# Carrier Relevance Study for Indoor Localization Using GSM

Iness Ahriz, Yacine Oussar, Bruce Denby, *Senior Member*, *IEEE*,
Gérard Dreyfus, *Senior Member*, *IEEE*

***Abstract*— A study is made of subsets of relevant GSM carriers for an indoor localization problem. A database was created containing power measurement scans of all available GSM carriers in 5 of 8 rooms of a second storey laboratory in central Paris, France, and a statistical learning algorithm developed to discriminate between rooms based on these carrier strengths. To optimize the system, carrier relevance was ranked using either Orthogonal Forward Regression or Support Vector Machine – Recursive Feature Elimination procedures, and a subset of relevant variables obtained with cross-validation. Results show that the 60 most relevant carriers are sufficient to correctly localize 97% of scans in an independent test set.**

***Index Terms*—Indoor localization, GSM networks, variable selection.**

## I. INTRODUCTION

DESPITE continued research, accurate localization in dense urban and indoor environments remains a difficult task. Fading and mask effects considerably degrade the performance of GPS in such situations. Numerous solutions for indoor localization have been proposed, predominantly based on WiFi [1] or Bluetooth [2] networks. These alternatives however suffer the inconvenience of requiring installation and maintenance of the chosen network by the user.

The widespread availability of GSM networks makes the possibility of their use for localization an attractive alternative. Such a solution can profit from the existing network infrastructure, and, in principle, from the mobile handsets already widely in use. Such methods may be based on network information such as the serving cell location, or on physical information, for example direction of arrival of the signal [3], [4]. These techniques, however, provide limited precision (of the order of 100 meters), and may be further compromised by multipath effects.

Manuscript received December 1st, 2009. This work was supported in part by CNFM (Comité National de Formation en Microélectronique)

I. Ahriz is with the Signal Processing and Machine Learning Laboratory, ESPCI – ParisTech, 10 rue Vauquelin, 75005 Paris, France; (e-mail: iness.ahriz@espci.fr).

Y. Oussar is with ESPCI – ParisTech (e-mail: yacine.oussar@espci.fr).

B. Denby is with Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France, and the Signal Processing and Machine Learning Laboratory, ESPCI – ParisTech (e-mail: denby@ieee.org).

G. Dreyfus is with the Signal Processing and Machine Learning Laboratory, ESPCI – ParisTech (e-mail: gerard.dreyfus@espci.fr).

The database correlation method using GSM fingerprints (power scans over a set of carriers) was presented in [5]. A database is first constructed by obtaining GSM fingerprints at a variety of positions in the area under study. New fingerprints can then be localized based on their "similarity" to certain previously taken measurements. A common approach is to use standard GSM Network Measurement Reports, which contain power measurements of the serving cell and the 6 strongest neighbor cells [5]. The use of fingerprints containing larger numbers of carriers was proposed, for example, in [6].

In [7], accurate indoor localization was obtained using fingerprints including measurements of *all possible* GSM carriers. That work made use of Support Vector Machine, or SVM, classifiers having large numbers of inputs (488 carriers). In the present study, we focus on the part inside the dashed square in the fig. 1 and we examine the question of whether all the GSM carriers used are actually necessary in order to obtain such a performance, or if a much smaller subset containing only the most "relevant" carriers might suffice. This would provide a number of practical benefits, were it to prove the case. A localization system could then be based on the carrier subset, resulting in a simpler implementation and reduced dimensionality for the SVM classifiers. The computational complexity and memory storage requirements of the algorithm would also be reduced, which are important considerations for a mobile platform.

In this study, two algorithms for ranking input variable were tested: Orthogonal Forward Regression, OFR, (using Gram-Schmidt orthogonalization [8]) and SVM Recursive Feature Elimination, or SVM-RFE [9]. The number of relevant carriers to be kept for classification is determined in a cross validation procedure. We shall demonstrate that good discrimination between the rooms in an indoor environment can indeed be obtained using a reduced number of relevant GSM carriers. We also examine the proportion of strongest carriers and beacon carriers in the selected subsets.

The article is organized as follows. The classification method used, based on SVMs, is presented in section II. The two carrier relevance ranking methods are described in section III. Section IV describes the database and the analysis approach adopted. The obtained results are presented and discussed in section V.

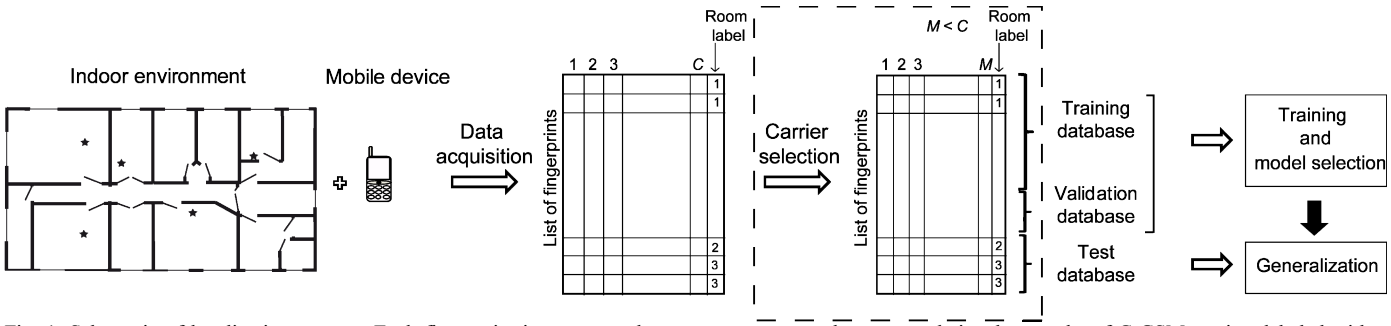


Fig. 1. Schematic of localization process. Each fingerprint is a vector whose components are the measured signal strengths of C GSM carriers labeled with a room number.

## I. SUPPORT VECTOR MACHINES

To discriminate between two classes whose examples are linearly separable, i.e. can be separated without error by a hyperplane, the SVM learning algorithm searches for the maximum margin hyperplane, that is, the hyperplane that separates correctly all the examples of the classes, while being situated as far as possible from the examples of the two classes [10], [11]. In the *M*-dimensional representation space of the items to be classified, a hyperplane has the following equation

$$f(\mathbf{x}) = \sum_{i=1}^{M} w_i x_i + b = \mathbf{w} \cdot \mathbf{x} + b \tag{1}$$

where $\mathbf{x}$ is the vector that describes an item to be classified, and where $\{w_i\}$ and $b$ are parameters that are estimated by training from examples. After training, the label assigned to the item described by $\mathbf{x}$ is +1 if $f(\mathbf{x})$ is positive, and -1 otherwise.

Training can be carried out by minimizing $\|\mathbf{w}\|^2$, under the constraint that all examples are correctly classified. That constrained optimization problem can be implemented in either a primal formulation, in which the estimated parameters are $\{w_i\}$ and $b$, or in a dual formulation. In the latter, a new set of parameters $\alpha_k$ are defined by

$$\mathbf{w} = \sum_{k=1}^{N_T} \alpha_k y_k \mathbf{x}_k \tag{2}$$

where $N_T$ is the number of training examples, $\mathbf{x}_k$ is the vector that describes example $k$, and $y_k = \pm 1$ is the class label of example $k$. In that framework, training consists of solving a quadratic constrained optimization problem with respect to the new variables $\{\alpha_k\}$. The equation of the hyperplane becomes:

$$f(\mathbf{x}) = \sum_{k=1}^{N_T} \alpha_k y_k (\mathbf{x}_k \cdot \mathbf{x}) + b \tag{3}$$

If the examples are not linearly separable, the constrained optimization problem has no solution. In such cases, a transformation of the variables is sought, such that, in the resulting new representation space, the examples are linearly separable. As a result of that transformation, the separation surface, in the original representation space, is no longer a hyperplane; its equation becomes

$$f(\mathbf{x}) = \sum_{k=1}^{N_T} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) + b \tag{4}$$

where $K(\cdot, \cdot)$ is an appropriate nonlinear function called a kernel function. In the framework that has been described, no training example is allowed to sit within the margin. The resulting classifiers are called "hard-margin" classifiers; otherwise (soft-margin SVM), the position of the separating surface is a compromise between the width of the margin and the number of examples that are within the margin. This compromise is implemented via a regularization control hyperparameter.

The classifiers described above are intended to separate two classes. For multiclass classification problems such as the localization problem addressed in the present article, the so-called "one versus one" strategy was chosen. This approach is used under the hypothesis that pairwise separation of classes will be simpler than separating each class individually from the others (the "one versus rest" strategy). To separate $P$ classes, $P(P\text{-}1)/2$ classifiers are necessary, where each classifier is intended to separate two of the $P$ classes. To classify a test example, it is presented to the $P(P\text{-}1)/2$ classifiers and a voting system attributes to the example the label most frequently appearing over the ensemble of classifiers. In our study, where the number of classes $P$ is 5, the localization system is composed of 10 classifiers.

## II. VARIABLE RELEVANCE RANKING METHODS

### A. Orthogonal forward regression, OFR

Orthogonal forward regression using Gram-Schmidt orthogonalization allows to rank variables (here, carriers) by order of their relevance. The method is constructive in that it begins with an empty set, to which relevant variables are added iteratively. The degree of relevance of a variable is estimated by calculating the squared cosine of the angle between a vector composed of the measured values of the considered variable and the measured vector of outputs (i.e., the position labels) in the space of observations

$$\cos^2(\mathbf{x}_k, \mathbf{y}) = (\mathbf{x}_k \cdot \mathbf{y})^2 / (\mathbf{x}_k \cdot \mathbf{x}_k)(\mathbf{y} \cdot \mathbf{y}) \quad (5)$$

where $\mathbf{x}_k$ is the vector of the measured values of the $k$-th variable, and $\mathbf{y}$ is the vector of the measured values of the quantity of interest; in the present case (classification problem), the components of $\mathbf{y}$ take on values -1 or +1.

The variable which exhibits the greatest value of the cosine squared forms the smallest angle with, and thus best "explains", the quantity of interest. It is therefore considered the most relevant variable. The vectors of the remaining variables, as well as the output vector, are projected onto the subspace which is orthogonal to the vector of the selected variable in order to eliminate components parallel to it. The calculation of the squared cosine is then repeated in this new space in order to determine the second most relevant variable. The procedure is iterated until all variables have been ranked. This approach assumes a model which is linear in its parameters. In the case of a nonlinear model, polynomial ranking (polynomial model of degree 2 for example), can be envisioned.

### B. *Support Vector Machines - Recursive Feature Elimination (SVM - RFE)*

SVM-RFE is a ranking method intended for the design of SVM classifiers [9]. Its ranking criterion is the change in the cost function that occurs when a variable is withdrawn from the model.

As discussed in [9], and since we use linear SVM classifiers whose cost function J is quadratic with respect to parameters $w_i$, withdrawing variable $i$ from the set of variables of the model results in a variation $\Delta J(i)$ of the cost function that is proportional to the squared weight $w_i$:

$$\Delta J(i) \propto w_i^2 \quad (6)$$

Thus, the ranking criterion becomes the magnitude of the weight $w_i$. As a result, the variables weighted with the smallest values $w_i$ can be considered as the least relevant ones.

Contrary to orthogonal forward regression, SVM-RFE is a "backward" approach, in that it starts with the full set of variables and iteratively removes the less relevant ones. Let **r** be the ordered vector of variables, which initially is empty. On the first iteration, all variables are used to train the SVM classifier. For each variable $i$, a weight $w_i$ is calculated. The least relevant variable is the variable whose weight $w_i$ is smallest. This variable is removed and placed into the vector **r**. The procedure is iterated until all carriers are ranked in ascending order of relevance in **r**. It should be noted that this method requires as many training sessions as the number of variables, making it much more costly numerically than the OFR procedure. The method does however have the advantage of being specifically designed for use with SVMs, which we have adopted for our localization system.

## III. EXPERIMENTAL TECHNIQUE

In this section we describe the data acquisition procedure and the technique adopted to perform the localization.

### A. *Data acquisition*

Measurements of the radio environment were carried out over a period of one month in 5 of 8 rooms of a second-floor laboratory in central Paris, France, using the TELIT GM-862 modem [12]. The laboratory layout and the points where measurements were taken are indicated in fig. 2.

The TELIT GM-862 is capable of measuring carrier power over the entire GSM band, for a total of 548 channels. The device reports the ARFCN (Absolute Radio Frequency Channel Number) and RXLEV (Received Signal Level) for every carrier detected. If a channel is determined to be a beacon channel, or BCCH (Broadcast Control Channel), the BSIC (Base Station Identity Code) is also returned when possible. In our data, 534 different carriers were detected, of which 234 were beacons. Carriers are detected only if their power is above a threshold of -108 dBm. Our database consists of 601 measurements, with an equal number of measurements in each of the 5 rooms indicated in the fig. 2. For this study, the measuring device was always placed at the same position in each room, indicated by the star symbols in the figure. Furthermore, to simplify the analysis, only the RXLEV values, and not the BSIC codes, were used as inputs to the localization system.

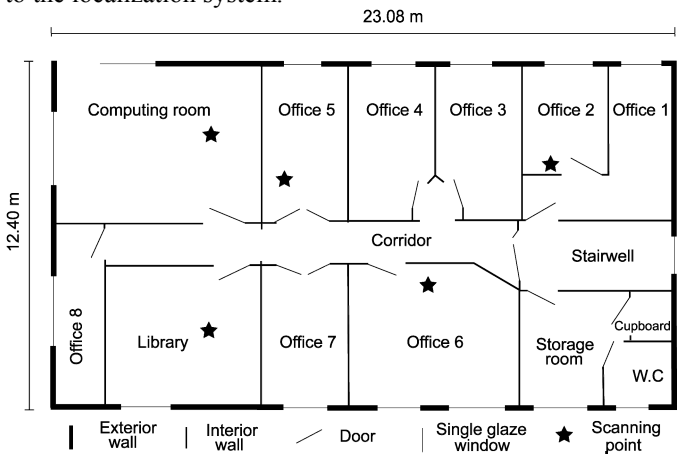


Fig. 2. Laboratory floor plan.

### B. *Localization technique*

The problem undertaken in this article is to discriminate between 5 rooms of an indoor environment using only the most relevant GSM carriers. It is thus a classification problem, with rooms identified as classes, which can be broken down into subproblems each involving only two of the classes. The solution thus requires the use of a system of 10 "one versus one" classifiers, each used to separate one pair of classes. A voting mechanism is then applied, in which the output class label is that receiving the largest number of "votes".

By the arguments in section III, a ranking by relevance of the input variables of a classifier will be applicable only to that classifier. The procedure adopted here was to carry out a

ranking for each of the 10 classifiers. The set of carriers $I$ appearing at the input of the final classifier was defined as

$$I = \bigcup_{j=1}^{10} I_j^N \tag{7}$$

where $I_j^N$ is the set of the $N$ most relevant carriers for the classifier $j$. Therefore, the number of relevant carriers taken into account by the classifier is the cardinal number of $I$, which depends on $N$: $M(N) = |I|$. This approach guarantees that each classifier will always find its own $N$ most relevant carriers among the inputs. A potential drawback of the method is that each classifier will also see the carriers needed by the others, but our tests demonstrated that the additional carriers did not lead to any degradation in performance as compared to a system in which each classifier had only its own $N$ most relevant carriers at the input. In addition, the approach allows for a greatly simplified implementation. An overview of the localization system is shown in fig. 3.

The number of carriers contained in the set $I$ as a function of $N$ is represented in fig. 4. We do not consider values of $N$ above 10 since, as shown in fig. 5, the system performance does not improve above $N = 10$.

Application of the Ho - Kashyap algorithm [13] showed that all training examples were pairwise linearly separable for $N \geq 3$. For $N = 1$ and $N = 2$, nonlinear classifiers with a Gaussian kernel $K$ were used.

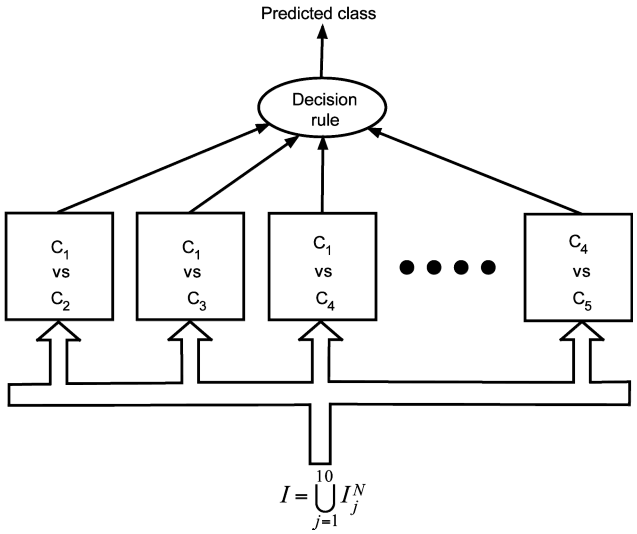


Fig. 3. Localization system composed of 10 "one versus one" classifiers.

During the training phase, the choice of SVM hyperparameters (regularization parameter and Gaussian kernel parameter for nonlinear classifiers), as well as the value of $N$, were obtained in a 10-fold cross-validation procedure based on a set of 500 measurements drawn randomly from the 601 available measurements. Once the best validation score was obtained, the corresponding value of $N$ was considered as the most appropriate given the available data, and the number of variables $M(N) = |I|$ was computed. A final SVM is trained, with $M(N)$ variables, on the whole set of 500 measurements. The remaining 101 measurements were subsequently used as a test set for estimating the classification performance. Both linear and Gaussian SVMs were implemented using *The Spider* toolbox [14].

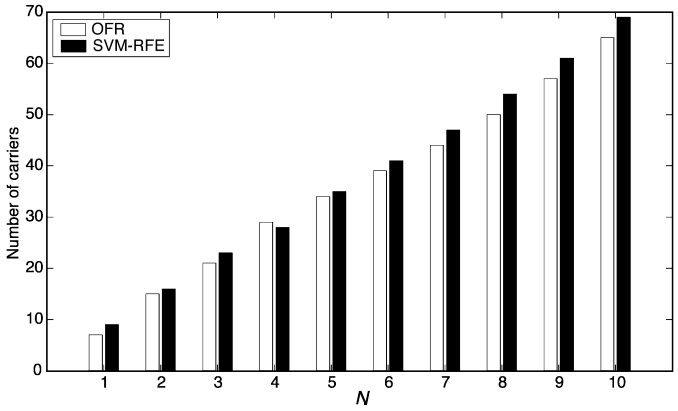


Fig. 4. Number of carriers $M(N) = |I|$ taken into account by the classifier as a function of $N$.

## IV. Results and Discussions

### A. Ranking and classification results

The cross-validation scores of the 10 classifier system, using ranking by orthogonal forward regression or SVM-RFE, are shown in fig. 4. Also shown in the figure, for comparison, are scores obtained when the input carriers are ranked simply by order of signal power. The number of carriers retained in this case is taken as the value of $M(N)$ obtained with the OFR procedure (OFR and SVM-RFE can give slightly different values of $|I|$, as shown on Fig. 3; for example, for $N = 10$, $M(N) = |I| = 65$ for OFR, while $M(N) = |I| = 69$ for SVM-RFE).

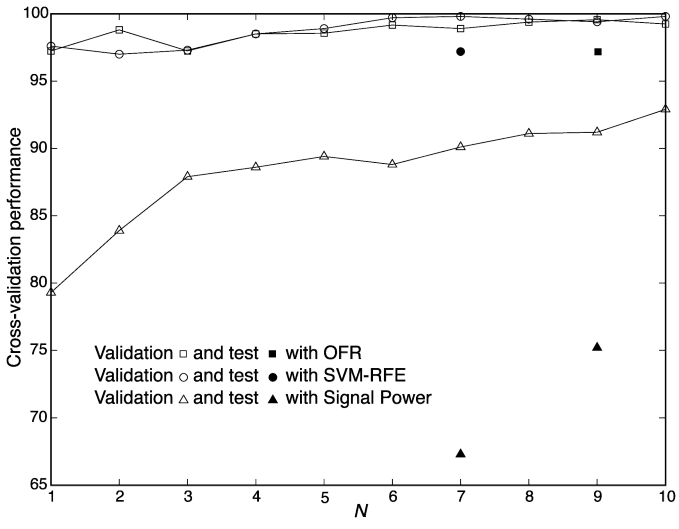


Fig. 5. Validation (symbols and lines) and test set (symbols) performance as a function of $N$.

The cross-validation scores obtained using the most relevant carriers ranked by OFR or SVM-RFE are far superior to those obtained with equivalent numbers of carriers ranked by signal

power. For example, a validation score of 99.4% is obtained by OFR relevance ranking with $N = 9$, and 99.8% by SVM-RFE with $N = 7$, while such levels of performance are never achieved using carriers ranked by signal power in the range of values of $M$ investigated here. The performance of the classifier giving this optimum validation score was then estimated on the test set. The results are indicated by the symbols in fig. 5, and summarized in table I.

TABLE I
TEST SET PERFORMANCE USING SELECTED CARRIERS

| Ranking method | $N$ | Number of carriers $M$ | Test set performance |
|---|---|---|---|
| OFR | 9 | 57 | **97%** |
| SVM-RFE | 7 | 47 | **97%** |
| Signal power level | 9 | 57 | 75.2% |
| | 7 | 47 | 67.3% |

As in the case of the validation, the test set results obtained using relevance carrier ranking are much better than those achieved using simple signal power ranking. Apparently the set of relevant carriers contains more information which is useful for discriminating between classes than does the set selected on signal power alone. Put another way, the carrier set selected on power does not contain enough of the relevant carriers to provide for good localization. This hypothesis is further supported by the result shown in fig. 6, which shows the fraction of relevant carriers contained in the power-ranked carrier set, as a function of $N$, for OFR ($N = 9$) and for SVM-RFE ($N = 7$). The curves show that in order to include the 57 (OFR ranking) or 47 (SVM-RFE ranking) most relevant carriers, it is necessary to take the strongest 470 carriers ranked on signal power. This explains the poor validation performance using the 57 (or 47) strongest carriers, which include only 20% of the most relevant carriers. These plots also support the result presented [7], which states that accurate GSM localization requires measuring all available carriers. Our results show that certain carriers of low power are nonetheless relevant, and necessary for good indoor localization performance.

### B. *Importance of beacon channels*

We may also examine the role of beacon channels in the set of carriers which are relevant for localization. In our study, these are identified by the presence of a BSIC.

Fig. 7 displays the percentage of beacon channels in the set of relevant carriers for different values of $N$. These percentages must be interpreted as lower bounds, since the absence of a BSIC does not guarantee that a channel is not a beacon (low power, multipath effects, etc., may also play a role in the non-detection of a BSIC).

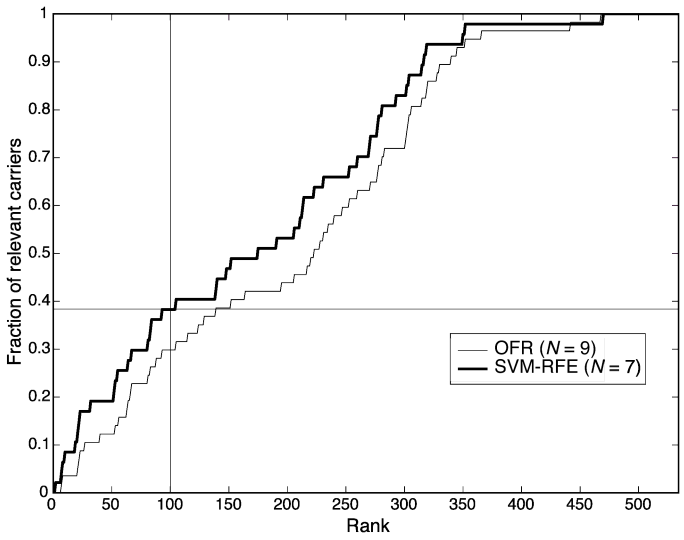


Fig. 6. Cumulative percentage of relevant carriers as a function of their rank in the power ranking of all available carriers, for OFR and SVM-RFE relevance ranking methods. For instance, the graph shows that approximately 40% of the relevant carriers selected by SVM-RFE are among the 100 strongest carriers.

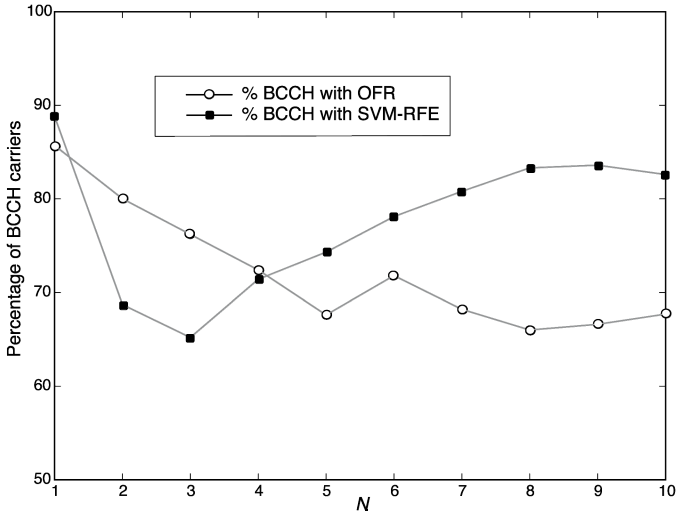


Fig. 7. Percentage of beacon channels in the set of relevant carriers, for several values of $N$ and for the two ranking methods.

The curves in fig. 7 show that at least 2/3 of the carriers selected by OFR or SVM-RFE are indeed beacon channels. This result implies that, beyond being crucial for the functioning of a GSM communications network, beacon channels are also valuable for localization. This is, to a certain extent, to be expected, since beacon channels are emitted at constant power. However, the power of a beacon is not a valid relevance criterion for localization.

## V. CONCLUSION

In this study, several methods of ranking GSM carriers, by relevance or by signal power, have been tested in an indoor environment, with the goal of identifying the room in which a mobile terminal is located. The study has allowed us to identify, among the hundreds of carriers detected, those which are the most relevant for this type of localization.

The results obtained show that a subset of some 60 carriers among the 534 measured permit to obtain very good localization performance. Analysis of these relevant carriers has demonstrated that they are not simply the strongest ones. The OFR and SVM-RFE methods have shown that, contrary to a communications scenario where signal power (relative to noise) is the critical parameter, algorithm development and variable selection for localization do not necessarily depend upon the availability the strongest carrier signals only. Nevertheless beacon channels, emitted at constant power, are predominant in the set of carriers found to be most relevant for the purposes of localization.

Tests on larger data sets will be necessary in order to further validate our localization approach. To this end, our laboratory has constructed a set of independent measuring devices which can be operated in parallel in order to position-label carrier power scans with a minimum of human intervention. Localization on an x-y grid, with a finer position resolution, making use of the ranking and classification tools outlined in the present article, are also envisioned.

## REFERENCES


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

[2] L. Aalto, N. Gothlin, J. Korhonen, T. Ojala, "Bluetooth and WAP push based location-aware mobile advertising system", in *Proc. 2nd International Conference on Mobile Systems, Applications, and Services*, Boston , 2004, pp. 49–58.

[3] B. Denby, Y. Oussar, I. Ahriz, "Geolocalisation in Cellular Telephone Networks", in *Proc. NATO Advanced Study Institute on Mining Massive Data Sets for Security*, Gazzada, 2007, F. Fogelman-Soulié, D. Perrotta, J. Piskorski & R. Steinberger, Eds., IOS Press, pp. 357–365, Amsterdam, Netherlands.

[4] H. Laitinen et al., "Cellular Location Technology", internal report of EU IST project "Cellular network optimization based on mobile location", available from http://www.telecom.ntua.gr/cello/documents/CELLO-WP2-VTT-D03-007-Int.pdf, October 2001.

[5] 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.

[6] V. Otsason, A. Varshavsky, A. LaMarca, E. de Lara, "Accurate GSM indoor localization", in *Proc. 7th International Conference on Ubiquitous Computing,* Tokyo*,* 2005, M. Beigl et al., Eds., pp. 141-158, Springer-Verlag, Berlin, Heidelberg.

[7] B. Denby, Y. Oussar, I. Ahriz, G. Dreyfus, "High-Performance Indoor Localization with Full-Band GSM Fingerprints", in *Proc. IEEE International Conference on Communications*, SyCoLo Workshop, Dresden, 2009.

[8] 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.

[9] I. Guyon, J. Weston, S. Barnhill, M.D, V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines"*, Machine Learning*, vol. 46, no 1–3, pp. 389–422, 2002.

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

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

[12] Telit GM862 module [Online]: http://www.gm862.com/en/products/gsm-gprs.php

[13] Y.-C. Ho, R.L. Kashyap, "An algorithm for linear inequalities and its applications", *IEEE Trans. Elec. Comp*, vol. 14, No. 5, pp. 683–688, 1965.

[14] The Spider. [Online]: www.kyb.tuebingen.mpg.de/bs/people/spider/