rather than deploying fixed sensors such as wireless magnetic sensors on the road [1] or using household questionnaires allows covering large areas of the road network efficiently.

Current proposals for accurate positioning heavily rely on using GPS enabled devices periodically recording the current time and position, and sending these data to a central server over a wireless network for further processing [2]. GPS however is not available in a large percentage of today’s mobile phones, especially in developing regions, leaving their users out of the many location-based applications. Moreover, GPS has poor performance in urban areas near high-rise buildings and consumes a lot of energy quickly draining the battery of mobile devices, which may not necessarily be connected to a charger within the vehicle. Other alternatives such as Wi-Fi [3] and augmented sensor [4] based localization in the cell phones suffer from the low availability similar to GPS. GSM on the other hand is available on all GSM based cell phones and consumes minimal energy in addition to the standard cell phone operation. Among the GSM based localization algorithms including time of arrival (TOA), time difference of arrival (TDOA), angle of arrival (AOA) and RSSI based methods, RSSI based localization algorithms are mostly preferred since RSSI is measured in the course of the normal operation of GSM-based mobile phones so does not require any specialized hardware [5].

According to GSM standards, each cell phone records a vector of the detected power levels of the pilot signals from at most seven cellular base stations one of which the cell phone is associated with. Positioning algorithms using this RSSI vector follow two approaches: triangulation and fingerprinting. In triangulation, first the RSSI reading from each base station is matched to a distance based on an experimentally determined path loss model then the position of the mobile phone is estimated based on the location of each base station and the distances to them. The accuracy of triangulation based approaches however has been shown to be very low due to the inaccuracy of the path loss models [6], [7]. In fingerprinting approach, RSSI measurements from the cellular base stations are recorded along with the GPS measurements during the database construction phase. The position is then determined by comparing the measured RSSI vector to those in the database identifying the GPS location of the closest RSSI matches determined by minimizing a cost function [8]. The accuracy of fingerprinting also has been shown to be very poor requiring stochastic post-processing such as Kalman filtering [9]. Recently proposed trajectory based methods using the RSSI readings of the neighboring base stations apply Hidden Markov Models (HMM) to improve the localization accuracy by taking into account route constraints however employing completely random behavior of the vehicles in each direction such as equal probability of moving to each of the adjacent cells in a grid-based map without any training [10], [11], [12]. None of the previously proposed HMM techniques developed for GSM RSSI based localization derive the statistics of the driver behavior on the road and analyze the effect of including such statistics on the localization accuracy. Furthermore, these techniques mostly use deterministic approaches to estimate the location of the mobile phones without analyzing the probabilistic distribution of the observed RSSI vectors on the road segments as a function of the amount of the training data nor comparing such a probabilistic model based approach to the commonly used cost functions in the literature [10], [11]. The only probabilistic fingerprinting approaches in the literature on the other hand either consider the associated base station only [12] or do not take the route constraints into account [13].
There has been extensive amount of work on RSSI based localization systems employing HMM using indoor WLAN (Wireless Local Area Networks) base stations [14], [15]. The movement models used for the indoor environments considering the building’s floor plan and expected pedestrian speeds however are not suitable for modeling the movement of the vehicles. HMM and RSSI based localization algorithms using outdoor WLAN base stations [3] and sensor networks [16] again cannot be used in GSM based vehicle localization since they assume completely random behavior of the nodes in the network and the propagation characteristics of the short range WLAN and sensor node communication are very different from those of the long range GSM communication.

In this paper, we propose an HMM based localization algorithm exploiting GSM RSSI data of the mobile phones located within the vehicles trained by the statistics of the average driver behavior on the road and the probabilistic distribution of the RSSI vectors in each road segment. The original contributions of this paper are as follows:

- We derive the statistics of the average driver behavior and the probabilistic distribution of the RSSI vectors in each road segment of urban and suburban roads in Istanbul, Turkey; Berkeley, USA and New Delhi, India; for the first time in the literature.
- We demonstrate that the proposed HMM trained by these statistics improves the accuracy of the previous HMM techniques significantly and provides a localization performance robust over different road segment lengths based on extensive cellular data collected in Istanbul, Turkey; Berkeley, USA and New Delhi, India.

II. SYSTEM OVERVIEW

A. Localization using HMM

We assume that the vehicle follows the same road. We divide the road into segments representing the hidden states of the HMM. Transitioning among these hidden states are governed by transition probabilities. Given the segment of the road, the RSSI vector has a particular conditional probability distribution called the emission probability distribution of the observables in HMM. The HMM then traverses its states to produce its outputs, which are the observables emitted at each state.

The algorithm has offline training and online tracking phases. In offline training phase, we collect GPS and RSSI vectors periodically recorded while the vehicle is moving on the road. These data are then transmitted via any available wireless network to the central server. The collected data from multiple users are used to estimate the transition probability and emission probability of the HMM. During the online tracking phase, a sequence of periodically recorded RSSI vectors are sent to the central server and input to the HMM to estimate the most probable sequence of states so road segments.

B. Mathematical Model of HMM

The HMM can be mathematically represented as \( \lambda = (S, V, A, B, \pi) \):

- \( S = \{ s_1, s_2, \ldots, s_N \} \) is the set of states, each state representing a road segment on the road, where \( N = |S| \) is the number of road segments on the road.
- \( V = \{ v_1, v_2, \ldots, v_M \} \) is the set of observation vectors, where \( M = |V| \) is the total number of possible observation vectors on the road. \( v_k \) is a \( Z \times 1 \) vector for \( k \in [1, M] \) such that the \( i \)-th element of \( v_k \) denoted by \( v_{k,i} \) is the received power from the \( i \)-th base station represented by an integer value in the range \([-110, -40]\) dBm if the \( i \)-th base station is one of the seven strongest base stations and 0 otherwise, where \( Z \) is the total number of basestations observed on the road.
- \( A \) is an \( N \times N \) matrix such that the element in the \( i \)-th row and \( j \)-th column of \( A \) denoted by \( a_{ij} \) is the probability of the transition from state \( s_i \) to state \( s_j \), i.e. \( a_{ij} = P(q_{n+1} = s_j | q_n = s_i) \) where \( q_n \) is the state at time \( n \tau \) and \( \tau \) is the sampling period. First, we assume that the vehicle follows the same road and the training of the HMM is performed by using the data collected from the vehicles following that same road. This assumption eliminates the requirement for considering the effect of the variation of the routes of different drivers to their destination. Second, we assume that the vehicles move with the same average speed in the same direction along the road used in the training and tracking phases meaning that the transition probability \( a_{ij} \) is independent of the segment \( i \) but only depends on the difference between \( i \) and \( j \), i.e. \( a_{ij} = a_k \) where \( k = j - i \) and \( a_k = 0 \) for \( k < 0 \). The matrix \( A \) needs to be derived for each road satisfying these conditions. This approach can be generalized to general vehicle tracking where different drivers have different routes to their destination by dividing the road into multiple segments where these assumptions hold. The first novelty of this paper is the derivation of \( a_k \) values based on the training data summarizing the statistics of the average driver behavior on the road. The previous HMM RSSI based GSM localization algorithms allow the transitions only to the current or following road segments, i.e. \( a_0 = 1/2, a_1 = 1/2 \) [10], [11], [12].
- \( B \) is an \( N \times M \) matrix such that the element in the \( j \)-th row and \( k \)-th column of \( B \) denoted by \( b_{jk} \) is the emission probability of \( v_k \) on the \( j \)-th road segment corresponding to state \( s_j \), i.e. \( b_{jk} = P(v_k | s_j) \). The emission probability distributions derived for WLAN systems [3], [14], [15] and sensor networks [16] cannot be used for GSM systems due to their different propagation characteristics. The HMM techniques developed for GSM RSSI based localization on the other hand use either a deterministic function of the Euclidean distance and number of matching base stations between the observed vector and each RSSI vector in the training database as the cost of the observed RSSI vector [11] or the histogram of the RSSI readings of the associated base station in each road segment [12]. The second novelty of this paper is proving that the histogram of the RSSI readings of each base station within each road segment converges to the Gaussian distribution if enough RSSI samples are collected.
- \( \pi \) is an \( N \times 1 \) vector such that the \( i \)-th element of \( \pi \) denoted by \( \pi_i \) is the probability of state \( s_i \), initially, i.e. \( \pi_i = P(q_1 = s_i) \).

III. DATA COLLECTION

We collected data over four different roads shown in Fig. 1. The first road of length 12 km between Sariyer and 4. Levent, Istanbul shown in Fig. 1 (a) represents an urban road in Istanbul. The second road of length 4 km between Koc University and Sariyer, Istanbul shown in Fig. 1 (b) represents a suburban road in Istanbul. The third road of length 8.2 km between University of California Berkeley and Temescal, USA shown in Fig. 1 (c) represents a suburban road in USA. The fourth road of length 2.9 km in New Delhi shown in Fig. 1 (d) represents a suburban road in India.
The monitoring device for collecting RSSI signatures was a standard Nokia handset, operating in the 900MHz band, connected to a laptop running network monitoring software. A sample of the RSSI signature containing the power levels and identifiers of the seven strongest base stations is reported once every two seconds along with the GPS latitude and longitude information, i.e., $\tau = 2$ sec. 20 test runs are performed at low, medium and high traffic conditions on the urban road in Istanbul shown in Fig. 1 (a) whereas 20 test runs are performed on the suburban roads shown in Fig. 1 (b), (c) and (d) since the traffic conditions do not change significantly during the day: 10 runs are used for training and 10 runs are used for tracking.

### IV. Estimation and Usage of HMM Statistics

#### A. Training Phase

During the training phase, GPS and RSSI vector data observed from the cell phones are collected at the central server to create the database. The parameters of the HMM are then estimated using this database.

Figs. 2 and 3 show the probability of the transition to the following road segments, $a_k$, for the urban road and suburban roads of different segment lengths respectively. $L$ denotes the length of the segment. The calculation of $a_k$ given the training database including the sequence of states recorded by GPS, i.e., $Q^r = (q^r_1, q^r_2, \ldots, q^r_{T^r})$ for $r \in [1, R]$ where $q^r_i$ is the $i$-th state in the $r$-th run, $R$ is the number of runs in the training data and $T^r_i$ is the number of data points corresponding to the $r$-th run in the training data, is given as

$$a_k = \frac{\sum_{r=1}^{R} \sum_{j=1}^{T^r_i-1} \delta_{q^r_{j+1}-q^r_{j}, k}}{\sum_{r=1}^{R} \sum_{j=1}^{T^r_i-1} 1} \quad (1)$$

where $\delta_{q^r_{j+1}-q^r_{j}, k}$ is the Kronecker delta function that takes value 1 if $q^r_{j+1} - q^r_{j} = k$ and 0 otherwise, the numerator is the number of transitions of magnitude $k$ and the denominator is the total number of transitions. We observe that the transition probability varies significantly depending on the location and traffic conditions of the roads contradicting the common assumption of the HMM-based localization literature that the mobiles move only to the adjacent grid cells or adjacent road segments.

The analysis of many histograms of the observed RSSI values from many base stations shows that the histogram has a Gaussian shape if enough RSSI samples are collected from the base station in the corresponding road segment. The number of RSSI samples required for the Kolmogorov-Smirnov test with a 95% confidence interval to decide for the Gaussian distribution of these RSSI samples is less than 30 in all the segments of different lengths over all the roads used for the evaluation (Kolmogorov-Smirnov test is a
nonparametric test for the equality of continuous one-dimensional probability distributions comparing a sample distribution with a reference probability distribution). The number of samples required to estimate the mean and variance of the Gaussian distribution of the RSSI samples with 90% accuracy is around 30. Given that the number of samples collected from the segment lengths greater than or equal to 10 m, which are used in the evaluation of the algorithm, in each run of the training data is more than 4, the total number of samples from 10 runs of the training data, i.e., 40 samples, will be enough to estimate the mean and variance of the Gaussian distribution with 90% accuracy.

Let us denote the mean and variance of the best Gaussian fit to the RSSI histogram of the i-th base station at the j-th road segment corresponding to state s_j by m_{j,i} and \sigma_{j,i}^2, respectively. Recall that b_{jk} is the emission probability of an observation vector v_k in state s_j. Let us define the cost of observing v_k in state s_j denoted by c_{jk} as the normalized logarithm of b_{jk}. Based on the assumption that the samples from different base stations are independent, c_{jk} is given by

\[ c_{jk} = \frac{1}{Z_p} \sum_{i=1, m_{j,i} < v_k, v_k < 0} \left( \frac{1}{2} \ln(2\pi) - \ln(\sigma_{j,i}) - \frac{(v_k - m_{j,i})^2}{2\sigma_{j,i}^2} \right) \]

(2)

where \(Z_p\) is the number of base stations for which both \(m_{j,i}\) and \(v_k\) are negative. This is very similar to the previously proposed cost functions defined as the distance between RSSI vectors as in [3], [11] except the weighting factor as a function of the variance, giving less weight to the observation with high variance.

The sampling period \(\tau\), the segment length \(L\), the road and traffic conditions on the road affect the values of both the transition and emission probabilities but the algorithm works as long as the same \(\tau\) and \(L\) values, the same road and the same traffic conditions are used in both the training and tracking phases.

B. Tracking Phase

During the tracking phase, the mobile phone only sends the RSSI vector information to the central server so that the server identifies the most likely sequence of the road segments traveled by the user, i.e., states \(Q = (q_1, q_2, ..., q_T)\), given the sequence of the observed RSSI vectors, i.e., observations \(O = (O_1, O_2, ..., O_T)\), where \(T\) is the number of observations. Since the parameters of the HMM are estimated during the training phase, Viterbi algorithm is used to estimate the most likely sequence of the road segments [17].

V. PERFORMANCE EVALUATION

A. Localization Error Comparison at Different Traffic Conditions in Different Locations

Figs. 4 and 5 compare the accuracy of the proposed HMM based technique using the statistics of the driver and RSSI measurements at each road segment, denoted by S-HMM, with the previously proposed HMM technique called CTrack [11] for urban and suburban roads respectively. CTrack is actually developed as a two-pass HMM first determining a sequence of grid cells corresponding to an input sequence of GSM fingerprints then matching grid cells to a road. The cost of the observed RSSI vector is a weighted function of the Euclidean distance and number of matching base stations between the observed vector and each RSSI vector in the training database. CTrack assumes uniform transition probability from the current grid cell to each of the adjacent cells and is adapted here to be used for the localization on the road. We observe that the median localization error for Ctrack is very large for small and large segment lengths whereas S-HMM has a stable localization error due to the inclusion of the driver and RSSI statistics. Moreover, the median error of S-HMM is smaller than that of Ctrack for all segment lengths.

Fig. 6 shows the cumulative distribution function (cdf) of the localization error on the suburban road in Berkeley for S-HMM, Ctrack and deterministic algorithm called KNN (K-Nearest Neighbor) [18]. In KNN algorithm, all the collected RSSI vectors and the corresponding GPS locations are stored during the training phase. During the tracking phase, the received RSSI vector at an unknown location is compared to all the RSSI vectors stored in the database and its location is estimated as the average of the GPS locations of the K-closest RSSI vectors in terms of Euclidean distance. We observe that the localization error for S-HMM is always to the left of the Ctrack and KNN algorithms for all the segment lengths. The cdf of the localization error for the urban road and suburban roads in Istanbul and New Delhi exhibit similar behavior.
B. Effect of Including Driver and RSSI Statistics

Fig. 7 shows the effect of including the driver and RSSI statistics on the localization error by comparing the performance of S-HMM to variations of S-HMM including different transition probabilities and different cost functions for the urban road at medium traffic in Istanbul. Since the behavior is similar for the other traffic conditions and locations, we did not include separate graphs for them. S-HMM uses driver and RSSI statistics to determine the transition probability matrix and emission probability of HMM respectively. “S-HMM w/ CTrack cost” and “S-HMM w/ Manhattan cost” use driver statistics to determine the transition probability matrix but CTrack cost function and Manhattan cost for emission score respectively. Manhattan cost of the observed RSSI vector is defined as the Manhattan distance between the observed vector and each RSSI vector in the training database [8]. “S-HMM w/ equal trans.prob.” on the other hand uses RSSI statistics to determine the emission probability but uniform transition probability to the adjacent segments. We observe that CTrack cost performs worst as the segment length increases whereas Manhattan cost is more stable but still performs worse than S-HMM. The cost function is the main reason for the increase in the localization error of CTrack as the segment length increases. When we compare S-HMM with “S-HMM w/ equal trans.prob.”, we see that the assumption of uniform distribution for the transition probability is the main reason for the steep increase of the localization error of CTrack as the segment length decreases. However, interestingly, the localization error of “S-HMM w/ equal trans.prob.” is smaller than that of S-HMM when the segment length is above 40 m. The main reason is that our implicit assumption of memoryless transitions between road segments in determining the statistics of the driver behavior loses validity as the segment length increases. As the segment length increases, we may need to include the state duration density in HMM [19].

C. Energy Consumption Comparison

The cell phone is expected to operate in the tracking phase after a short training period. The training can also be done by using some voluntary mobile phones only. Since cell phones continuously track base stations as part of their normal operation, the marginal energy cost of the tracking phase is driven by the CPU load to process base station signatures and send this information to the central server. Processing a cell tower signature might require at most 100,000 instructions, which costs at most 30 µJ on ARM Cortex-A8 processor. The energy required to send the packets depends on how frequently the data is uploaded to the central server. Negligible amount of additional power is consumed if the collected data is compressed using simple gzip compression and sent every minute [11]. Acquiring cell phone signatures with period 2 seconds therefore results in power consumption of 15 µW. The power consumption of GPS on the other hand is measured to be around 400 mW [11], which is several orders larger than that of the GSM based localization.

VI. CONCLUSION

In this paper, we propose a statistical fingerprinting approach to position a mobile device that follows any predetermined route by using RSSI measurements recorded from at most seven cellular base stations in the course of the normal operation of GSM-based mobile phones with minimal energy consumption. We incorporate the route constraint for the motion of the vehicles by using HMM. We demonstrate that training the HMM based on the average driver statistics on the road and the probabilistic distribution of the RSSI vectors in each road segment improves the accuracy of the previously proposed HMM based techniques considerably and provides localization performance robust over different segment lengths by using extensive cellular data collected in Istanbul, Turkey; Berkeley, USA and New Delhi, India.

REFERENCES


