Estimation of Average Vehicle Speeds Traveling on Heterogeneous Lanes Using Bluetooth Sensors

Jorgos Zoto†, Richard J. La†, Masoud Hamedi§ and Ali Haghani§

***Abstract*—We investigate the problem of estimating the average speeds of vehicles traveling in different types of lanes, e.g., express lanes and local lanes, *without* knowing which lanes individual vehicles were traveling in. In our study, we use the recordings at two spatially separated Bluetooth sensors, which contain the globally unique Bluetooth device addresses of Bluetooth-enabled devices inside vehicles. These recordings are used to compute the vehicle speeds. Unfortunately, these recordings do not tell us which lanes the vehicles were traveling in and hence do not allow us to directly estimate the *separate average* speeds of vehicles, for example, in express lanes and local lanes or in high occupancy vehicle (HOV) lanes and regular lanes. We propose a novel estimation scheme that can provide *separate* average speeds of the vehicles in different types of lanes. We demonstrate the feasibility and accuracy of our proposed scheme, using real data collected by Bluetooth sensors alongside highways.**

***Index Terms*—Average speed estimation, expectation maximization, parameter estimation.**

I. Introduction

Traffic congestion, especially in and around major cities and their surrounding areas, has experienced a significant increase over the past decade or two. This increase in congestion has placed considerable strain on the transportation infrastructure and results in many lost hours for commuters. Regrettably, increasing the capacity of the transportation system, while it sounds attractive, is expensive and, in some cases, may not even be feasible due to lack of available land. For these reasons, there has been a growing interest in making use of advanced information technologies in order to improve the efficiency of the existing transportation system.

Effective management of a transportation system often demands accurate travel time information or average speeds of vehicles over different (segments of) roads in the system, which reflect real-time traffic conditions or a current state of the system. There are several existing approaches to estimating the average speeds of vehicles traveling on certain roads. These include use of GPS data [10], cameras [3], [6], loop detectors [11], Bluetooth sensors that record Bluetooth device addresses [2], [8] and received signal strength from cell phones [4], [5].

In this paper, we focus on the approach that utilizes Bluetooth sensors deployed alongside roads [2], [8]. Bluetooth is a wireless technology developed for short to medium range, low rate communication between wireless devices, such as cellular phones and other portable electronic devices [9]. A Bluetooth device has a *globally unique* 48-bit Bluetooth device or medium access control (MAC) address, which is electronically engraved on the device. The Bluetooth sensors deployed in the field wake up periodically and scan the wireless medium to pick up broadcast messages from which they extract the MAC addresses of Bluetooth devices in the vicinity.

A. Current limitation of Bluetooth sensor-based approach and motivation for our study

Many parts of the highways in the U.S. have two or more separate sets of lanes. For instance, in many urban and suburban areas surrounding metropolitan areas (e.g., the Washington D.C. metropolitan area), some of the lanes are dedicated for high occupancy vehicles, which are called HOV lanes, during peak hours. In some other parts, the lanes are divided into express lanes and local lanes (e.g., New Jersey Turnpike). The (average) speed of the vehicles will likely be different in express lanes and local lanes (or regular lanes and HOV lanes). Hence, the average speed of *all* vehicles may not be a good estimate of the average speed on either type of lanes when the average speeds differ significantly, for example, after an accident or construction on local lanes.

Unfortunately, when the Bluetooth sensors are deployed in the field along the highways to record the MAC addresses of Bluetooth devices in vehicles that pass by the sensors [8], e.g., GPS devices or cellular phones, it is difficult, if possible at all, to determine which lane each vehicle was traveling in from sensor recordings. Even if one could estimate the *distance* of the vehicle from the sensor using received signal strength, which is a challenging task, this does not provide enough information to determine the lane of the vehicle without knowing the angle of arrival of the received signal.

This provides a motivation for our study. More specifically, we are interested in the problem of estimating the average speeds of the vehicles on different types of lanes (e.g., regular lanes vs. HOV lanes or express lanes vs. local lanes) *without* knowing which lanes vehicles were traveling in beforehand. There are several possible approaches one can employ to tackle this problem. For example, a reasonable first approach is to *attempt to identify the lanes* in which the vehicles are traveling and then estimate the average speed in each type of lanes. This essentially reduces to a *classification* problem. In order for this approach to work well, vehicles in different types of lanes should travel at different speeds so that we can *classify* the

J. Zoto and R.J. La are with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park. M. Hamedi and A. Haghani are with the Department of Civil and Environmental Engineering, University of Maryland, College Park.

978-1-4673-1881-5/12/$31.00 ©2012 IEEE

vehicles effectively, based on the speeds of individual vehicles. However, in general it is difficult to classify a vehicle based on its speed. The reason for this is that a vehicle moving at some speed could be either in, for example, a regular lane or an express lane. Therefore, if there is no simple way to *classify* the vehicles based on their speeds, this approach is unlikely to perform well in practice.

II. Description of our approach

In this paper, we take the viewpoint that the speeds of individual vehicles recorded in the collected data sets come from one or more distinct distributions. In other words, we assume that the speeds of the vehicles are *random variables* (rvs) generated according to a *mixture distribution*,[1] and our goal is to determine (i) the number of mixture components, (ii) the mixture weights and (iii) the *means* of mixture components. In particular, we use *Gaussian mixture distributions* to estimate the average speeds and mixture weights.

a) Selection of the number of mixture components: First, in order to determine the number of mixture components, we make use of *Akaike information criterion* (AIC) [1]. Although there are many measures for assessing the "goodness" of a fitting distribution for a given data set, we use the AIC to carry out a trade-off between *complexity* and *accuracy* of the fitting distribution. In our problem, the complexity is proportional to the number of mixture components. Taking into account the complexity of the fitting distribution is important when the data set is limited, as they are in our problem over a short interval, and we are interested in identifying only the principal mixture components.

The aforementioned trade-off between complexity and accuracy is carried out through the expression

$$AIC = 2\ K - 2\ \log(L), \tag{1}$$

where K is the *number of parameters* in the distribution and L is the *maximized likelihood* for the given parameters. Thus, while the second term prefers a more accurate distribution, the first term penalizes the distribution with a larger number of parameters to be estimated. From (1), we are interested in finding a fitting distribution with the smallest value of AIC.

b) Parameter estimation and expectation-maximization (EM) algorithm: Once the number of mixture components is determined, we use the estimated *means* of the mixture components and the mixture weights to approximate the average speeds of the vehicles and the percentage of vehicles traveling in different types of lanes, respectively. This step can be formulated as an *estimation* problem in which the parameters to be estimated are the means and weights of the mixture components. While there are several well known estimators, our study reveals that the EM algorithm [7] performs better than many popular estimators, such as maximum likelihood estimators (Section IV).

The EM algorithm is an iterative statistical estimation procedure in which each iteration consists of two processes - an *expectation* step and a *maximization* step. The EM algorithm is often employed for parameter estimation when, besides observable variables $\mathbf{X}$, there are hidden or latent variables $\mathbf{Z}$ that cannot be observed directly: suppose that we are interested in estimating parameters $\Theta \in \mathcal{S}_\Theta$. The estimate of the parameters at the n-th iteration, $n = 1, 2, \ldots$, is denoted by θ_n.

1) *Expectation* step: At the $(n+1)$-th iteration, along with the observed value $\mathbf{X} = \mathbf{x}$, the estimate θ_n is used as the true value of Θ to compute the conditional distribution of $\mathbf{Z}$. This conditional distribution of $\mathbf{Z}$ is in turn employed to compute the expected value of the log-likelihood $\mathbf{E}_\mathbf{Z}\left[\ln\left(\mathbf{P}\left[\mathbf{x}, \mathbf{Z} \mid \theta\right]\right)\right]$, $\theta \in \mathcal{S}_\Theta$, where the expectation is taken with respect to the conditional distribution of $\mathbf{Z}$ given $\{\mathbf{X} = \mathbf{x},\ \Theta = \theta_n\}$.

2) *Maximization* step: In the maximization step, the new estimate of Θ at the $(n+1)$-th iteration is set to

$$\theta_{n+1} = \arg\max_{\theta \in \mathcal{S}_\Theta} \mathbf{E}_\mathbf{Z}\left[\ln\left(\mathbf{P}\left[\mathbf{x}, Z \mid \theta\right]\right)\right]. \tag{2}$$

If there is more than one solution to the maximization problem in (2), one of them is chosen arbitrarily.

In our problem, there is a latent variable associated with each pair of a sample point, i.e., a vehicle speed, and a mixture component. These latent variables determine which mixture component is responsible for generating the sample point. Hence, rather than *classifying* the sample points first, the EM algorithm will make use of *conditional probabilities* of these latent variables to determine how well the current estimate of the parameters fits the given sample points.

III. Numerical Results

In this section we present the numerical results from the experiments we conduct using the *real measurements* collected along several different segments of the interstate I-95 highways in the U.S. First, we explain how the measurements are collected and how we use the measurements to compute the (average) speeds of vehicles. Then, we provide the numerical results we obtain using our proposed scheme.

A. Bluetooth sensor measurements

The overall experimental setup is illustrated in Fig. 1. In our setup, two Bluetooth detectors or sensors are placed along a road that is at least 1 mile long. The sensors wake up every 5 seconds to scan the wireless medium. As vehicles equipped with Bluetooth devices pass by the sensors, the scanned MAC addresses and the time of scanning are recorded by the sensors. A typical sensor can have a scanning radius of up to 300 feet. When the same MAC address is recorded by both sensors, using the known distance between the two sensors and the recorded times of scanning, we calculate the average speed of the vehicle between the sensors by dividing the distance by the travel time. Throughout the rest of the section, we refer to these average speeds of individual vehicles simply as speeds, in order to distinguish them from the *average* speeds of more than one vehicle.

[1] Here, we allow a mixture distribution to have a single component.

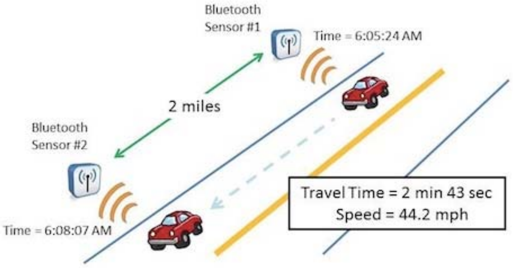

Fig. 1. Bluetooth sensor measurements and speed estimation.

B. *Experimental Results*

We examine three different scenarios in order to evaluate the performance of our proposed scheme. The first two are based on the Bluetooth sensor measurements on two distinct segments of the I-95 corridor without any modification to the data sets. The last scenario represents a *controlled* experiment, which is used to validate the accuracy of our scheme. In all cases, a file that contains the speeds of recorded vehicles is given to our proposed scheme as input. Our scheme then produces i) the number of mixture components with the corresponding value of AIC, ii) the means of mixture components, and iii) the mixture weights for Gaussian mixture distributions with the three smallest values of AIC (as the number of mixture components is varied).

1) Experiment #1 – Express and Local Lanes: In the first experiment, we consider a segment of New Jersey Turnpike with both express lanes and local lanes in Ridgefield Park, NJ. The measurements were taken on the 21st of April, 2011 from 6:00 AM to 7:00 AM. Fig. 2 plots the speeds of the vehicles we computed (as described in Section III-A) using the recordings by the sensors. There are a total of 107 sample points. Clearly, it would be difficult, if possible at all, to classify the vehicles into express lane vehicles and local lane vehicles based solely on their speeds plotted in the figure.

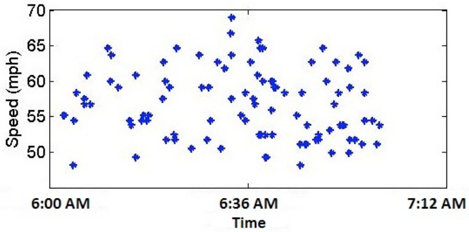

Fig. 2. Plots of vehicle speeds between 6:00 AM and 7:00 AM on April 21, 2011 (Ridgefield Park, NJ).

Fig. 3 plots the histogram generated from the computed vehicle speeds (blue $\star$) and three different fitting Gaussian mixture distributions with the three smallest values of AIC. For this experiment, the fitting distributions with 1, 2 and 3 mixture components have the three smallest values of AIC. These values of AIC are also provided in the figure. In addition, we plot the *smoothed* histogram we obtain by taking the average of three consecutive histogram points - the original value and the two nearest neighbors in the histogram with equal weights of 1/3 (green *). As evident from Fig. 3, while the original histogram is very noisy due to a limited sample size, the *smoothed* histogram provides a clearer picture of the distribution of vehicle speeds.

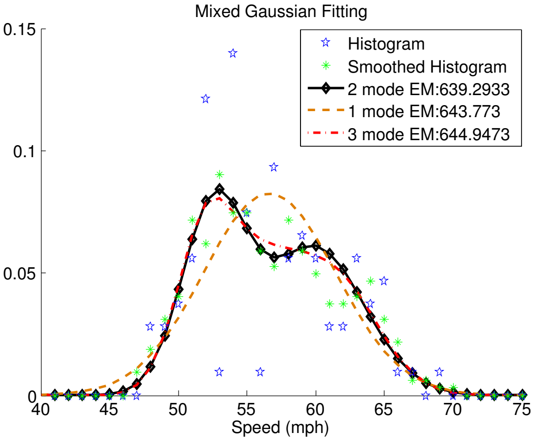

Fig. 3. Plots of histograms and fitting Gaussian mixture distributions with 1, 2, and 3 mixture components.

As one can see from Fig. 3, the Gaussian mixture distribution with two mixture components has the smallest value of AIC. The estimated means of the two mixture components are 52.5 miles per hour (mph) and 59.9 mph with the mixture weights of 0.443 and 0.557, respectively. Hence, according to our proposed scheme, approximately 44.3 percent of the recorded vehicles were traveling in the local lanes with the average speed of 52.5 mph, while 55.7 percent of recorded vehicles were in the express lanes with the average speed of 59.9 mph. Unfortunately, while these estimates appear reasonable, we cannot verify their accuracy because we do not know which lane each vehicle was in. This issue will be addressed in the last scenario (Experiment #3).

It is obvious from Fig. 3 that the histogram of the vehicle speeds is quite noisy and it is hard to determine the distribution of the speeds directly from the histogram. However, our fitting Gaussian mixture distribution with two mixture components closely resembles the *smoothed* histogram of the speeds.

2) Experiment #2 – Toll Plaza: In the second scenario, we examine the measurements obtained on a segment of I-95 highways that includes a toll plaza in Iron Hill Park, DE, near the border between Delaware and Maryland. The measurements were taken on the 1st of July, 2011 between 5:30 PM and 6:00 PM. There are a total of 93 samples in this scenario.

The reason we are interested in studying this data set is as follows: the toll plaza has separate lanes for E-ZPass holders different from cash lanes. Since E-ZPass holders do not have to stop to pay the toll, one expects them to travel at higher speeds over the segment than the vehicles that go through the cash lanes, which often have lines of vehicles waiting to pay the toll. Hence, we expect to see noticeable differences in the (average) speeds of the vehicles going through the E-ZPass lanes and the cash lanes. We show that indeed this is true and our proposed scheme can easily recognize this fact and offers two separate estimates - one for the vehicles through the E-ZPass lanes and the other for the vehicles through the cash lanes.

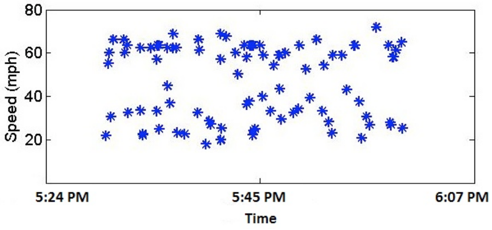

Fig. 4. Plots of vehicle speeds between 5:30 PM and 6:00 PM on July 1, 2011 (Iron Hill Park, DE).

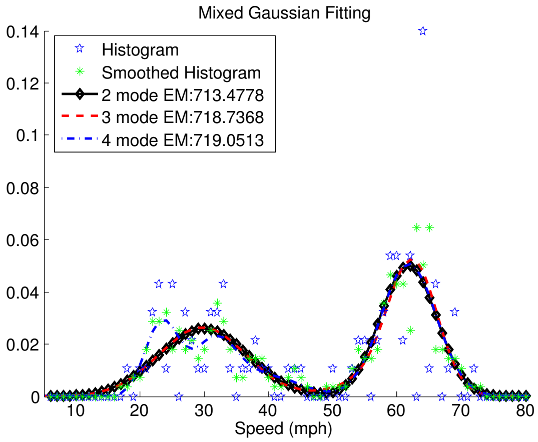

Fig. 5. Plots of histograms and fitting Gaussian mixture distributions with 2, 3 and 4 mixture components.

We plot the samples, i.e., vehicle speeds, in Fig. 4. One can easily see from the figure that indeed there are two groups of vehicles traveling at different speeds, one concentrated around 60 mph and the other around 30 mph. Both the histograms and the fitting distributions are shown in Fig. 5. Once again, the Gaussian mixture distribution with two mixture components has the smallest value of AIC and appears to approximate the smoothed histogram of the speeds reasonably well, although even the smoothed histogram seems somewhat noisy in this case, in part due to a smaller sample size. The estimated means of the mixture components are 61.7 mph and 29.9 mph with the mixture weights 0.546 and 0.454, respectively. Hence, roughly 55 percent of the recorded vehicles used the E-ZPass lanes according to our estimate.

Note that there is an outlier in the histogram around 63 mph, which can distort the average value if one tries to estimate the average speed of the vehicles using the *sample average*. Our proposed scheme, however, is not affected by this outlier significantly because it recognizes such outliers that occur with small probability.

3) Experiment #3: Controlled Experiment: In the final scenario, instead of taking a single set of measurements collected by the sensors over a single period, we merge two sets of measurements taken over two *disjoint* periods into one. Both sets of measurements were taken from the same segment of I-95 in New Jersey, on two different dates. The first measurement set was gathered between 4 PM and 5 PM on the 20th of April, 2011, and the second measurement set was collected between 5:30 PM and 6 PM on the 21st of April, 2011. The first set contains 72 samples and the second set has 22 samples. We merge these two measurement sets into one data set and use it as an input file to our proposed scheme. Thus, the first data set and the second data set account for 76.6 percent and 23.4 percent, respectively, of all samples.

The goal of this exercise is to create a *controlled* experiment for which we know the answers and investigate how accurately our proposed scheme can estimate the average speeds when vehicles travel in two different types of lanes, in particular, an HOV lane and regular lanes, where a smaller number of vehicles are expected to use an HOV lane. The first data set (with 72 samples) serves as the set of samples that come from the regular lanes in our *controlled experiment*, while the second data set (with 22 samples) provides the samples that (we pretend) come from an HOV lane.

Since we expect the vehicles in an HOV lane to move faster during rush hours than the vehicles in regular lanes, which tend to be more congested, we add a *fixed* constant to all the samples in the second data set to mimic the higher anticipated speeds of vehicles in an HOV lane during rush hours. After this modification to the samples, the average speed of the vehicles in the second data set is 72.6 mph, and that of the vehicles in the first data set is 58.0 mph.

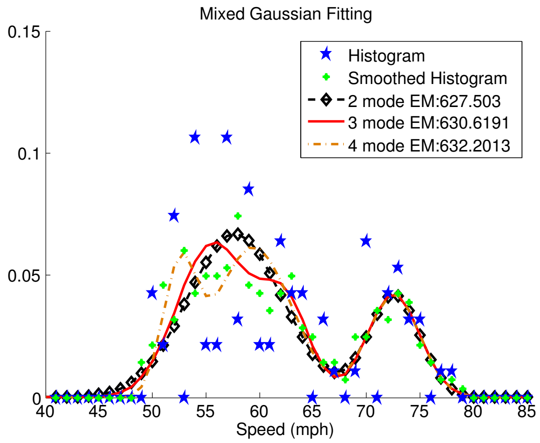

Fig. 6. Plots of histogram and fitting distributions.

In Fig. 6, we plot the Gaussian mixture distributions with the three smallest values of AIC. As shown in the figure, the Gaussian mixture distribution with two mixture components has the smallest value of AIC. Fig. 6 also shows the histogram of the vehicle speeds. The means of the two mixture components estimated by our proposed scheme are 57.7 mph and 72.6 mph with mixture weights [0.747 0.253]. Clearly, these estimates are close to the true values that we use in this controlled experiment, demonstrating the accuracy of our proposed scheme.

IV. Comparison against other maximum likelihood estimators

Our proposed scheme is based on the EM algorithm, which is well suited for our purpose and has a provable convergence property [7]. However, there are other estimators, including the popular maximum likelihood estimator (MLE), and other mixture distributions one can employ to fit the measurements. In this section, due to the space constraint, we briefly discuss their performance and mention some problems one may encounter with the MLE utilizing different mixture distributions.

First, for a given mixture distribution whose parameters need to be estimated, the MLE simply chooses the parameters that are *most likely* to generate the observed measurements. Hence, when the measurements are noisy due to the limited sample size, the MLE often believes that there is a single mixture component even when there is more than one mixture component.

Second, the output of the MLE tends to be sensitive to the initial starting point. More specifically, when the *overall* sample mean and variance of the vehicle speeds are used as the initial means and variances of mixture components with evenly distributed initial mixture weights, the MLE does not always produce (close to) correct values.

Type of estimator	Mixture distribution	Means of two components
EM	Gauss-Gauss	57.7, 72.6
MLE	Weib-Weib	57.2, 71.5
MLE	Gamma-Gamma	$63.1, 2.61 \cdot 10^5$
MLE	Gauss-Weib	57.6, 71.7
MLE	Gauss-Gamma	$61.5, 6.99 \cdot 10^3$
MLE	Weib-Gamma	$61.3, 5.99 \cdot 10^3$

TABLE I
SUMMARY OF CONTROLLED EXPERIMENT (WITH MEANS [58.0 72.6] MPH) – GAUSS = GAUSSIAN, WEIB = WEIBULL.

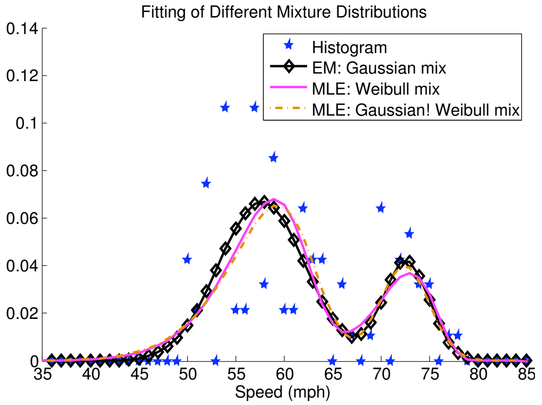

Fig. 7. Plots of histogram and different fitting distributions.

These are illustrated in Table I. In this example, we use the same controlled experiment in the previous subsection, i.e, experiment #3 with average speeds of 58.0 mph and 72.6 mph. We use the MLE to find the parameters of different mixture distributions with two mixture components. As shown in the table, except for two cases (Weib-Weib and Gauss-Weib), the MLE puts a mixture weight of one on a single mixture component and outputs the *overall* average speed of the vehicles as its mean. In addition, an (almost) arbitrary number is chosen as the mean of the second mixture component with zero mixture weight, as it has no meaning. The mixture distributions for the two cases that produce two mixture components (i.e., Weib-Weib and Gauss-Weib) are plotted in Fig. 7 along with the fitting Gaussian mixture distribution selected by our proposed algorithm. As we can see from Fig. 7 and Table I, while they are relatively close, our proposed scheme provides more accurate estimates.

V. Conclusion

We studied the problem of estimating the average speeds of vehicles traveling on different types of lanes. The problem is formulated as one of finding the parameters of mixture distributions (and corresponding mixture weights). We proposed a novel estimation scheme based on the well known expectation-maximization (EM) algorithm, which estimates the means of mixture components and the mixture weights. We conducted experiments using real measurements and demonstrated the accuracy of our proposed scheme.

References

[1] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. on Automatic Control*, 19(6):716-723, Dec. 1974.

[2] C. Bakula, W.H. Schneider IV and J. Roth, “Probabilistic model based on the effective range and vehicle speed to determine Bluetooth MAC address matches from roadside traffic monitoring,” *Journal of Transportation Engineering*, 137(1):43-49, Jan. 2012.

[3] “A novel technique to dynamically measure vehicle speed using uncalibrated roadway cameras,” Proc. of IEEE Intelligent Vehicles Symposium, pp. 777-782, Jun. 2005.

[4] G. Chandrasekaran, T. Vu, A. Varshavsky, M. Gruteser, R.P. Martin, J. Yang and Y. Chen, “Vehicular speed estimation using receiver signal strength from mobile phones,” Proc. of the 12th ACM International Conference on Ubiquitous Computing (Ubicomp), Copenhagen (Denmark), Sep. 2010.

[5] G. Chandrasekaran, T. Vu, A. Varshavsky, M. Gruteser, R.P. Martin, J. Yang and Y. Chen, “Tracking vehicular speed variations by warping mobile phone signal strengths,” Proc. of IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 213-221, Seattle (WA), Mar. 2011.

[6] D.J. Dailey, F.W. Cathey and S. Pumrin, “An algorithm to estimate mean traffic speed using uncalibrated cameras,” *IEEE Trans. on Intelligent Transportation Systems*, 1(2):98-107, Jun. 2000.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, 39(1):1-38, 1977.

[8] A. Haghani, M. Hamedi, K. Sadabadi, S. Young and P. Tarnoff, “Data collection of freeway travel time ground truth with Bluetooth sensors,” *Journal of the Transportation Research Board*, 2010.

[9] B.A. Miller and C. Bisdikian, *Bluetooth Revealed*, Prentice Hall PTR, 2001.

[10] S. Poomrittigul, S. Pan-ngum, K. Phiu-Nual, W. Pattara-atikom and P. Pongpaibool, “Mean travel speed estimation using GPS data without ID number on inner city road,” Proc. of the 8th International Conference on ITS Telecommunications, pp. 55-61, Phuket (Thailand), Oct. 2008.

[11] Y. Wang and N.L. Nihan, “Freeway traffic speed estimation using single loop outputs,” *Transportation Research Record*, 1727(1):120-126, 2000.