This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings

A Hidden Markov Model for Localization Using Low-End GSM Cell Phones

Mohamed Ibrahim
Wireless Intelligent Networks Center (WINC)
Nile University
Smart Village, Egypt
Email: m.ibrahim@nileu.edu.eg

Moustafa Youssef
Dept. of Comp. Sc. and Eng.
School of Elect. and Comp. Eng.
Alexandria University & EJUST, Egypt
Email: moustafa.youssef@ejust.edu.eg

***Abstract*—Research in location determination for GSM phones has gained interest recently as it enables a wide set of location based services. RSSI-based techniques have been the preferred method for GSM localization *on the handset* as RSSI information is available in all cell phones. Although the GSM standard allows for a cell phone to receive signal strength information from up to seven cell towers, many of today's cell phones are low-end phones, with limited API support, that gives only information about the associated cell tower. In addition, in many places in the world, the density of cell towers is very small and therefore, the available cell tower information for localization is very limited. This raises the challenge of accurately determining the cell phone location with very limited information, mainly the RSSI of the associated cell tower. In this paper we propose a Hidden Markov Model based solution that leverages the signal strength history from *only the associated cell tower* to achieve accurate GSM localization. We discuss the challenges of implementing our system and present the details of our system and how it addresses the challenges. To evaluate our proposed system, we implemented it on Android-based phones. Results for two different testbeds, representing urban and rural environments, show that our system provides at least 156% enhancement in median error in rural areas and at least 68% enhancement in median error in urban areas compared to current RSSI-based GSM localization systems.**

I. Introduction

As cell phones become more ubiquitous in our daily lives, the need for context-aware applications increases. Knowing the location of a cell phone enables a wide set of location-based services including navigation, location-aware social networking, and security applications. Many technologies have been proposed to address the localization problem in cell phones including the GPS system, cellular-based system, and city-wide WiFi-based systems. GPS is considered one of the most well known localization techniques [1]. However, GPS is not available in many cell phones, requires direct line of sight to the satellites, and consumes a lot of energy. Therefore, research for other techniques for obtaining cell phones' location has gained momentum fueled by both the user needs for location-aware applications and government requirements, e.g. FCC [2]. City-wide WiFi-based localization for cellular phones has been investigated in [3], [4] and commercial products are currently available [5]. However, WiFi chips, similar to GPS, are not available in many cell phones and not all cities in the world contain sufficient WiFi coverage to obtain ubiquitous localization. Similarly, using augmented sensors in the cell phones, e.g. accelerometers and compasses, for localization have been proposed in [6]–[9]. However, these sensors are still not widely used in many phones. On the other hand, GSM-based localization, by definition, is available on all GSM-based cell phones, which presents 80-85% of today's cell phones [10], works all over the world, and consumes minimal energy in addition to the standard cell phone operation. Many research work have addressed the problem of GSM localization [2], [4], [11]–[13], including time-based systems, angle-of-arrival based systems, and received signal strength indicator (RSSI) based systems. Only recently, with the advances in cell phones, GSM-based localization systems have been implemented [4], [11]–[13]. These systems are mainly RSSI-based as RSSI information is easily available to the user applications.

According to the GSM standards, each cell phone can receive signals from at most seven cell towers, one of them is the current tower the cell phone is associated with and the others are the neighboring cell towers. Current localization techniques for GSM networks are designed to work with all these information. However, many cell phones, even advanced ones, do not provide the API to get access to the neighboring cell towers' information. This information is even worse in many places in the world, especially in developing countries, where low-end phones are the norm and cell towers' density is very low. This brings up the challenge of providing accurate GSM localization with minimal information, mainly the RSSI form only the associated cell towers. Therefore, new techniques are needed to address this new problem.

In this paper we propose a Hidden Markov Model (HMM)-based GSM localization scheme using only the associated with cell tower information. The main idea is to use the HMM to leverage the history of the associated cell towers and their signal strength to obtain accurate localization [1]. We describe the details of our HMM and how we estimate its parameters and how given a sequence of RSSI readings from the associated cell towers only we can estimate the phone's location.

To evaluate our system, we implemented it on Android-

This research is supported in part by a Google Research Award.

[1]Note that the associated cell tower may change as the user moves.

enabled cell phones, ***using only the associated cell tower information***, and compare its performance to both deterministic [11], [12] and probabilistic [13] fingerprinting techniques, and to Google's MyLocation service [14] under two different testbeds representing rural and urban environments. Our results show that our system provides at least 156% enhancement in median error in rural areas and at least 68% enhancement in median error in urban areas compared to other systems.

The rest of the paper is organized as follows: Section II gives a background on the current techniques for RSSI-based localization in GSM networks. In Section III we discuss our new approach. Section IV presents the performance evaluation of our system. Finally, Section V concludes the paper and gives directions for future work.

II. Background

This section presents a brief background on the current RSSI-based techniques for GSM localization that we use for comparison with our HMM-based technique, namely: cell-ID based techniques and fingerprinting techniques.

A. Cell-ID based Techniques

Cell-ID based techniques, e.g. Google's MyLocation [14], do not use RSSI explicitly, but rather estimate the cell phone location as the location of the cell tower the phone is currently associated with. This is usually the cell tower with the strongest RSSI. Such techniques require a database of cell towers' locations and provide an efficient, though coarse grained localization method.

B. Fingerprinting Techniques

Fingerprinting based techniques store the RSSI signature of cell towers at different locations in the area of interest in a database during an offline phase. This database is searched during the tracking phase for the closest location in the RSSI space to the unknown location. Fingerprints are usually constructed by war driving, where a car drives the area of interest continuously scanning for cell towers and recording the cell tower ID, RSSI, and GPS location for the associated and neighboring cell towers.

Current fingerprinting techniques for GSM localization are either deterministic [11], [12] or probabilistic [13] techniques. Deterministic techniques does not take the signal strength distribution into account. For example, each location in the fingerprint of [11] stores a vector representing the RSSI value from each cell tower heard at this location. During the tracking phase, the K-Nearest Neighbors (KNN) classification algorithm is used, where the RSSI vector at an unknown location is compared to the vectors stored in the fingerprint and the K-closest fingerprint locations, in terms of Euclidian distance in RSSI space, to the unknown vector are averaged as the estimated location. Probabilistic techniques, on the other hand, store information about the RSSI distribution in the fingerprint and try to estimate the most probable user location during the online phase. For example, in CellSense [13], the system stores the RSSI histogram for each cell tower at a particular location and uses Bayesian-based inference to estimate the user location.

Fingerprinting techniques require searching a larger database than cell-ID based techniques but provide higher accuracy. Note that the overhead of constructing the fingerprint is the same as constructing the cell ID database as both require war driving.

III. A HMM for GSM Localization

In this section, we present our HMM-based technique for GSM phones localization using only the RSSI information from the associated cell tower. We start by an overview of the system followed by the details of the offline training and online tracking phases.

A. Overview

We assume that the area of interest is divided into a grid as shown in Figure 1. Our technique works in two phases: an offline phase and and online tracking phase. The offline phase is used to construct the HMM and estimate its parameters. As we describe below in more details, each state represents a location in the discrete physical space and an observation from a state represents a RSSI reading from the associated cell tower. During the online tracking phase, a sequence of observations, representing the history of RSSI readings from the associated cell towers, is input to the HMM to estimate the most probable sequence of states (locations). The last state in the most probable sequence of states is used as the estimated location. In the following subsections, we give details about our system.

B. Mathematical Model

Without loss of generality, let $\mathbb{L}$ be a two dimensional physical space where each location represent one grid cell. A HMM, λ can be represented as $\lambda = (S, V, A, B, \pi)$ [], where:

- $S = \{S_1, S_2, S_3, ..., S_N\}$ is the set of possible states and $N = |S|$. In our case, each state represents a grid location in the physical space $\mathbb{L}$.
- $V = \{v_1, v, 2, v_3, ..., v_M\}$ is the set of observations from the model and $M = |V|$. In our case, each observation is an ordered pair of (Associate cell tower ID, RSSI).
- $A = \{a_{ij}\}$ is the state transition probability distribution, where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i], i < i, j < N$ and q_t is the state at time t.
- $B = \{b_j(k)\}$ is the observation symbol probability distribution in state j, where $b_j(k) = P[v_k \,\mathrm{at}\, t | q_t = S_j], i < j < N, 1 < k < M$ and v_t is the output symbol at time t.
- $\pi = \{\pi_i\}$ is the initial state distribution, where $\pi_i = P[q_1 = S_i]$.

Therefore, the problem becomes, given a sequence of observations $O = (O_1, ..., O_T)$, where T is a system parameter and each $O_i \in V, 1 < i < T$, we want to find the most probable sequence of locations (states) $Q = (q_1, ..., q_T)$, where each $q_i \in S, 1 < i < T$.

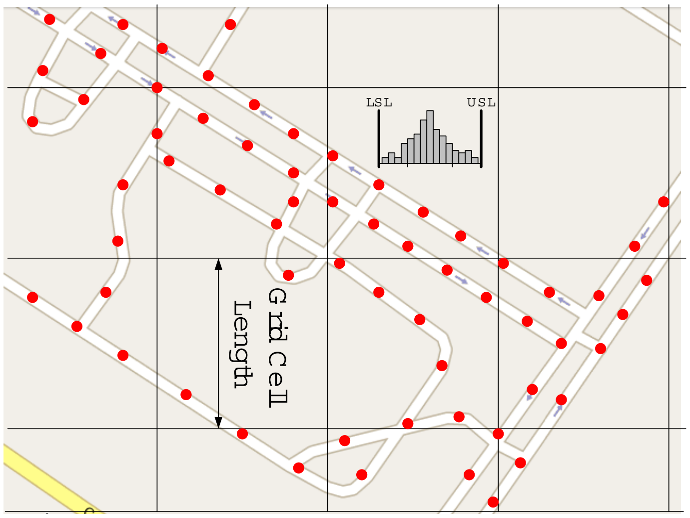

Fig. 1. *CellSense* approach for fingerprint construction. The area of interest is divided into grids and the histogram is constructed using the fingerprint locations inside the grid cell. No extra overhead is required for fingerprint construction. The grid cell length parameter can be used to tradeoff accuracy and scalability.

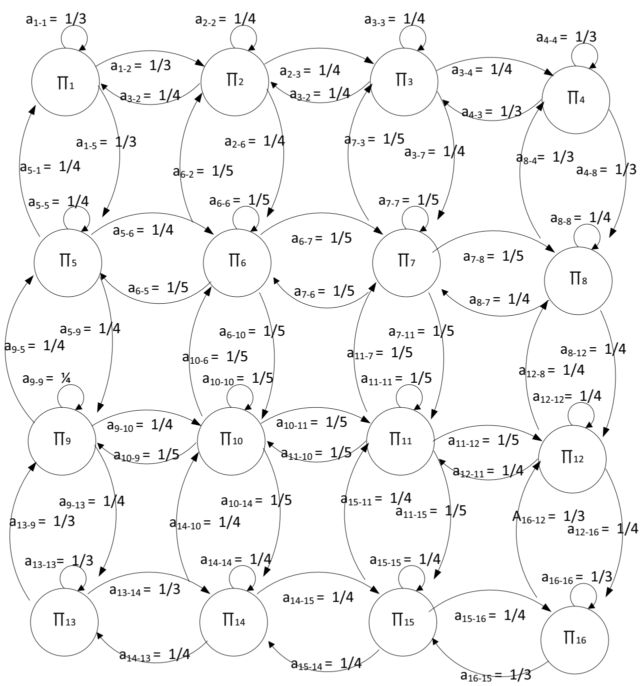

Fig. 2. The equivalent HMM for Figure 1. There is a transition from each state (grid cell) to its four neighboring states only.

C. Offline Phase

The purpose of this phase is to construct the HMM and estimate its parameters, namely (S, V, A, B, π). For our GSM localization system, the parameters are estimated as:

- S: Since each state in our model represents a physical grid, the number of states, N, is the number of grid cells. Note that the grid cell size can be tuned to balance accuracy and complexity of implementation.
- V: At each state our set of observations corresponds to the different (Associated cell tower, RSSI) pairs that can be received inside this cell.
- A: To estimate the state transition matrix, we take advantage of the grid structure. Since a user can only move between adjacent grid cells, and each state represents a cell, the transition probability is taken to be unifrom over all the neighbors of a given cell and zero otherwise as shown in Figure 2. Note that, in general, the transition matrix should depend on the shape of the road network. This can be obtained from digitized road maps, or for simplicity, it can be assumed that each cell is connected to its four or eight neighbors.
- B: To estimate the observation probability distribution at each cell, we use a technique similar to our previous work in [13], where all the data points collected during war driving inside a given cell are used to estimate the RSSI histogram inside this cell. However, contrary to [13], the histogram is constructed for the (Associated cell tower, RSSI) pairs and not for each individual cell tower, since in [13] we assume that we have information about all neighboring cell towers.
- π: If the initial state distribution is known, it can be used as is. If this information is not available, the steady state probability distribution, π_{ss} of the states can be used as an estimate for the initial state distribution. This can be estimated from the transition probability matrix, A as $\pi_{ss}A = \pi_{ss}$.

Once the HMM parameters are estimated, the system is ready for the online tracking phase.

D. Discussion

The grid cell size parameter can be used to trade-off accuracy and overhead. The larger the grid cell size, the lower the overhead as the number of grid cells decreases. However, the grid cell size increases reducing accuracy. We quantify the effect of the grid cell size parameter on the system performance in the next section.

In order to estimate the observation probability distribution (B), war driving is required. Since war driving is currently performed by many entities, such as Google, there is no overhead for constructing the observation probability distribution.

E. Online Phase

The idea of the technique is to use the history of RSSI values from the associated cell tower to estimate the user location. In particular, during the online phase, the user is moving in the area of interest receiving signal strength information from the associated cell tower only. Given a sequence of observations of length T, $O = (O_1, ..., O_T)$, we want to find the location where the user exists in at the end of the sequence. To estimate this location, we compute the most probable sequence of states $Q = (q_1, ..., q_T)$ given the sequence of observations seen by the user using the Viterbi algorithm [15]. q_T is returned as the estimated user location. Note that increasing the observation sequence length adds more information and hence should

Testbed	Area covered	Training set size	Testing set size	Avg. num. towers/loc.
One (Rural)	1.958Km^2	1198	301	5.16
Two (Urban)	5.451Km^2	2890	1051	5.97

TABLE I
COMPARISON BETWEEN THE TWO TESTBEDS.

increase accuracy. However, this comes with an increase in latency. We quantify this tradeoff in the next section.

IV. PERFORMANCE EVALUATION

In this section, we study the effect of different parameters on our system and compare its performance to other RSSI-based GSM localization systems described in Section II.

A. Data Collection

We collected data for two different testbeds. The first testbed covers the Smart Village in Cairo, Egypt which represents a typical rural area. The second testbed covers a 5.5 Km^2 in Alexandria representing a typical urban area. Data was collected using a T-Mobile G1 phone which has a GPS receiver (used as ground truth for location) and running the Android 1.6 operating system. Although the phone provides information about all neighboring cell tower, we ***did not*** use this information in evaluating the techniques to simulate the low-end cell phones.

We implemented the scanning program using the Android SDK. The program records the (cell-ID, signal strength, GPS location, timestamp) for the cell tower the mobile is connected to as well as the other six neighboring cell towers information as dedicated by the GSM specifications. The scanning rate was set to one per second. Two independent data sets were collected for each testbed: one for training and the other for testing. Table I summarizes the two testbeds.

B. Effect of Grid Cell Size

Figure 3 shows the effect of increasing the grid cell length on accuracy. The observation sequence length parameter was fixed at ten. The figure shows that as the grid cell length increases, the median error decreases and then increases again. This is attributed to two opposing factor: (a) As we increase the grid cell size, we have a better estimate for the observation probability distribution (B) as we have more samples inside each cell. This has a positive effect on accuracy. (b) As we increase the grid cell size, there is more error in location estimation inside each cell which has a negative effect on accuracy.

On another hand, as we increase the grid cell size, the computational overhead is reduced due to the decrease of the number of cells. For the rest of the paper, we fix the grid cell size at 400 for the rural and urban cases as they provide the best accuracy.

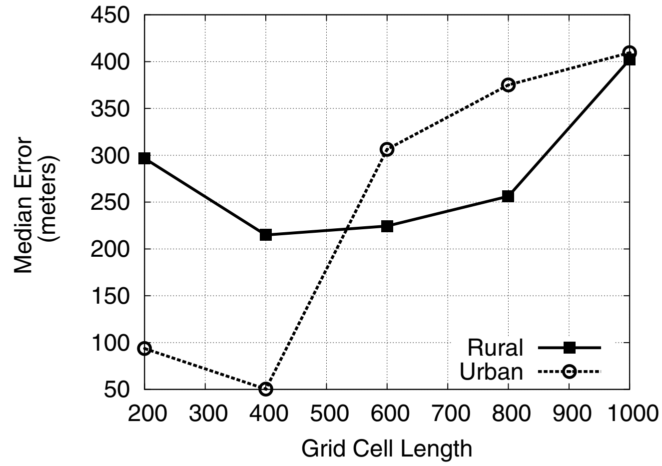

Fig. 3. Effect of changing the grid cell length on our system's median error.

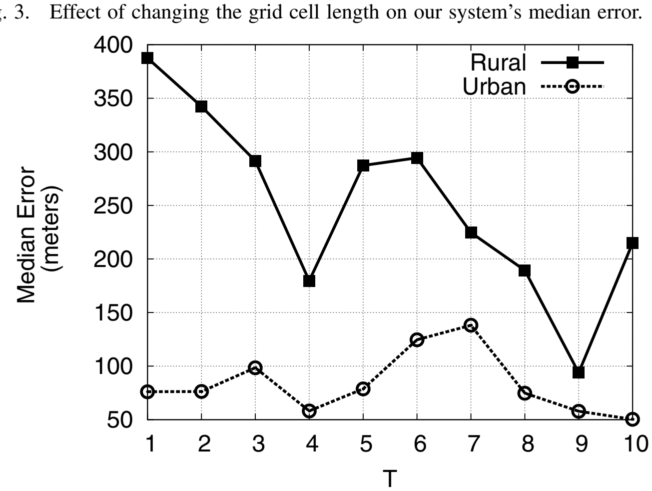

Fig. 4. Effect of the observation sequence length (T) on our system's median error.

C. Effect of Observation Sequence Length

Figure 4 shows the effect of changing the length of observation sequence, T, on the median error. The figure shows that, as expected, as the window size increases, the accuracy increases. This is due to increasing the amount of information that can be used by the HMM model. However, increasing the length of the observation sequence means that we need to wait for more samples before estimating the user location, increasing the latency of location estimation. Therefore, a balance has to be made between the accuracy and latency depending on the required application.

D. Comparison with Other Techniques

In this section, we compare the performance of our system, in terms of localization error, to other RSSI-based GSM localization techniques described in Section II. Figure 5 shows the CDF of distance error for the different algorithms for the two testbeds. The parameters that give the best median error were used for **all** algorithms. Table II summarizes the results. The table shows that our system's accuracy is significantly better than any technique achieving 93.85m median error in rural

Algorithms	Google's MyLocation	Deterministic	*CellSense*	HMM
Testbed 1-Rural Median Error(meters)	656.37 (260.52%)	263.56 (599.38%)	240.76 (156.53%)	93.85
Testbed 2-Urban Median Error(meters)	503.89 (900.9%)	89.12 (77.03%)	84.75 (68.36%)	50.34

TABLE II
COMPARISON BETWEEN DIFFERENT TECHNIQUES USING THE TWO TESTBEDS. NUMBERS BETWEEN PARENTHESIS REPRESENT PERCENTAGE DEGRADATION COMPARED TO THE PROPOSED HMM-BASED TECHNIQUE.

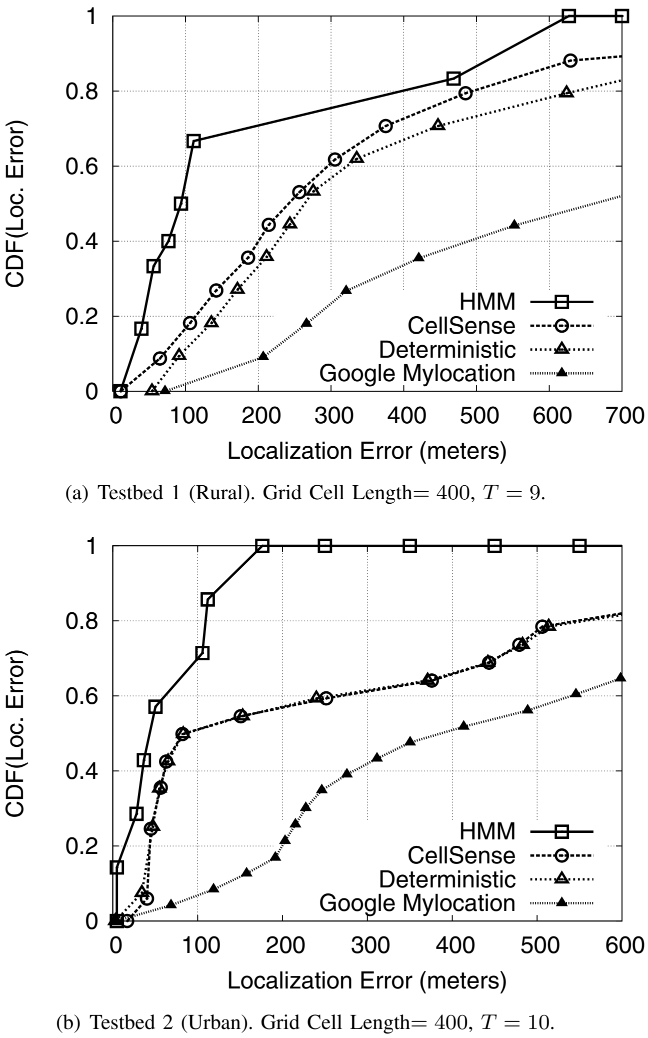

Fig. 5. CDF's of distance error for different techniques under the two testbeds. Note that the CDFs for techniques other than HMM has been truncated due to their heavy tail.

areas and 50.34m median error in urban areas. This is better than any of the other techniques by more than 156% and 68% for the rural and urban testbeds respectively. All techniques perform better in urban areas than rural areas due to the higher density of cell towers and the more differentiation between fingerprint locations due to the dense urban area structures. This is the same reason for the significant enhancement in performance of our technique in rural areas.

V. CONCLUSION

We proposed a HMM-based localization system for GSM cell phones that uses only the associated cell tower information. We presented the details of the system and how it constructs the HMM model and estimates its parameters and how we leverage the history information to enhance the accuracy. We also implemented our system on Android-based phones and compared it to other GSM-localization systems under two different testbeds. Our results show that our system accuracy is better than all other techniques achieving 93.85m median error in rural areas and 50.34m in urban areas, an enhancement over the current GSM localization techniques of at least 156% and 68% for rural and urban areas respectively. Currently, we are working on extending our system in different directions including using para- metric distributions, clustering of fingerprint locations, experimenting with larger datasets, among others.

REFERENCES

[1] P. Enge and P. Misra, "Special issue on GPS: The Global Positioning System," *Proceedings of the IEEE*, pp. 3–172, January 1999.

[2] S. Tekinay, "Special issue on Wireless Geolocation Systems and Services," *IEEE Communications Magazine*, April 1998.

[3] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm, "Accuracy characterization for metropolitan-scale wi-fi localization," in *MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM, 2005, pp. 233–245.

[4] I. Smith, J. Tabert, A. Lamarca, Y. Chawathe, S. Consolvo, J. Hightower, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, P. Powledge, G. Borriello, and B. Schilit, "Place lab: Device positioning using radio beacons in the wild," in *Proceedings of the Third International Conference on Pervasive Computing*. Springer, 2005, pp. 116–133.

[5] Skyhook wireless, "http://www.skyhookwireless.com."

[6] R. R. C. Ionut Constandache and I. Rhee, "Towards mobile phone localization without war-driving," in *IEEE Infocom*, 2010.

[7] R. S. Andrew Offstad, Emmett Nicholas and R. R. Choudhury, "Aampl: Accelerometer augmented mobile phone localization," in *ACM MELT Workshop (with Mobicom 2008)*, 2008.

[8] I. C. Martin Azizyan and R. R. Choudhury, "Surroundsense: Mobile phone localization via ambience fingerprinting," in *ACM MobiCom*, 2009.

[9] M. Youssef, M. A. Yosef, and M. N. El-Derini, "GAC: Energy-efficient hybrid GPS-accelerometer-compass gsm localization," in *IEEE Globecom*, 2010.

[10] Wikipedia, "Comparison of mobile phone standards — Wikipedia, The Free Encyclopedia," [accessed 25-March-2010]. [Online]. Available: http://en.wikipedia.org/wiki/Comparison_of_mobile_phone_standards

[11] M. Y. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. Lamarca, F. Potter, I. Smith, and A. Varshavsky, "Practical metropolitan-scale positioning for GSM phones," in *Proceedings of the Eighth International Conference on Ubiquitous Computing (UbiComp*. Springer, 2006, pp. 225–242.

[12] A. Varshavsky, M. Y. Chen, E. de Lara, J. Froehlich, D. Haehnel, J. Hightower, A. LaMarca, F. Potter, T. Sohn, K. Tang, and I. Smith, "Are GSM phones THE solution for localization?" in *WMCSA '06: Proceedings of the Seventh IEEE Workshop on Mobile Computing Systems & Applications*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 20–28.

[13] M. Ibrahim and M. Youssef, "CellSense: A probabilistic RSSI-based gsm positioning system," in *IEEE Globecom*, 2010.

[14] Google Maps for Mobile, "http://www.google.com/mobile/maps/."

[15] G. Forney Jr, "The viterbi algorithm," vol. 61, no. 3, 1973, pp. 268–278.