The ARPEGEO Project

A New Look at Cellular RSSI Fingerprints for Localization

Iness Ahriz[1], Bruce Denby[2,1], Gérard Dreyfus[1], Rémi Dubois[1], Pierre Roussel[1]
[1]SIGMA (Signal processing and Machine learning) Laboratory, ESPCI ParisTech
[2]Université Pierre et Marie Curie
Paris, France
[2]denby@ieee;org; [1]firstname.lastname@espci.fr

Abstract—**A new technique developed at ESPCI ParisTech should allow cellular received signal strength fingerprints to play an important role in localization systems for regions which are not well covered by GPS. The article describes the ARPEGEO project, initiated to evaluate the impact of full-band GSM fingerprints analyzed with modern machine learning techniques. Results on indoor localization, as well as techniques to facilitate practical implementation of the method, are presented.**

Keywords-GSM; fingerprint; localization; indoor; machine learning

I. Introduction

Development of localization techniques for regions where GPS does not work well is currently a very active area of research. Substantial literature exists, for example, on methods making use of UWB nodes, WiFi RSSI (Received Signal Strength Indicator) signals, accelerometers or other inertial devices, magnetometers, and the like [1-4]. Recently, dynamic multi-sensor approaches combining two or more localization technologies have also become rather common [5-7].

Although radiotelephone base stations, like WiFi access points, provide fixed-power beacon signals that may also be exploited for localization, cellular-based approaches have predominantly been limited to outdoor applications due to their low accuracy – for example, commercial location based services (LBS) based on 7-carrier GSM network measurement report (NMR) RSSI fingerprints, with an accuracy of about 150 meters [8]. Applications of cellular RSSI fingerprints to indoor localization have also appeared in the literature [9], and there is evidence that fingerprints with higher-carrier-counts are useful here [10]. Nevertheless the prevailing logic in the localization community has remained that additional carriers beyond the first few strongest ones will be irrelevant or redundant; difficult to analyze should they begin to number in the hundreds; and in any case impossible to obtain using the hardware available in everyday electronic devices.

The ARPEGEO project (Analysis of Radioprints for Enhanced Geolocalization) at the SIGMA Laboratory of ESPCI ParisTech was initiated to study the localization capability of GSM fingerprints containing *all* carriers in the GSM band – more than 500 channels in most installations. Recent work performed at our laboratory has indicated that by using machine learning techniques to manage the high-dimensionality of such full-band fingerprints, localization performance far superior to that obtained with more standard-sized RSSI vectors – of the order of a few meters – is in fact possible [11]. It furthermore appears evident that the GSM frequency scans necessary to obtain the required fingerprints will be able to be performed on a standard cellular telephone with appropriate software. Thus, cellular RSSI fingerprints, contrary to traditional reasoning, may indeed be able to play an important role as part of a modern indoor localization system.

In what follows, an overview of the technological solutions being developed in ARPEGEO is presented. Currently employed and possible future hardware RSSI acquisition platforms are described in the following section. Section III presents indoor localization results on two test datasets, which clearly show the importance of including large numbers of carriers in the fingerprint scans, while a detailed description of the classification and variable selection algorithms employed in our work appears in section IV. An oft-cited criticism to machine learning approaches is the "black-box" nature of such techniques, which make it difficult to ascertain "how" the system is performing its localization and to understand what the relevant system parameters actually are. The discussion in section V uses our variable selection procedure to shed light on precisely what information is being used for localization – with sometimes surprising results. Some shortcomings of our method, as well as a study of its temporal stability, are also presented in that section. Finally, as mentioned earlier, a consensus is emerging in the localization community that the "ultimate" indoor solution will likely be a hybrid of several different technologies. A discussion of how those developed in ARPEGEO might integrate into such a vision appears with our concluding remarks in section VI.

II. Hardware Platforms

The starting point for a system using full-band cellular fingerprints (we limit our discussion to GSM here; similar techniques are possible in 3G networks), is a hardware platform capable of monitoring all frequencies in the band and recording the information to disk. The two most common approaches, both of which have been tested in ARPEGEO, are:

- **Trace Mobiles**. Cellular engineers have for years used so-called "trace mobiles" to analyze and troubleshoot the radio network interface. To limit development costs, most manufacturers use the same hardware for trace mobiles and standard cellphones, simply disabling monitoring mode in the consumer units. Frequency scanning capability is one standard feature of such devices. The TEMS trace mobile system [12] is used in ARPEGEO.

- **M2M Modules**. Beyond its utility for personal communications, the ubiquity and simplicity of the GSM system has made it an attractive alternative for industrial communications interfaces as well. These make use of so-called Machine-to-Machine, or M2M, modules, which are implemented using cellphone chipsets configured as Hayes-compatible modems. Some of these, such as the Telit GM862-GPS [13] used in ARPEGEO, also have scanning capability.

The TEMS handset used for the data in this article (an older SH-888 model) required more than a minute to scan an entire GSM band, and the scan time of the GM862 module used is similar. Although this permits us to evaluate test scenarios, realistic dynamic localization is not feasible with these two hardware platforms. More recent chipset, however, allow significantly more processing to be performed on the telephone. The Sony-Ericsson W995, for example, is available as a TEMS Pocket [14] trace mobile with a form factor identical to the standard W995, and is able to scan 1600 carriers per second (without Base Station Identity Code (BSIC) decoding) [15]. With such a scan rate, a full band GSM scan in a possible future device would require only about 300 milliseconds to execute. These scans, furthermore, are carried out in idle mode, so that no actual connection to any cell tower is required for localization.

III. Datasets, Results

The most stringent test of any proposed localization technology is an evaluation of its performance as a stand-alone system in a static, memory-less scenario. Tests of this type performed in two indoor localization settings, as described below, demonstrate that full-band GSM fingerprints give performance far superior to that obtainable with more "standard" fingerprints, and are able to identify the room in which the hardware acquisition platform is located with more than 95% accuracy.

A. Home and Lab Datasets

Data were recorded over a one-month period in 5 rooms of a research laboratory (the *Lab* set) and 5 rooms of a private apartment (the *Home* set). The *Lab* set, recorded with the Telit GM862-GPS, contained 600 full-band GSM scans, and the *Home* set, which used a TEMS trace mobile, 241. For both datasets, the scans, labeled by room number from 1 to 5, contained approximately 500 carriers. Of these, only a fraction correspond to fixed-power beacon channels, the rest being traffic channels which, due to their variability, are normally not expected to be useful for localization. Theoretically, beacons can be identified by the presence of a BSIC; however, due to attenuation or multipath effects, these are sometimes not decoded. We chose to ignore BSICs in our study, for three reasons:

1. The decoding may fail, as mentioned;
2. Scanning is more rapid when BSICs are not requested;
3. To remain open to the possibility that non-beacon channels could be useful for localization.

At the same time, this choice will of course require our analysis algorithms to handle significant numbers of potentially noisy inputs, in addition to the beacon signals.

Three types of fingerprints were defined: 1) "Standard", containing about 40 carriers, which includes all carriers occurring at least once in the top 7 ranked by mean power over the training set (see section IV); 2) "All", containing all carriers (about 500 channels); and 3) "Relevant", containing about 30 carriers which were selected for their "relevance" by an algorithm called Orthogonal Forward Regression, which we describe in section IV.

B. Results

The three fingerprints were compared on our two databases in a static (i.e., no memory of earlier positions or sensor inputs), indoor, room-level classification problem. The results presented here are based on a 5-class classifier using Linear Support Vector Machines (SVM). The operation of this classifier is detailed in section IV. For each database, 80% of the data were used for SVM training and validation (see section IV); the performances presented in Table I were then computed by applying the trained classifier to the remaining 20% of the data. Both the train and test fingerprints are uniformly distributed in time over the one-month period (see section V.B for a discussion of the temporal stability of the method).

TABLE I. Linear SVM results on *Lab* and *Home* sets

Data set	**Fingerprint**		
	Standard (≈40 carriers)	All *(≈500 carriers)*	Relevant (≈30 carriers)
Home	68.9%	96.7%	93.4%
Lab	59%	99%	95%

The table shows that the performance of the "Standard" fingerprints is unacceptably poor, whereas including all available carriers in the fingerprint leads to very good performance. Nevertheless, a solution requiring 500 input variables presents some problems of interpretation. When, however, a subset of carriers selected by their "relevance" for localization is employed, as in column 3 of the table, a much simpler solution is obtained, at the price of only slightly reduced efficiency. We shall return to this point in the discussion in section V. Below, we detail the variable selection procedure used, as well as the functioning of the SVM algorithm.

IV. Algorithms

A. Support Vector Machine classifiers (SVM)

Our method aims at indicating in which room among 5 the data are being recorded. To perform this task, 10 pairwise classifiers, i.e., classifiers that discriminate room i from room j, $i, j = 1,\dots,5,\ i \neq j$) were designed.

It was first ascertained, by running the Ho-Kashyap algorithm [16], that the examples available in the training-validation set were linearly separable pairwise. This result allowed us to make use of linear Support Vector Machine (SVM) classifiers [17, 18]. A linear SVM provides, from examples, the optimal separating hyperplane in feature space, i.e., the separating hyperplane that classifies all examples without error, while lying as far as possible from the closest examples. Denoting by $\mathbf{x}$ the vector of features describing the items to be classified (in our case, the powers of all received carriers, or of a subset thereof), and by $\boldsymbol{\theta}$ the vector of parameters of the model, the equation of the hyperplane is of the form

$$\mathbf{x}.\boldsymbol{\theta}=0 \tag{1}$$

Training is the process whereby the values of the parameters are estimated from the examples. It is cast in the form of a constrained optimization problem, where the function to be minimized is the norm of the vector of parameters, under the constraint that all examples be correctly classified (”hard-margin” SVMs).

The central problem in machine learning is the ability of the trained models to generalize, i.e. to correctly classify examples that are not present in the training set. The fact that the magnitude of the vector of parameters is kept as small as possible minimizes the risk of poor generalization. However, allowing some examples of the training set to be misclassified may further improve the generalization ability of the model. This leads to “soft-margin SVMs”, where the function to be minimized contains, in addition to the norm of vector $\boldsymbol{\theta}$, a term that is roughly proportional to the number of misclassified examples, with a proportionality coefficient (termed “regularization constant”), which must be determined by the model designer.

In the present study, the value of the regularization constant was found by cross-validation: the training set of each pairwise classifier was divided into ten folds; one of them was used in turn as a validation set, on which the performance of the classifier trained on the other 9 folds was estimated. Thus, for each pair of rooms $\{i, j\}$, 10 classifiers with the same value of the regularization constant C_{ij} were trained, and the cross-validation score was computed as the average classification score on the validation sets. The procedure was iterated for different values of C_{ij} in a prescribed range, and the value of C_{ij} that yielded the best cross-validation score was retained.

Finally, each pairwise classifier was trained on the data contained in all ten folds with the value of the regularization constant found by cross-validation, and the resulting classifier was tested on the test set, i.e. on fresh data that were not used during the cross-validation procedure.

B. *Overall system*

The final classification decision was made by a vote on the basis of the results of the 10 two-room classifiers designed as described in the previous section: the predicted class was the most frequently chosen room. An overview of the algorithm is presented in figure 1. It is the localization performance obtained with this procedure that is presented in Table I.

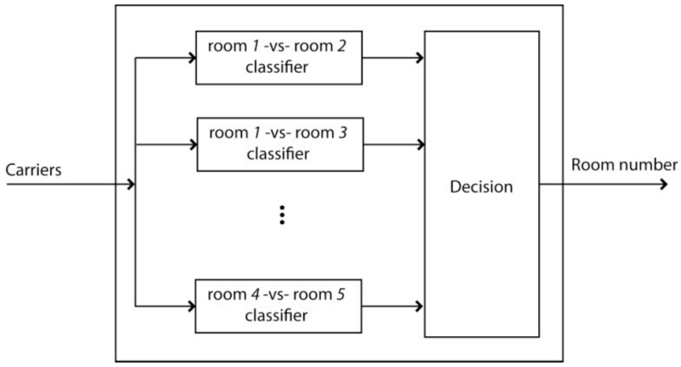

Figure 1. Overview of the system: the input “carriers” of the classifier can be all available carriers, “All”; or a subset thereof, “Standard” or “Relevant”.

C. *Selection of the relevant carriers*

This section focuses on the feature selection technique used to obtain the “Relevant” fingerprint subset. The procedure, called Orthogonal Forward Regression (OFR), is based on Gram-Schmidt orthogonalization [19] and relies on the correlation between the target and the features.

For the classifier that discriminates room i from room j, the target is the label of the room (+1 for room i or −1 for room j) where the carrier power measurements were made. The first feature selected is that which exhibits the highest correlation with the target. The remaining inputs and the target are then orthogonalized with respect to this first selected feature and the process is iterated until some termination criterion is met, thereby resulting in a list of carriers ranked in order of decreased relevance. This feature ranking technique, based on linear correlations, is well suited to the design of a linear classifier such as the linear SVMs used here.

The optimal number of carriers for each classifier was selected by cross-validation together with the value of the regularization constant, as explained above. The performances of the individual classifiers on the *Lab* dataset, after the variable selection procedure, are presented in Table II, which amounts to a classifier-by-classifier breakdown of the 95% overall score for the *Lab* set given in Table I (column 3). The cross-validation score and the test score are very similar, thereby showing that the classifiers were not over-fitted to the training/validation set, and generalize as expected. Most classifiers exhibit very good performance with a small number of input carriers, except for 3-vs-4 and 4-vs-5, which are somewhat worse. These results are discussed further in the next section.

TABLE II. TEST PERFORMANCES ON *LAB* SET FOR EACH CLASSIFIER

	Room1 vs Room2	Room1 vs Room3	Room1 vs Room4	Room1 vs Room5	Room2 vs Room3	Room2 vs Room4	Room2 vs Room5	Room3 vs Room4	Room3 vs Room5	Room4 vs Room5
Number of input carriers	4	4	4	4	3	4	3	7	4	3
Cross-validation score (%)	98.8	99.6	99.4	99.5	98.9	98.4	99.1	98.8	98.6	95
Test score (%)	100	100	100	100	96.3	98.1	97.9	90	97.2	91.2

V. DISCUSSION

The proposed method has been shown to provide good results in our static, stand-alone tests carried over a period of one month. In this section, we examine this performance in more detail, concentrating on observed failure modes, the temporal stability of the solution, and an interpretation of the variable selection procedure from an engineering practice standpoint. These discussions are based on the *Lab* dataset, using OFR variable selection followed by a linear SVM.

A. Room-by-room breakdown of results

The confusion matrix in table III demonstrates that the deviation of the performance from 100% is dominated by the localizations errors that occur when the acquisition device in room 4 is predicted as having been in room 3 or room 5. This observation is also reflected in the poorer generalization scores, 90% and 91.2% respectively, obtained for the 3-vs-4 and 4-vs-5 classifiers in Table II.

TABLE III. CONFUSION MATRIX

	Predictions				
	Room 1	Room 2	Room 3	Room 4	Room 5
Room 1	100	0	0	0	0
Room 2	0	97	0	0	3
Room 3	0	0	95.2	4,8	0
Room 4	0	0	5.3	84.2	10.5
Room 5	0	0	0	0	100

A local performance loss such as this is problematical. As all GSM carriers have already been included in our study, it seems reasonable to conclude that, in order to further improve performance, additional information will be required. This could take the form of additional RSSI measurements from other frequency bands, for instance, or of complementary data imported from other "imperfect" sensors, such as magnetometers, accelerometers, and the like. Adding memory to the system, to enable the use of dynamic trajectory approaches such as particle filters or Markov models [5, 20], will undoubtedly also be useful.

B. Stability in time

It is well known that RSSI measurements suffer from long term drift caused by seasonal and other environmental factors. Network modifications by the cellphone operator may also be a cause. Tests in ARPEGEO have indeed confirmed that a system trained on RSSIs at a particular date will be practically unusable for prediction six months later if no updates are made to the classifier.

The results presented here have confirmed, however, that system coherence over a time scale of one month is indeed possible. Still, in a real implementation, the system will only have access to measurements that have been made in the past. It is interesting to ask the question, over what time scale can past measurements be used to make good predictions, without retraining the classifier?

To test this, a new training set was created, from the *Lab* dataset, containing the measurements taken in the first three weeks of the one-month period, setting aside the last week as a test set. A classifier system was built as before, using cross-validation and feature selection procedures. Good generalization capability was again observed for most classifiers, and a correct room localization performance of 94% was obtained, to be compared to 95% for the previous system (Table I) in which training and test fingerprints were uniformly distributed in time.

C. Interpretation of selected carriers

In section IV, it was demonstrated that good localization performance can be achieved using fewer than ten carriers per classifier. It is interesting to examine the properties of the carriers that have been selected as being relevant for localization.

Assembling all of the carriers required by the 10 one-vs-one classifiers results in a master set of only 29 "relevant" carriers, out of an original 500 scanned. Of these, 22 carriers were identified as beacons via the BSIC code. This means that about one quarter of the carriers found most relevant for localization (the remaining 7 carriers), *never* exhibited a valid BSIC code in an entire month of data acquisition. Thus, these carriers either are not beacons, or for some reason do not identify themselves as such. One interpretation is that they are traffic channels, which *a priori* were not expected to be useful for localization. In any case, a system basing itself on a pre-selection by BSIC code will suffer the handicap of missing the information supplied by this type of channel.

In order to compare the results of the OFR selection procedure to the more traditional choice of selecting the strongest carriers, we plot, in figure 2, the mean signal strength in dBm versus its standard deviation for all carriers scanned in room 1, where beacons are represented by crosses, and non-beacons (no BSIC) by filled circles. The carriers that were selected as being relevant for discrimination for the Room1-vs-Room2 classifier are at the centers of the four larger, open circles. The "relevant" carriers follow the same general distribution as "non-relevant" carriers, and are concentrated at lower standard deviations. Indeed, it may be due to their smaller variances that these variables are useful for discrimination. It is also clear from the figure that the relevant carriers certainly do *not* tend to be the strongest ones.

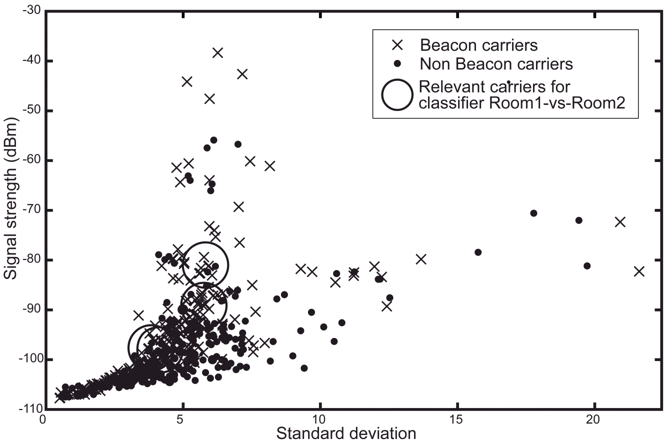

Figure 2. Signal strength and standard deviation of 500 available carriers scanned in room 1.

VI. CONCLUSION

Work performed in the ARPEGEO project has shown that full-band GSM RSSI fingerprints, when analyzed with a statistical learning methodology, provide vastly improved static localization performance as compared to standard fingerprints having much lower carrier counts. Such fingerprints, acquired in idle mode, are available today using trace mobiles or M2M modems at repetition rates of about a minute, and in the near future, should be obtainable on standard cellphone platforms at much higher rates. Variable selection techniques demonstrate that the most relevant carriers for localization purposes tend not to be the strongest carriers, and in some cases fail to be identified as beacons due to non-decoding of a BSIC. The ability to perform localization reliably with our method, using training data taken a few weeks previously, has been demonstrated. A confusion matrix analysis shows that, despite a global room classification performance above 95%, poorer performance (for example, 84.2%) can occur in certain locations, suggesting the need to include other sensors and/or trajectory modeling methods. In this context, full-band GSM fingerprints can be expected to take their place as one component of an "ultimate" indoor localization solution integrating several different technologies.

REFERENCES

[1] P. Meissner, C. Steiner, K. Witrisal, "UWB Positioning with Virtual Anchors and Floor Plan Information," in Proc. of the 7th Workshop on Positioning, Navigation and Communication, Dresden, Germany, 2010.

[2] Q. Yang, S. J. Pan, V. Wenchen Zheng, "Estimating Location Using Wi-Fi," IEEE Intelligent Systems, vol. 23, no. 1, pp. 8–13, 2008.

[3] A. Ofstad, E. Nicholas, R. Szcodronski, and R. R. Choudhury. "AAMPL: Accelerometer Augmented Mobile Phone Localization," In Proc. of International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, San Francisco, USA, 2008.

[4] G. Glanzer, U. Walder, "Self-Contained Indoor Pedestrian Navigation by Means of Human Motion Analysis and Magnetic Field Mapping," in Proc. of the 7th Workshop on Positioning, Navigation and Communication, Dresden, Germany, 2010.

[5] L. Klingbeil, R. Reiner, M. Romanovas, M. Traechtler, Y. Manoli, "Multi-Modal Sensor Data and Information Fusion for Localisatino in Indoor Environments," in Proc. of the 7th Workshop on Positioning, Navigation and Communication, Dresden, Germany, 2010.

[6] F. Ababsa, "Advanced 3D Localization by Fusing Measurements from GPS, Inertial and Vision Sensors," in Proc. of the IEEE international conference on Systems, Man and Cybernetics, San Antonio, USA, 2009, pp. 871–875.

[7] C. Fritsche, and A. Klein, "On the Performance of Hybrid GPS/GSM Mobile Terminal Tracking," in Proc. of the International Conference on Communications, International Workshop on Synergies in Communications and Localization, Dresden, Germany, 2009.

[8] D. Zimmerman, J. Baumann, M. Layh, F. Landstorfer, R. Hoppe, G. Wölfle, "Database Correlation for Positioning of Mobile Terminals in Cellular Networks using Wave Propagation Models," in Proc. IEEE 60th Vehicular Technology Conference, Los Angeles, 2004, vol. 7, pp. 4682–4686.

[9] W. ur Rehman, E. de Lara, S. Saroiu, "CILoS: A CDMA Indoor Localization System," in Proc. of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, 2008.

[10] V. Otsason, A. Varshavsky, A. LaMarca, E. de Lara, "Accurate GSM Indoor Localization," in Proc. of 7th International Conference on Ubiquitous Computing, M. Beigl et al, Eds., pp. 141-158, Springer-Verlag, Berlin, Heidelberg.

[11] B. Denby, Y. Oussar, I. Ahriz, G. Dreyfus, "High-Performance Indoor Localization with Full-Band GSM Fingerprints," in Proc. of the International Conference on Communications, International Workshop on Synergies in Communications and Localization, Dresden, Germany, June 2009.

[12] Tems Mobile System. [Online]: http://www.ericsson.com/solutions/tems/

[13] Telit GM862-GPS module. [Online]: http://www.telit.com/en/products/gsmgprs.php

[14] Tems Pocket System. [Online]: http://www.ascom.com/en/index/products-solutions/our-solutions/product/tems-pocket-3

[15] Scanning with Sony Ericsson TEMS Phones, Ascom Corporation Technical Paper, 2009.

[16] E. Ho, R.L. Kashyap, "An Algorithm for Linear Inequalities and its Applications," IEEE Transactions on Electronic Computers, vol. 14, pp. 683 – 688, 1965.

[17] N. Cristianini, J. Shawe-Taylor, Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000.

[18] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121–167, 1998.

[19] S. Chen, S.A. Billings, W. Luo, "Orthogonal Least Squares Methods and their Application to Non-Linear System Identification," International Journal of Control, vol. 50, pp. 1873-1896, 1989.

[20] J.Seitz, T. Vaupel, J. G. Boronat, J. Thielecke, "A Hidden Markov Model for Pedestrian Navigation," in Proc. of the 7th Workshop on Positioning, Navigation and Communication, Dresden, Germany, 2010.