Application des techniques d'apprentissage à la géolocalisation par radio fingerprint

Iness Ahriz Roula

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Électronique

Présentée par

Mme Iness AHRIZ ROULA

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Application des techniques d'apprentissage à la géolocalisation par *radio fingerprint*

soutenue le 1er Décembre 2010

devant le jury composé de :

M. Bruce DENBY	Professeur	Directeur de thèse
M. Giuseppe ABREU	Professeur	Rapporteurs
Mme. Geneviève BAUDOIN	Professeur	
M. Aziz BENLARBI	Professeur	Examinateurs
M. Richard DUSSEAUX	Professeur	
M. Emmanuel VIENNET	Professeur	
M. Gérard DREYFUS	Professeur	

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 44 27 28 10
Fax : 01 44 27 23 95
Tél. pour les étudiants de A à EL : 01 44 27 28 07
Tél. pour les étudiants de EM à MON : 01 44 27 28 05
Tél. pour les étudiants de MOO à Z : 01 44 27 28 02
E-mail : scolarite.doctorat@upmc.fr

Remerciements

Je tiens en tout premier lieu à remercier Monsieur Bruce Denby pour avoir dirigé cette thèse et dont l'aide précieuse m'a été indispensable sur le plan scientifique. Je tiens également à le remercier pour la confiance et la sympathie qu'il m'a témoignées au cours de ces trois années de thèse.

Je remercie également Monsieur Gérard Dreyfus, directeur du laboratoire SIGMA de l'ESPCI, pour m'avoir accueilli au sein de son équipe.

J'adresse également mes remerciements à Madame Genvième Baudoin et à Monsieur Giuseppe Abreu pour avoir accepté d'être les rapporteurs de cette thèse, ainsi qu'à Messieurs Aziz Benlarbi-Delaï, Richard Dusséaux et Emmanuel Viennet de l'avoir examinée.

Pendant une partie de me thèse j'ai eu la chance de travailler avec Messieurs Pierre Roussel et Rémi Dubois, Maîtres de conférences au laboratoire SIGMA, je tiens à les remercier pour leurs encouragements et soutien constant. Je remercie également tous les autres membres du laboratoire SIGMA pour leur accueil et leur bonne humeur quotidienne.

Un grand merci à tous mes amis, qui se reconnaitront, de m'avoir soutenu et supporté durant ces trois années de thèse.

Parce qu'elle représente beaucoup pour moi, je tiens à exprimer ma profonde gratitude à ma mère qui m'a accompagné, soutenu et encouragé tout au long de mon parcours, ainsi qu'à tous les membres de ma famille qui ont toujours cru en moi.

Enfin, merci à mon époux Adlene qui m'a toujours donné le coup de pouce dont j'ai besoin dans les moments difficiles.

Résumé

L'objectif de la thèse est d'améliorer la précision de localisation des personnes (ou objets) dans les environnements où le signal GPS est faible, voire inexistant, par exemple à l'intérieur de bâtiments.

L'originalité de ce travail est d'utiliser des signaux radio reçus de sources extérieures pour effectuer la localisation à l'intérieur des bâtiments. Durant ce travail, des mesures des puissances reçues ont été effectuées dans diverses conditions. Étant données les difficultés liées aux processus expérimentaux (incertitude et bruit liés aux mesures), ce dernier point a constitué une première partie de ce travail de thèse. Le traitement et l'interprétation des mesures des puissances de signaux radio dans un environnement de propagation instable ont été un véritable défi durant cette étude. En effet, toutes les porteuses du réseau GSM ont été considérées et aucune hypothèse n'a été posée, à priori, sur leurs pertinences. Une seconde partie de la thèse a été justement consacrée à l'estimation de ces pertinences. Le problème de la détermination de la pièce dans laquelle se trouve le mobile a été considéré comme un problème de classification automatique : des méthodes d'apprentissage statistique (supervisé et semi-supervisé) ont donc été mises en œuvre. Le choix des méthodes utilisées a été fait sur la base des études menées sur les mesures des puissances.

Des performances très satisfaisantes ont été obtenues, dans les pièces de deux bâtiments différents. Ces résultats ont ainsi confirmé l'apport des méthodes d'apprentissage statistique au problème de localisation par *fingerprints*.

Mots clés

localisation en intérieur, *fingerprints* GSM, apprentissage statistique.

Abstract

The objective of this thesis is to improve the localization precision in indoor environments where GPS signal is weak or non-existent.

The originality of the research lies in the design of indoor localization system based on radio signals received from external sources. In this work, real measurements of received GSM signals were used. Given the difficulties associated with these experimental procedures (uncertainty and noise related to the measurements), the first part of the thesis is dedicated to their description. The processing and interpretation of radio GSM measurements in a real propagation environment were a challenge in this study. Indeed, all carriers of the GSM network were considered with no a priori hypothesis made regarding their relevance. The second part of this thesis then describes a study of relevance, as well as the statistical learning approaches (supervised and semi-supervised) which were implemented to predict positions from the measurements. A choice of learning method was made based on the studies conducted on these measurements.

The promising room-level localization performance reached in this thesis demonstrates clearly that good quality indoor localization can be obtained by applying a machine learning strategy to GSM radio fingerprints.

Keywords

Indoor localization, GSM fingerprints, machine learning.

Table des matières

Liste des figures

Introduction générale

Contexte de la thèse

L'utilisation étendue des réseaux téléphoniques sans fil a donné naissance naturellement à l'idée de fournir des services en fonction de la position de l'utilisateur : un besoin de localisation est donc apparu. L'information sur la position a été, dans un premier temps, utilisée pour la sécurité des usagers (par exemple, l'assistance aux personnes en danger). Elle a, ensuite, servi pour une utilisation optimale des services proposés sur le réseau tels que l'aide à la navigation. La qualité de ces services est étroitement liée à la précision de positionnement. Un intérêt considérable s'est alors développé au sein des communautés scientifique et industrielle pour les techniques de localisation.

Les méthodes de localisation ont été l'objet d'encore plus d'attention quand la *United States Federal Communications Commission* (FCC) a exigé que les opérateurs de téléphonie mobile soient capables de localiser leurs abonnés appelant le numéro d'urgence 911 avec une erreur de 50 mètres dans 67% des cas. Les mêmes exigences de sécurité ont également été adoptées par l'Europe concernant les appels vers le numéro d'urgence 112. Cette nécessité légale de pouvoir localiser les utilisateurs a donné naissance à de nombreuses applications.

Grâce à la localisation des utilisateurs, les opérateurs de téléphonie mobile ont pu proposer des tarifs préférentiels en fonction de la position du mobile. C'est le cas pour l'offre « Happy Zone » de SFR, où les utilisateurs peuvent bénéficier d'appels illimités vers les fixes quand ils sont à leurs domiciles. Des systèmes de transport « intelligents » [1] ont aussi été proposés. Outre l'aide à la navigation et les services de proximité, de tels systèmes permettent la gestion du trafic routier (en particulier en cas d'accident de la route), le paiement électronique (péage urbain à Londres, par exemple), une assistance à la conduite pour améliorer la sécurité des usagers (on peut citer les systèmes de limitation de vitesse développés dans le cadre du projet LAVIA - Limiteur s'Adaptant à la VItesse Autorisée) [2], la gestion de flotte pour les entreprises de transport, etc. Ces applications sont déjà déployées et reposent, pour la plupart, sur le *Global Positioning System* (GPS). Cette technologie fournit une précision de localisation suffisante pour les applications citées, dans les milieux dégagés (de l'ordre de 10 à 20 mètres). Cependant, pour la localisation des utilisateurs dans des milieux urbains denses ou à l'intérieur de bâtiments (résidences, aéroports, centres

commerciaux, musées, hôpitaux) l'efficacité du GPS est très limitée. Des méthodes de localisations plus adaptées sont donc nécessaires pour ces environnements.

Diverses méthodes permettent d'estimer la position des appareils mobiles à partir des signaux reçus des réseaux auxquels ils sont rattachés. À partir de la puissance du signal reçu, du temps de propagation ou de l'angle d'incidence, une triangulation permet de localiser le mobile. Néanmoins, pour des applications en environnements intérieurs, qui ont pour objet la santé et la sécurité des personnes, une localisation plus précise est requise ; c'est le cas notamment du suivi des personnes âgées ou atteintes de troubles cognitifs. Un système de localisation performant améliore l'autonomie de ces personnes tout en permettant une assistance en cas de besoin.

Cette thèse propose une méthode nouvelle de localisation de récepteurs GSM à l'intérieur des bâtiments, reposant sur l'application de l'apprentissage statistique aux puissances des signaux GSM, dont les résultats sont très prometteurs.

Organisation du manuscrit

Le manuscrit comprend quatre parties. Dans le premier chapitre, les méthodes de localisation existantes sont présentées, et leurs avantages et inconvénients respectifs sont discutés. Une présentation du réseau GSM ainsi qu'une étude des signaux reçus par les utilisateurs de ce réseau font l'objet du deuxième chapitre. La troisième partie est consacrée à la présentation du système de localisation et des techniques mises en œuvre. Les performances de ce système sont exposées et discutées dans le dernier chapitre.

Chapitre 1. Localisation des personnes et objets mobiles

1.1. Introduction

Avec le développement des réseaux mobiles, un marché de *Location Based Services* (LBS) [3] a fait son apparition. Ces services fournissent des informations qui dépendent de la position de l'utilisateur, telles que les observations ou prévisions météorologiques, le trafic routier, etc. La particularité des LBS est que l'utilisateur mobile est automatiquement localisé et n'a pas besoin d'indiquer lui-même sa position. Une forte activité de recherche sur les méthodes de localisation s'est alors développée, l'objectif étant de fournir une bonne précision de localisation afin d'assurer une bonne qualité de ces services. La localisation d'un objet ou d'une personne revient à trouver sa position dans un espace de coordonnées cartésiennes, ou simplement situer le mobile dans un environnement défini (une rue, une pièce ou un étage par exemple) ; le choix dépend des applications visées et des techniques de localisation utilisées.

De nombreuses solutions pour la localisation des personnes (et objets) ont été proposées. Elles devaient répondre à des exigences de coût, de complexité et de précision. La plus connue des solutions est l'utilisation des récepteurs GPS. La précision de 10 à 20 mètres qu'offre le GPS dans les milieux dégagés a permis un large déploiement des services de localisation en extérieur (l'aide à la navigation ou recherche d'itinéraire). Cependant, l'extension de ces services en intérieur reste limitée à cause de la faible précision des méthodes de localisation dans ces environnements. Afin de se libérer des limitations du système GPS dans les environnements urbains denses ou intérieurs, des solutions reposant sur des capteurs ou des réseaux sans fils ont fait leur apparition ; nous allons en présenter quelques exemples.

Des études de géophysique ont montré que l'intensité et la direction du champ magnétique terrestre dépendent de l'endroit où l'on se trouve. Cette propriété a été reprise en géolocalisation pour caractériser chaque position par l'amplitude de ce champ magnétique. Une telle approche consiste à enregistrer une « carte magnétique » de la zone étudiée, associant chaque position à l'intensité du champ mesurée. Un mobile souhaitant se localiser mesure, à l'aide d'un magnétomètre, l'intensité du champ magnétique, qui est

comparée aux enregistrements pour déduire sa position [4, 5]. L'utilisation de cette technique de localisation à l'intérieur des bâtiments reste limitée par les fluctuations du champ magnétique dues aux systèmes électriques, appareils électroniques, matériaux de la structure des bâtiments, etc.

L'estimation de la position par une double intégration de l'accélération peut aussi être envisagée comme solution de localisation. Ceci est possible si le mobile à localiser est en déplacement et possède donc une accélération. À partir des accélérations mesurées sur trois axes par des accéléromètres [6] une localisation en trois dimensions est possible. La précision de la localisation est néanmoins limitée par le bruit sur la mesure de l'accélération.

La possibilité de localisation par capteurs infrarouges et ultrasonores a aussi été explorée [7, 8]. Dans le cas de l'infrarouge, un badge porté par l'utilisateur émet périodiquement un signal qui est reçu par des capteurs installés à des positions connues. Ce signal permet de repérer la position de l'utilisateur par rapport à un point fixe. Dans le cas des ultrasons, le mobile peut aussi estimer sa position à partir du temps de propagation d'un signal provenant d'émetteurs ultrasonores dont les positions sont connues. Grâce à la courte portée des signaux infrarouges et ultrasonores, ces solutions offrent une précision de localisation acceptable. Cependant, leur utilisation ne peut être envisagée pour une large couverture en raison du coût d'installation d'un réseau de capteurs.

Enfin, la connaissance visuelle de l'environnement est également une piste pour la localisation des mobiles. Des images [9] sont enregistrées pour caractériser l'environnement des positions connues. L'image enregistrée par le mobile à localiser est ensuite comparée à l'ensemble des images préalablement enregistrées afin de prédire sa position.

Les ondes radio échangées sur des réseaux *Radio Frequency Identification* (RFID), Bluetooth, Zigbee ont fait l'objet de nombreuses études. Nous discuterons ces techniques de manière détaillée dans la section 1.2.2. Malgré la bonne précision de telles méthodes, l'installation d'un réseau de capteurs dédié à la localisation reste nécessaire. Cet inconvénient peut être contourné en mettant à profit les ondes radio des réseaux existants. En effet, un réseau déjà installé, entretenu et largement utilisé, tel que WLAN ou GSM, constitue l'infrastructure adéquate et la moins coûteuse pour les techniques de localisation. Dans ce mémoire, nous proposons de nouvelles méthodes de localisation fondées sur la réception de

ces ondes en l'absence d'informations exogènes telles que la cartographie, la vitesse ou la direction de déplacement du récepteur[1].

Dans la suite de ce chapitre, une première partie est consacrée au principe de fonctionnement du système GPS adopté pour la localisation extérieure : nous mettons en évidence les principales raisons pour lesquelles cette technologie est inadaptée à la localisation en intérieur. Une seconde partie traite les méthodes fondées sur les ondes radio, utilisées dans le cadre de la thèse, pour la localisation en intérieur. Selon la technique utilisée pour déduire la position à partir des ondes, on distingue trois catégories : localisation par proximité, localisation par calcul de distance, ou localisation par *fingerprints*. Le principe de chacune de ces catégories est décrit dans cette seconde partie.

1.2. Méthodes de localisation

1.2.1. Localisation par satellites

Le système de localisation par satellite le plus répandu est sans doute le *Global Positioning System* (GPS). Développé et entretenu par le *Department of Defense* (DoD) américain à des fins militaires, le GPS est également utilisé pour des applications civiles : aide à la navigation, gestion du trafic, localisation des téléphones portables dans des milieux dégagés, etc. Une autre application intéressante mais moins connue de ce système est la synchronisation en temps grâce aux horloges atomiques de grande précision embarquées dans les satellites.

Le système GPS est constitué de trois « secteurs ».

- Le secteur « spatial » : il est constitué d'une constellation de 24 satellites, situés à une altitude de 20 000 kilomètres, en mouvement sur six plans orbitaux inclinés de 55 degrés par rapport à l'équateur. Cette architecture assure qu'au moins quatre satellites sont visibles à tout instant en tout point du globe terrestre : c'est le nombre minimal de satellites nécessaire pour une bonne précision de

[1] Quand des informations exogènes sont disponibles, dans le cas du suivi de personnes, où la notion de trajet existe, les méthodes de filtrage particulaire[8], les modèles de Markov cachés [10] et les filtres de Kalman [11], etc. ont été mises en œuvre avec des résultats satisfaisants. Ces outils permettent une modélisation dynamique des trajets et ainsi la prédiction de la position à partir des positions précédentes et de grandeurs telles que la vitesse et la direction de déplacement. Pour les raisons qui seront exposées dans le chapitre 2, ces méthodes ne peuvent pas être appliquées à nos données.

localisation en trois dimensions (latitude, longitude et altitude). En pratique, le nombre de satellites visibles peut atteindre dix.

- Le secteur « au sol » permet le contrôle des satellites ; il est constitué d'une station principale située à Colorado Springs et de quatre stations de contrôle situées à Hawaii, Ascencion, Diego Garcia et Kwajalein. Ces stations effectuent des enregistrements continus des signaux GPS, et des mesures météorologiques. Ces données sont traitées par la station principale pour le calcul des paramètres qui constituent les signaux de navigation pour chaque satellite. Ces messages de navigation (définis ci-dessous) sont ensuite transmis au satellite en question.
- Pendant l'exploitation du GPS, ces messages de navigation, appelés aussi éphémérides, sont envoyés par les satellites au secteur « utilisateur » qui est constitué de l'ensemble des récepteurs GPS civils et militaires. Cet échange se fait sur les fréquences L1 = 1575.42 MHz et L2 = 1227.6 MHz, qui ont été choisies pour compenser les effets atmosphériques sur l'onde électromagnétique. Cependant, l'effet Doppler dû au mouvement des récepteurs par rapport aux satellites nécessite une phase de synchronisation afin de décoder les messages de navigation. Ce temps de synchronisation est appelé *Time To First Fix* (TTFF).

Les messages de navigation sont des signaux à spectre étalé par des codes pseudo-aléatoires. Cette technique permet d'envoyer des messages simultanément sur la même bande de fréquence. Chaque message est identifié grâce au code pseudo aléatoire qui le module. Deux types de codes existent, le code C/A ou *Coarse/Acquisition code* pour les applications civiles et le code P ou *Precision code* pour les applications militaires. Cet étalement du spectre assure également une protection contre le bruit et le brouillage, et permet au système de fonctionner avec des niveaux de puissances très bas, de l'ordre de -160 dBW.

Le message de navigation de chaque satellite contient :

- les éléments képlériens qui permettent de définir sa position dans un repère terrestre à l'instant de l'observation,
- les coefficients du modèle ionosphérique permettant de corriger l'effet de l'ionosphère sur les signaux envoyés par le satellite,
- l'état de santé du satellite,

- un modèle polynomial qui caractérise le fonctionnement de l'horloge,
- l'écart entre le temps GPS et le temps universel coordonné de l'US *Naval Observatory*,
- les codes C/A (pour la fréquence L1) et un code P (pour les fréquences L1 et L2).

Pour le calcul de sa position, le récepteur GPS met en œuvre la méthode du temps d'arrivée du signal, qui consiste à calculer la position d'un point en trois dimensions à partir des distances qui le séparent d'un certain nombre de satellites dont les positions sont connues. Ces distances sont calculées à partir du temps Δt de propagation du signal entre les satellites et le récepteur GPS. Ce temps de propagation est estimé en générant une réplique du code C/A (ou P) au niveau du récepteur et en retardant cette réplique jusqu'à l'alignement entre le code reçu et le code généré.

Les distances ainsi obtenues sont appelées pseudo-distances en raison du décalage entre les horloges des satellites et celle du récepteur GPS. La distance qui sépare un récepteur et un satellite *i* s'écrit sous la forme :

$$D_{R-S_i} = c.\Delta t = \sqrt{\left(X_{S_i} - X_R\right)^2 + \left(Y_{S_i} - Y_R\right)^2 + \left(Z_{S_i} - Z_R\right)^2} + c.\delta t \qquad (1.1)$$

où (X_R, Y_R, Z_R) sont les coordonnées (connues) du satellite dans le repère géocentrique, (X_{Si}, Y_{Si}, Z_{Si}) sont les coordonnées (inconnues) du récepteur GPS dans le même repère, δt est le décalage temporel (inconnu) entre l'horloge du satellite et celle du récepteur, $c = 3*10^8$ m/s est la vitesse de la lumière et Δt est le temps de propagation de l'onde entre le satellite et le récepteur. En supposant que les horloges des différents satellites sont synchronisées, les quatre inconnues à déterminer sont X_{Si}, Y_{Si}, Z_{Si} et δt. Ainsi, quatre satellites sont nécessaires pour obtenir le système de quatre équations requises pour déterminer la position en trois dimensions comme le montre la Figure 1.1.

La propagation des signaux transmis par les satellites à travers les différentes couches de l'atmosphère terrestre est la principale source d'erreur dans le système GPS. Par exemple, l'erreur sur les pseudo-distances causée par la traversée de l'ionosphère[2] est estimée à environ 20 mètres. La deuxième couche atmosphérique qui affecte le signal GPS est la troposphère[3]. L'erreur occasionnée par cette dernière est difficilement prédictible à cause des différentes conditions climatiques.

[2] L'ionosphère est la partie de l'atmosphère terrestre située entre 60 et 800 km d'altitude. Elle est constituée de gaz fortement ionisés.

[3] La troposphère est située entre la surface du globe et une altitude d'environ 8 à 15 kilomètres.

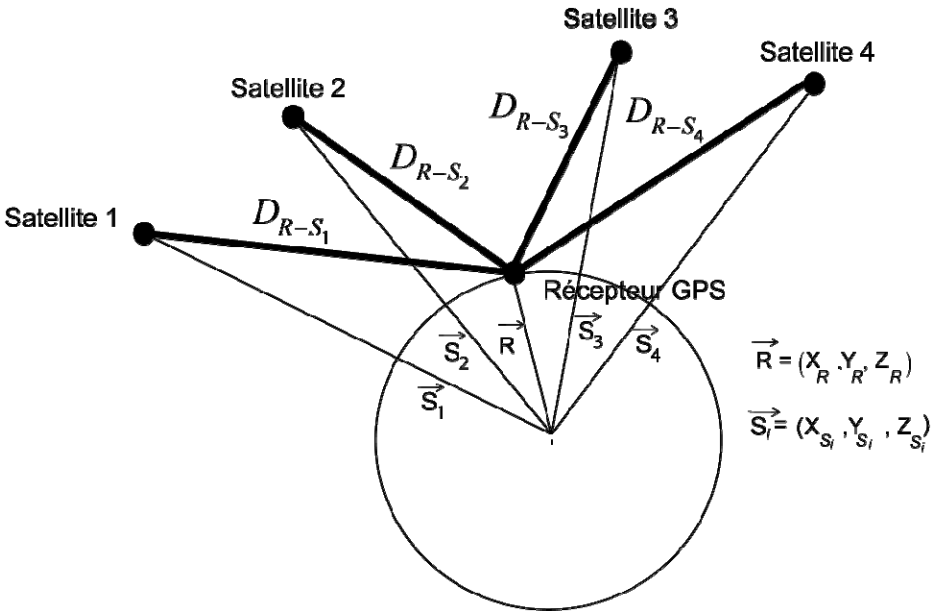

Figure 1.1. Positionnement par satellites.

Un autre élément important qui affecte la précision de localisation par le GPS est la géométrie de la constellation des satellites impliqués dans le calcul des pseudo-distances. La *Dilution of Precision* (DOP) est un paramètre contenu dans les messages de navigation et représente un indicateur de la géométrie. Il est déduit de la connaissance de la constellation des satellites. Une faible valeur du DOP indique une bonne précision de localisation. La valeur du DOP varie dans le temps à cause des changements de la géométrie des satellites engendrés par les déplacements le long de leurs plans orbitaux.

L'effet de masque causé par les obstacles rencontrés par le signal satellitaire, les trajets multiples dus aux réflexions des ondes pendant leur propagation et qui produisent un évanouissement au niveau du récepteur, constituent également des sources d'erreur sur le calcul des pseudo-distances. Ces aléas de propagation sont plus sévères quand le récepteur GPS est en milieu urbain dense ou à l'intérieur des bâtiments. Cette source d'erreur dégrade considérablement la précision du GPS. La non-visibilité entre le récepteur GPS et les satellites dans ce type d'environnement influe également sur la vitesse d'acquisition des messages de navigation nécessaires pour le calcul de la position.

Afin de minimiser l'impact de ces phénomènes, une variante du GPS, appelée *Differential*-GPS [12] a été proposée. Elle consiste à corriger la position du récepteur GPS mobile à partir de mesures effectuées par un récepteur de référence dont la position est connue. Comme

indiqué sur la Figure 1.2, chaque récepteur estime tout d'abord les pseudo-distances qui le séparent de chaque satellite. Le récepteur de référence calcule ensuite les différences entre les pseudo-distances mesurées et les véritables pseudo-distances. Le récepteur mobile peut ainsi corriger les pseudo-distances qu'il a mesurées à partir des différences qu'il a reçues du récepteur de référence.

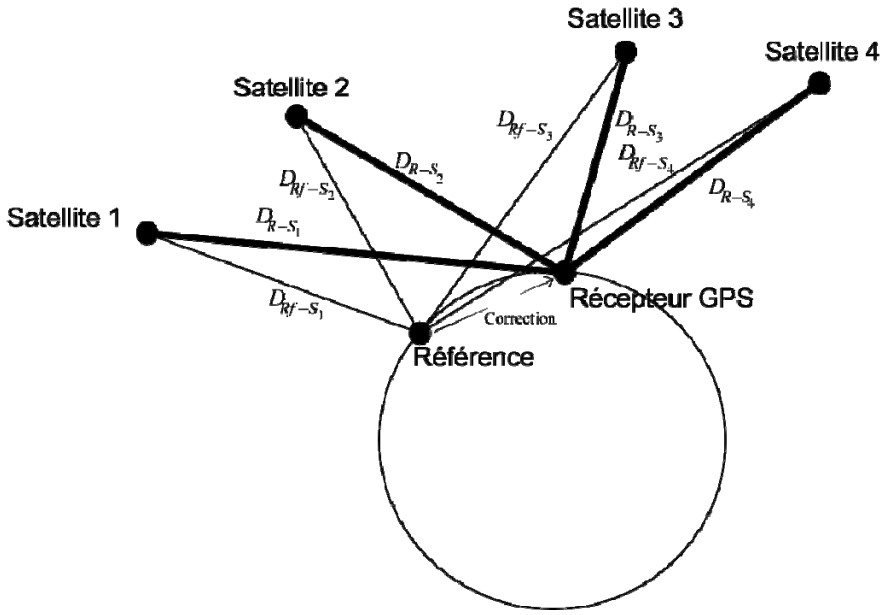

Figure 1.2. Positionnement par *Differential*-GPS.

Le D-GPS permet d'améliorer la précision de localisation quand le récepteur mobile est proche de la référence, situation dans laquelle les phénomènes de propagation engendrés par l'environnement sont sensiblement les mêmes. La contrainte liée à son utilisation, en contrepartie, est le déploiement des points de référence et le choix de leurs positions.

Afin de permettre l'utilisation du GPS dans des milieux non dégagés, un système constitué d'un réseau de stations qui peut scruter régulièrement les satellites et enregistrer des informations sur leurs positions et horloges a été proposé. Ces informations sont utilisées pour prédire les messages de navigation sur plusieurs jours. Ainsi un récepteur GPS connecté à ce réseau utilise directement ces messages pour le calcul de sa position. Cette approche est appelée *Assisted* GPS [13]. Cette technique réduit le TTFF en s'affranchissant de l'étape de synchronisation et décodage des messages de navigation, et, de plus, elle fournit au récepteur GPS les informations nécessaires pour le calcul de sa position même si ce dernier n'est pas en visibilité avec des satellites. Cependant, la précision de positionnement

de cette méthode est comparable à celle du GPS, et reste insuffisante pour les milieux intérieurs.

Dans ces milieux, les signaux satellitaires reçus sont de puissance trop faible pour être détectés par les récepteurs GPS classiques. Des récepteurs munis de mécanismes spéciaux permettant une synchronisation avec des signaux GPS de basse puissance [14] ont été proposés ; ils rendent possible l'accès aux messages de navigation en milieux bruités, mais l'implémentation de ce type de solutions nécessite une lourde modification au niveau des récepteurs. Malgré les études menées pour l'application du GPS pour la localisation dans les milieux urbains denses ou intérieurs, la précision de ce dernier demeure insuffisante. Pour ce type d'applications, des solutions plus adaptées ont pu être développées. La suite de ce chapitre sera consacrée à leur présentation.

1.2.2. Localisation par ondes radio

1.2.2.1. Localisation par proximité

La plus simple des techniques de localisation par ondes radio consiste à se localiser par rapport à un point fixe dont la position est connue. La position d'un mobile est approchée par la position connue du point fixe présentant le signal de plus forte puissance.

La *Radio Frequency IDentification* (RFID) est la plus connue des technologies utilisant la localisation par proximité. Celle-ci est basée sur l'échange des signaux radio entre un lecteur RFID et une radio-étiquette RFID. Le premier est constitué d'une antenne, d'un processeur et d'une alimentation. La radio-étiquette RFID est constituée d'une antenne et d'une mémoire ; elle peut être *active*, c'est-à-dire posséder sa propre alimentation, ou *passive*, c'est-à-dire être alimentée par le signal radio reçu. Le lecteur RFID émet régulièrement des requêtes d'informations qui sont captées par des radio-étiquettes. Ce signal fournit aux radio-étiquettes passives l'énergie dont elles ont besoin pour répondre à cette requête. La portée de ce signal radio étant limitée, les étiquettes qui répondent sont localisées dans la zone à proximité du lecteur RFID. Selon la taille de la zone à couvrir et le nombre de personnes à localiser, le déploiement de cette technologie peut être effectué de deux manières :

- dans le premier cas, les lecteurs RFID constituent la partie fixe et les personnes à localiser sont dotées de radio-étiquettes ;

- dans le second cas, la structure à couvrir est équipée de radio-étiquettes et les personnes à localiser de lecteurs RFID.

Le coût des lecteurs RFID étant plus élevé que celui des étiquettes, le choix de l'architecture de déploiement dépend du coût envisagé pour l'infrastructure. Le réseau constitué de lecteurs/étiquettes RFID peut être remplacé par un réseau *Bluetooth* [15] utilisé principalement pour la connexion sans fil entre des équipements par liaison radio de courte portée. Cette technologie est construite selon le principe Maître/Esclave, où chaque nœud du réseau « Maître » est chargé de gérer le trafic de sept « Esclaves » qui forment un pico-réseau. Afin d'assurer une continuité dans le réseau, un « Esclave » peut faire partie de plusieurs pico-réseaux. Connaissant la position et la portée du signal du nœud « Maître », on peut donc localiser les mobiles « Esclaves » qui lui sont rattachés par proximité. Les réseaux de capteurs utilisant la technologie *Zigbee* [16, 17] peuvent également remplacer la RFID. Ces technologies de courte portée permettent une bonne précision de localisation par proximité. Cependant, elles présentent l'inconvénient de nécessiter l'installation d'un réseau de capteurs.

Les réseaux WiFi installés dans certains lieux publics peuvent être considérés comme une infrastructure pour la localisation. Un réseau WiFi est constitué d'un ensemble de points d'accès, chacun permettant de couvrir une zone allant jusqu'à quelques centaines de mètres. La taille de cette zone dépend de l'environnement de propagation. Selon les spécifications de du WiFI (IEEE 802.11), chaque point d'accès diffuse régulièrement un signal sur un canal de contrôle qui permet au mobile de se connecter au réseau par le point le plus favorable. Ces canaux sont scrutés périodiquement par les mobiles afin de pouvoir utiliser un autre point d'accès si nécessaire. Parmi les informations diffusées sur les canaux de contrôle figure une identification du point d'accès. Les mobiles recevant cette information, peuvent alors se localiser à proximité (dans la zone de couverture) de ce point d'accès.

Le réseau de téléphonie mobile GSM (voir section 2.2) étant le plus déployé et le plus utilisé des réseaux sans fil (en Europe), il offre une plateforme matérielle et logicielle pour les services de localisation. Selon le principe cellulaire où chaque *Base Transceiver Station* (BTS) couvre une zone limitée, chaque mobile peut se localiser, dans le cas d'une localisation par proximité, par rapport à sa BTS de rattachement [18]. La Figure 1.3 illustre le principe de cette technique, où les positions de tous les mobiles d'une cellule sont approchées par la

position de la BTS servant cette cellule. L'identité de la cellule serveuse étant une information accessible sur le réseau GSM, cette méthode de localisation ne nécessite aucune modification sur les équipements existants. De plus aucun calcul n'est requis pour la localisation des mobiles, ce qui réduit considérablement le temps de réponse.

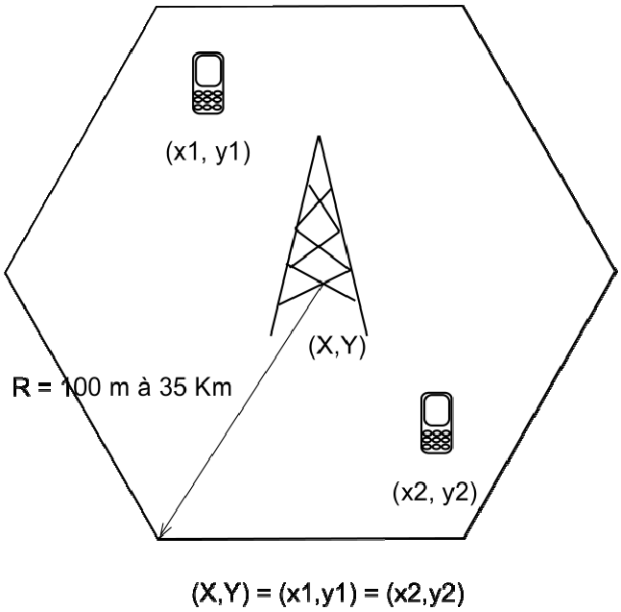

Figure 1.3. Principe de la localisation par proximité dans le cas du réseau GSM.

Cette solution présente donc des avantages incontestables, mais sa précision est étroitement liée à la taille de la cellule en question, qui peut atteindre une trentaine de kilomètres de rayon en GSM dans les zones rurales. Même si des pico-cellules (quelques dizaines de mètres) sont utilisées en milieux urbains, les performances d'une telle solution ne permettent pas pour autant son utilisation pour la localisation en intérieur.

1.2.2.2. Localisation par calcul de distance

Toujours fondée sur les signaux radio échangés au sein d'un réseau sans fil, cette technique exploite le fait que le mobile reçoit des signaux provenant de plusieurs points fixes. Le principe est d'utiliser des grandeurs physiques liées à ces signaux pour calculer les distances qui séparent le mobile de trois points (au minimum) du réseau. En supposant des conditions de propagation en espace libre et des antennes d'émission omnidirectionnelles, cette distance est le rayon d'un cercle autour d'un point fixe du réseau. L'intersection de trois

cercles, correspondant à trois points fixes de positions connues, permet de définir la position du mobile. La technique est illustrée pour un réseau de type GSM dans la Figure 1.4.

Comme pour la localisation par proximité, des réseaux de capteurs RFID peuvent être utilisés dans la localisation par calcul de distance. Un intérêt croissant est aussi manifesté pour la technologie ultra large bande ou *Ultra-WideBand* (UWB) [19]. Dans ce cas, les conditions de propagation en espace libre n'étant pas vérifiées, on fait appel à des modèles d'atténuation tels que le modèle d'Okumura-Hata [20] ou COST 231 [21] qui tiennent compte, pour l'estimation de la distance, de la hauteur de l'émetteur, de la hauteur du mobile et de la fréquence d'émission. Des solutions reposant sur le calcul de la distance entre le mobile est les points d'accès WiFi sont aussi couramment utilisées [22]. Ces réseaux sont le plus souvent installés dans les milieux intérieurs où les conditions de propagation sont difficiles à modéliser, ce qui constitue un inconvénient majeur pour ce type de techniques. En effet, plus la fréquence est élevée, moins la pénétration est importante et par conséquent plus les réflexions sont intenses [23]. Le RFID et le WiFi opérant à une fréquence de 2.4 GHz, et les émetteurs étant à l'intérieur des bâtiments, les ondes reçues par un mobile subissent plusieurs réflexions pendant la propagation du signal.

Le réseau GSM présente l'avantage que les puissances reçues de la BTS serveuse et des six voisines sont transmises régulièrement par le mobile dans des *Network Measurement Reports* (NMR), afin de prévoir un éventuel *handover* (voir section 2.2.3) vers une autre cellule en cas de baisse de la qualité de la liaison. Cette information pourrait alors être utilisée directement pour le calcul des distances séparant le mobile de ces BTS, dans l'hypothèse d'une ligne de visée directe vers chacune. En pratique, les effets des trajets multiples, atténuation, évanouissement et l'effet de masque, définis dans la section 1.2.1, sont plus sévères à cause des obstacles, présents en milieu urbain, qui favorisent ces phénomènes. Le signal reçu dans ce type d'environnement est donc influencé par une combinaison des phénomènes cités. Ceci rend difficile l'établissement d'un modèle théorique précis de la propagation des ondes radio dans ce type de milieu.

Le temps de propagation du signal entre l'émetteur et le récepteur est une autre grandeur physique qui permet le calcul de la distance. Ceci peut être obtenu par rapport à des points d'accès à un réseau Wifi [24], en estimant le temps de propagation à partir du *timestamp* contenu dans les signaux de contrôle. Dans le cas du GSM, la même méthode peut être

appliquée par rapport aux BTS, ou alors en se servant du paramètre *Timing Advance* (TA) qui est transmis régulièrement par le mobile [25] au réseau. Une exploitation réussie de cette approche nécessiterait toutefois une synchronisation temporelle entre le mobile et les points fixes du réseau, ce qui n'est généralement pas le cas dans les réseaux sans fil WiFi ou cellulaires.

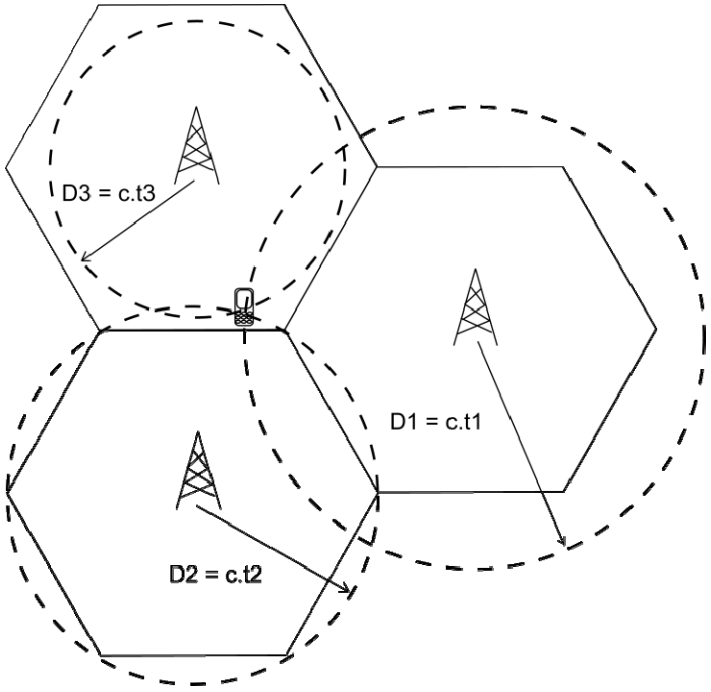

Figure 1.4. Principe de la localisation par calcul de distance dans le cas du GSM.

L'angle d'incidence du signal provenant du mobile au niveau des BTS ou des points d'accès constitue une troisième source d'information utilisable pour l'estimation de sa position. Le principe est illustré sur la Figure 1.5, où la position du mobile est déterminée par l'intersection de deux droites résultant de la mesure de l'angle d'arrivée du signal, reliant le mobile aux BTS. Un minimum de deux BTS est nécessaire pour appliquer cette méthode. Cependant une meilleure précision peut être atteinte avec plus de BTS. Cette méthode de localisation est surtout appliquée en GSM, mais peut aussi être implémentée aussi dans le cas d'un réseau WiFi ou d'un réseau de capteurs.

L'inconvénient majeur de cette méthode est qu'elle nécessite que le mobile et le point d'accès ou BTS soient en visibilité directe, ce qui n'est pas le cas dans les milieux urbains et encore moins dans les milieux intérieurs. Cette mesure implique également des antennes directives, donc une modification non négligeable du réseau existant pour une précision qui

n'est pas toujours acceptable. En effet, la mesure de l'angle d'incidence est sensible à la distance entre le mobile et la BTS. Plus cette distance est importante moins la mesure de l'angle est précise, ce qui influe sur la précision de la localisation.

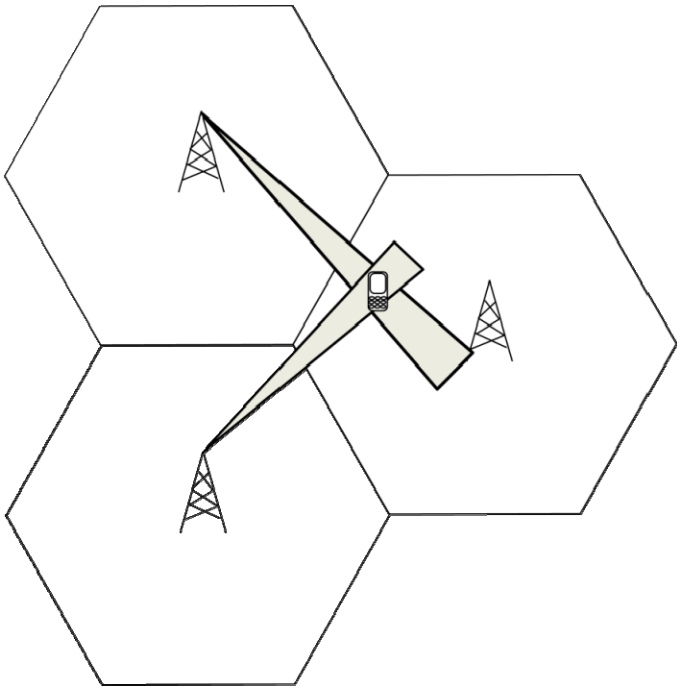

Figure 1.5. Principe de localisation par angle d'arrivée dans le cas du GSM.

En plus des inconvénients déjà cités, toutes les méthodes de localisation présentées dans cette section requièrent une connaissance des positions des BTS ou des points d'accès. Cette information peut difficilement être obtenue des opérateurs des réseaux de radiotéléphonie pour des raisons commerciales. Enfin, ces méthodes ne permettent pas non plus une utilisation dans le contexte de la localisation en intérieur.

1.2.2.3. Localisation par *fingerprints*

La technique des *fingerprints*, introduite dans [26, 27] pour les réseaux GSM et WiFi, permet de s'affranchir des inconvénients des approches citées jusqu'ici. Elle consiste, pendant une première phase, à enregistrer les caractéristiques des signaux reçus par le mobile à des positions connues. Ceci revient à faire une « carte radio » de la zone à couvrir. Pendant la phase d'exploitation, le mobile fait des mesures des mêmes caractéristiques, qui sont comparées aux enregistrements de la base de données. La position attribuée au mobile

est celle de la « plus similaire » des mesures dans la base. La procédure est illustrée par la Figure 1.6.

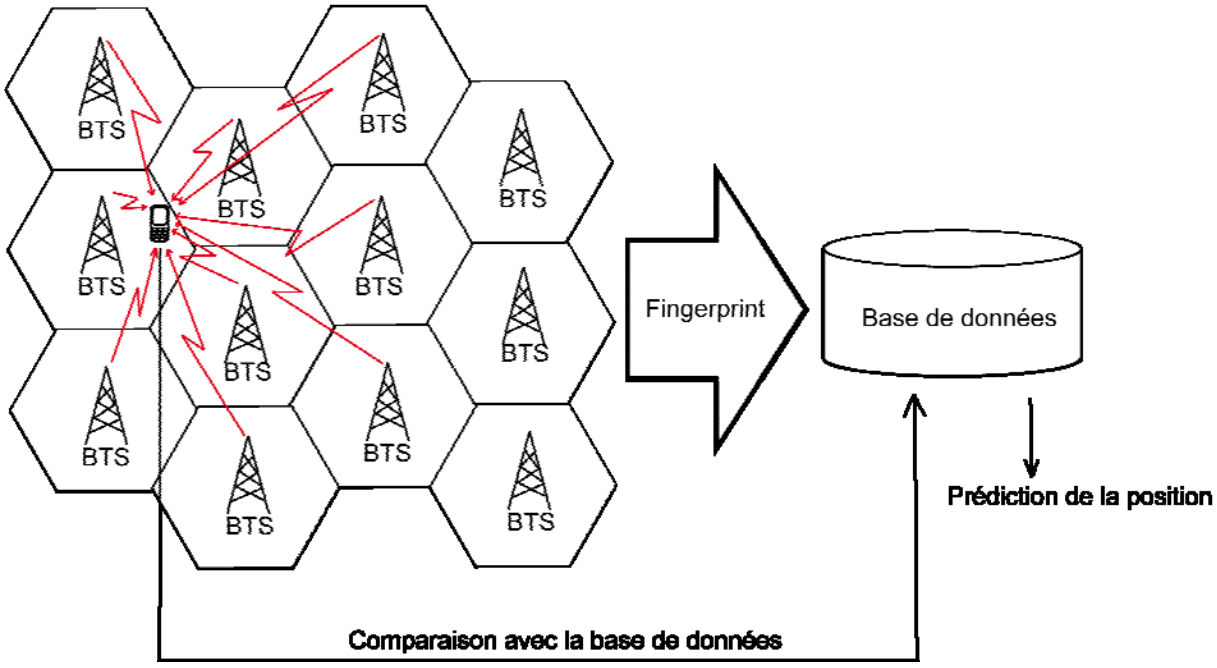

Figure 1.6. Principe de la localisation par *fingerprint* dans le cas du GSM.

Choix des caractéristiques du signal constituant les fingerprints

Les caractéristiques du signal qui sont choisies pour constituer les *fingerprints* ont évidemment une grande influence sur la précision de la localisation. Les caractéristiques les plus couramment utilisées sont les puissances reçues, ou *Received Signal Strength Indicators* (RSSI). On peut également utiliser, par exemple, le temps de propagation, l'angle d'incidence, ou la réponse impulsionnelle du canal de communication [28].

Dans le cas du réseau GSM, un sous-ensemble de ces informations est échangé régulièrement entre le mobile et le réseau sous forme des NMR prévus dans le standard GSM, qui permettent une localisation à environ 100 mètres près en milieu extérieur. Le NMR peut ainsi être interprété comme un *fingerprint* de 7 valeurs RSSI (cellule de rattachement et les 6 voisines les plus fortes). En intérieur, où les conditions de propagation sont plus compliquées, les NMR ne fournissent pas une précision adéquate. De plus, les NMR ne sont pas échangés pendant que le mobile est en veille, ce qui limite leur utilisation à la phase de communication. On peut également envisager d'étendre le *fingerprint* à des valeurs RSSI

provenant de plus que 7 BTS [29, 30], ou à un nombre plus élevé de points d'accès, dans le cas d'un réseau WiFi.

Dans le système SkyLoc [29], par exemple, les 36 plus fortes puissances mesurées par un mobile GSM sont utilisées pour une identification de l'étage à laquelle se trouve le mobile. Avec ce système, 73% des cas testés ont pu être localisés correctement, ce qui constitue déjà une performance intéressante pour une localisation en intérieur.

Localisation à partir des fingerprints

Le second aspect important dans le processus de localisation par *fingerprint* concerne le traitement apporté aux *fingerprints* et la méthode d'estimation de la position. Ces choix conditionnent largement la précision obtenue avec ces méthodes de localisation. On distingue deux types de traitement des *fingerprints* pour la localisation :

1. Mesurer la similitude entre les nouvelles acquisitions et les enregistrements de la base de données. Dans ce cas, on s'appuie sur l'hypothèse que plus les positions sont proches, plus les mesures sont semblables. Cette comparaison entre les *fingerprints* revient à un calcul de distance, le plus souvent euclidienne. Pour un vecteur **RSSI** = [RSS_1, RSS_2…RSS_n] des puissances reçues des n sources à une position inconnue, la distance d_j entre ce vecteur et le j-ième vecteur de la base $\mathbf{RSSI}_b(j)$ = [RSS_{b1}, RSS_{b2}…RSS_{bn}] est donnée par :

$$d_j = \sqrt{\sum_{i=1}^{n} \left(RSS_i - RSS_{bi}(j) \right)^2} \tag{1.2}$$

   La position associée au vecteur RSSI est alors, soit la position du vecteur j de la base pour lequel $d_j = \min(\mathbf{d})$, où $\mathbf{d}$ est le vecteur des distances entre RSSI et tous les vecteurs de la base [31], soit obtenue à partir d'une technique qui combine les positions des vecteurs de la base les plus proches dans l'espace des puissances [30]. Des métriques telles que la distance de Mahalanobis ont aussi été testées [32]. Pour une bonne précision, ces méthodes nécessitent une base de référence de grande taille. Le temps nécessaire à la localisation est proportionnel à la taille de cette base, puisque la localisation nécessite le calcul de la distance entre le *fingerprint* de la position inconnue et tous ceux de la base. Un compromis doit donc être trouvé entre la précision de la localisation et le temps de calcul nécessaire à celle-ci.

2. Trouver une correspondance fonctionnelle entre les enregistrements et les positions. Dans ce cas, des méthodes plus évoluées que le calcul des distances sont appliquées. L'idée est de construire un modèle capable de déterminer la position de l'utilisateur mobile à partir des *fingerprints* [33]. Ce modèle est construit à partir de la base de référence au cours d'une phase d'apprentissage. Ensuite, pendant la phase d'exploitation, ce modèle est appliqué au *fingerprint* du mobile à localiser pour prédire sa position. Bien qu'elle implique une étape supplémentaire d'apprentissage, cette approche a de nombreux avantages, qui seront décrits dans ce mémoire ; elle permet notamment un raccourcissement considérable du temps de réponse, puisqu'il n'est plus nécessaire de comparer le *fingerprint* à localiser à l'ensemble des vecteurs contenus dans la base de référence.

Lorsque les mesures nécessaires pour l'établissement des *fingerprints* ne sont pas disponibles, des *prédictions* des *fingerprints* peuvent également être envisagées [32], quoique avec une précision finale de localisation réduite. Ces prédictions sont faites à partir de modèles de propagation adaptés pour chaque type d'environnement [34]. Des systèmes basés sur le réseau WiFi, combinant les informations mesurées et prédites, sont commercialisés [35, 36].

1.3. Conclusion

Dans ce chapitre, les principales méthodes de localisation des utilisateurs mobiles ont été présentées. On distingue la localisation par proximité, qui est la plus simple, mais aussi la moins précise des techniques ; la localisation par calcul de distance, qui présente la difficulté de construire un modèle d'atténuation de l'onde radio prenant en considération tous les phénomènes de propagation ; et enfin, la localisation par *fingerprints*, qui semble aujourd'hui la plus prometteuse des techniques de localisation. Cette dernière peut être utilisée avec tous les réseaux sans fil ; cependant, le réseau GSM semble un choix très intéressant pour un système de localisation intérieur à base de *fingerprints*, car il est présent quasiment partout, utilisable en extérieur comme en intérieur sans nécessiter l'installation d'une nouvelle infrastructure.

Le travail de cette thèse s'inscrit dans le cadre de la localisation en intérieur par *fingerprints* GSM. Le chapitre suivant est consacré à l'identification et à l'extraction des informations pertinentes pour le positionnement des mobiles dans ce type d'environnement.

Chapitre 2. Étude des signaux GSM

2.1. Introduction

Rappelons que cette thèse est consacrée à la conception et à la réalisation d'un système de localisation de mobiles à l'intérieur de bâtiments par *fingerprints* GSM. Ce chapitre est donc consacré à l'étude des ces signaux GSM dans un tel environnement. Contrairement aux approches traditionnelles qui ne considèrent qu'un nombre limité des porteuses les plus fortes, nous considérons l'ensemble des porteuses de la bande GSM, afin de déterminer lesquelles d'entre elles sont les plus pertinentes pour la localisation des utilisateurs mobiles. Cela constitue une des originalités de notre travail.

La première partie de ce chapitre est consacrée à la présentation de quelques notions sur le fonctionnement du réseau GSM. Les appareils mobiles classiques ne permettant pas d'accéder aux types de mesures étudiés, deux dispositifs ont été mis en œuvre dans le cadre de cette étude : ils sont présentés dans la deuxième section de ce chapitre. Enfin, nous décrirons la forme des spectres mesurés, la relation entre les différentes porteuses ainsi que le comportement des mesures dans le temps.

2.2. Le réseau GSM

Le réseau GSM est incontestablement la première norme de téléphonie cellulaire et la référence mondiale pour les systèmes radiomobiles. Cette utilisation étendue du réseau GSM a permis l'apparition et le développement de différentes solutions de localisation basées sur cette technologie. Dans cette partie, nous présentons l'architecture et l'interface radio du réseau GSM, ainsi que les procédures de rattachement à ce réseau.

2.2.1. Architecture du réseau GSM

Pour assurer le bon acheminement des communications des abonnés et permettre l'exploitation et la maintenance par l'opérateur, le GSM est découpé en trois sous-ensembles comme le montre la Figure 2.1.

- Le sous-système radio ou *Base Station Sub-system* (BSS) : constitué de *Base Transceiver Stations* (BTS) et de *Base Station Controllers* (BSC), ce sous-système assure l'acheminement de communications entre le mobile et le réseau ainsi que la gestion

des ressources radio. La BTS assure l'interface entre le réseau fixe et les stations mobiles d'une zone géographique limitée appelée cellule. À chaque BTS est affecté un nombre de porteuses qui dépend du trafic estimé sur la cellule. La capacité maximale d'une BTS est de 16 porteuses. Chaque porteuse pouvant transmettre huit communications en même temps (voir section 2.2.2), une centaine de communications peuvent être prises en charge simultanément par une même BTS. Les BTS sont aussi chargées du traitement des signaux envoyés sur l'interface radio qui relie le mobile au réseau.

Le BSC contrôle un ensemble de BTS. Il assure principalement la gestion des ressources radio, l'établissement des appels ainsi que la libération des ressources à la fin de chaque appel. C'est également l'élément qui prend les décisions pour l'exécution des *handovers* (affectation du mobile à une BTS plus favorable).

- Le sous-système réseau ou *Network Sub-System* (NSS), qui a pour fonction l'établissement des appels et la gestion de la mobilité. Cette partie du réseau est constituée de plusieurs *Mobile-services Switching Centers* (MSC) essentiellement chargés des réservations des ressources pour l'établissement des communications, de l'acheminement des messages courts, de l'identification et de l'authentification des abonnés, et de l'exécution des *handovers*. À chaque MSC est associé un *Visitor Location Register* (VLR). Cette base de données contient les informations sur l'identité, le profil et les services autorisés des utilisateurs mobiles d'une zone géographique. Celle-ci contient également un numéro d'identité temporaire qui permet d'identifier chaque abonné dans le réseau. Des informations similaires concernant tous les abonnés du réseau mobile sont également enregistrées dans le *Home Location Register* (HLR) qui est la base de données de localisation et de caractérisation des abonnés au sein du réseau. Le HLR enregistre aussi le numéro du VLR auquel est rattaché chaque abonné. Cette information permet la localisation des utilisateurs du réseau.
- Le sous-système d'exploitation et de maintenance ou *Operation Sub-System* (OSS) permet l'administration du réseau GSM ; il est constitué de plusieurs *Operations and Maintenance Centers* (OMC). Ces équipements sont chargés de gérer les incidents mineurs au niveau des BTS, BSC et MSC. Les incidents plus importants sont gérés par le *Network Management Center* (NMC). Ce sous-système contient également deux

bases de données. La première, appelée *Equipment Identity Register* (EIR), contient les *International Mobile Equipment Identity* (IMEI) ; l'IMEI est le numéro d'identification internationale attribué lors de la souscription d'un abonnement mobile, qui permet d'identifier l'utilisateur, sur tous les réseaux GSM. Cette base est consultée afin de s'assurer que le mobile utilisé par l'abonné est autorisé à accéder au réseau. La seconde base est l'*AUthentification Center* (AUC) ; elle contient des informations permettant d'identifier chaque abonné au sein du réseau GSM et ainsi d'authentifier les services qui lui sont autorisés. La clé attribuée à chaque utilisateur, qui permet de chiffrer les communications, est également un élément de l'AUC.

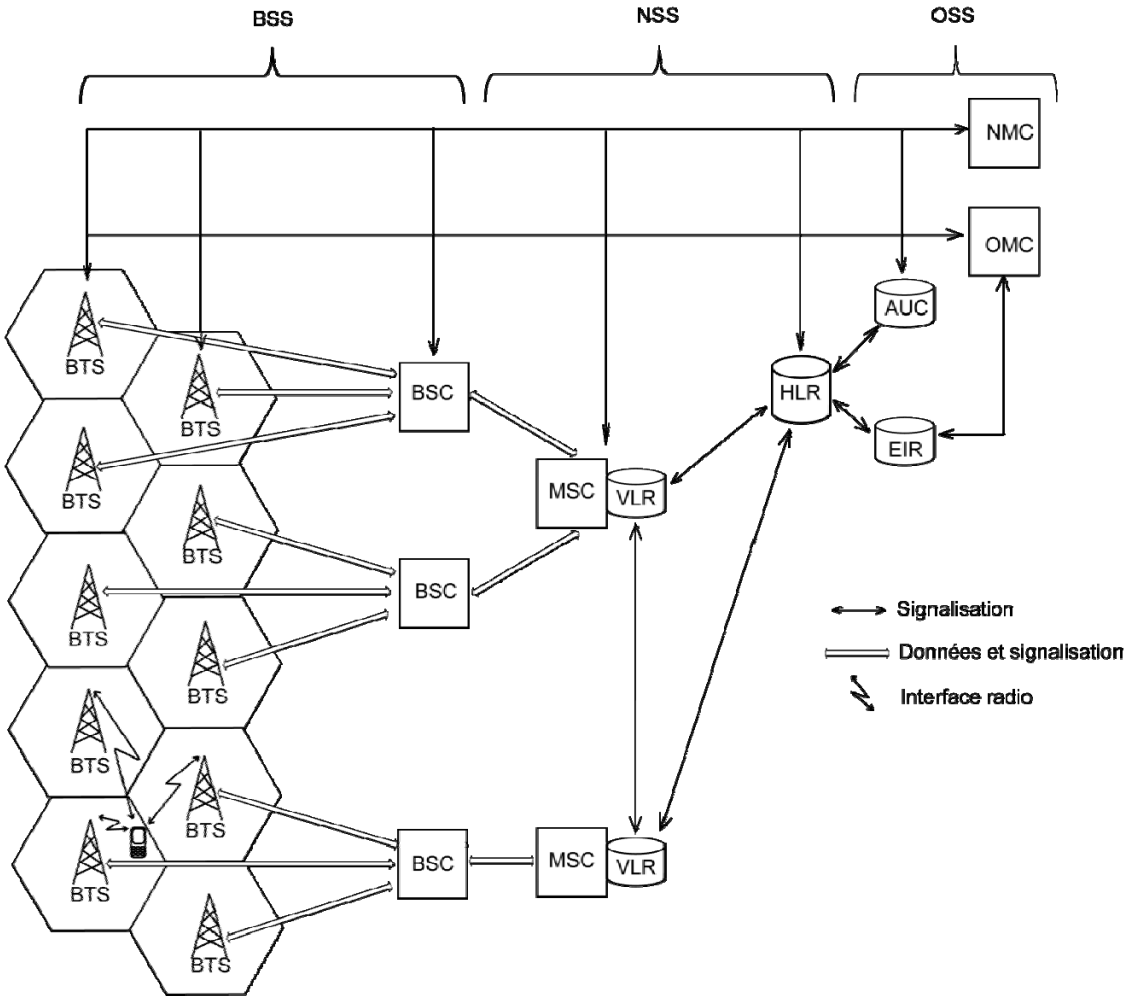

Figure 2.1. Architecture du réseau GSM.

2.2.2. L'interface radio du réseau GSM

Pour pouvoir assurer une qualité de service acceptable, le standard GSM a prévu une interface radio élaborée. Cette interface permet la communication entre le réseau et la station

mobile, et constitue l'élément sensible dans la chaîne de transmission. Pour cette interface, la norme GSM dispose de deux bandes de fréquences. Le Tableau 2.1 résume l'utilisation de ces bandes.

Tableau 2.1. Les bandes de fréquences GSM.

	GSM 900	DCS 1800
Bandes de fréquences MHz	880 – 915* (liaison montante) 925 – 960 (liaison descendante)	1710 – 1785 (liaison montante) 1805 – 1880 (liaison descendante)
Ecart duplex**	45 MHz	95 MHz

*la bande 880-890 et 925-935, appelée GSM étendu, a été ajoutée par la suite pour faire face à la forte demande des fréquences.

**L'écart duplex est le décalage entre une voie montante et une voie descendante

Le GSM utilise deux blocs de fréquences dans la bande de 900 MHz et deux autres dans la bande 1800 MHz. Dans chaque bande, un bloc est réservé aux liaisons montantes (mobile vers réseau) tandis que le second est dédié pour les liaisons descendantes (réseau vers mobile). L'utilisation des fréquences est schématisée sur la Figure 2.2.

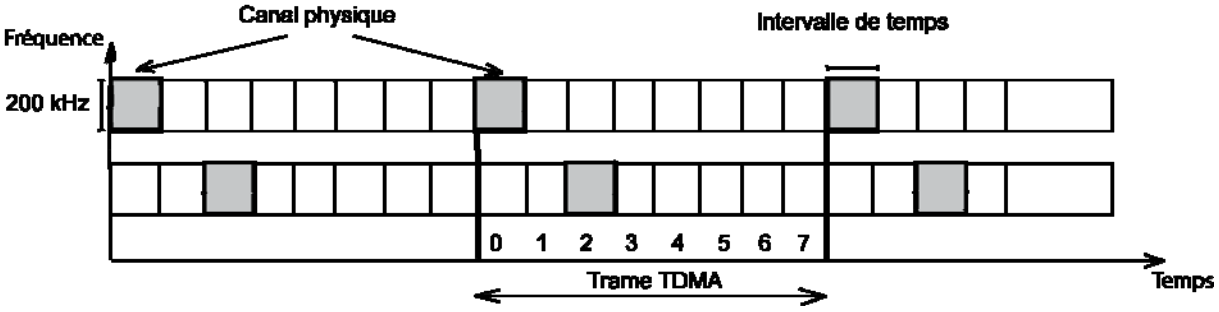

Figure 2.2. Partage des ressources radio.

Afin d'optimiser l'utilisation de la ressource radio, la bande de fréquences est divisée en canaux de 200 kHz (partage en fréquence *Frequency Division Multiple Access*). Chaque canal comprend des intervalles de temps ou *timeslots* de 577µs, groupés par huit et numérotés de 0 à 7 pour former une trame *Time Division Multiple Access* (TDMA). Un canal physique est alors défini comme un intervalle de temps sur une tranche de 200 kHz. Cette approche de partage à la fois en temps et en fréquence permet une utilisation optimale de la ressource radio. Dans la suite de ce mémoire, on utilisera le terme « porteuses » pour désigner les canaux de largeur 200 kHz. À chaque porteuse est attribué un *Absolute Radio-Frequency Channel Number*

(ARFCN). Dans la bande 900 MHz, les porteuses sont numérotées de 1 à 124 et 975 à 1024 pour le GSM étendu. La bande 1800 MHz est numérotée de 512 à 885. Ceci représente 548 porteuses GSM disponibles.

Comme indiqué plus haut, chaque BTS est chargée de rattacher les mobiles d'une cellule au réseau GSM, par l'intermédiaire des porteuses qui lui sont affectées. Une porteuse particulière est réservée à la diffusion d'informations permettant au mobile de se connecter au réseau par une BTS favorable. Cette porteuse particulière, appelée « voie balise » est émise à puissance constante. Chaque BTS possède donc une voie balise permettant au mobile de se synchroniser en temps et en fréquence, et de connaître les caractéristiques de la cellule.

Malgré le partage en temps et en fréquence, la ressource radio reste limitée, ce qui nécessite d'utiliser une même fréquence dans deux cellules différentes. Afin de minimiser les interférences, le standard GSM a prévu une distance minimale entre deux cellules utilisant la même fréquence, appelée « distance de réutilisation » D :

$$D = R\sqrt{3N} \tag{2.1}$$

où R est le rayon de la cellule et N le nombre de cellules contenant une seule fois la totalité des porteuses (appelé motif cellulaire). Pour discriminer plusieurs BTS ayant la même fréquence balise, un *Base Station Identity Code* (BSIC) différent est affecté à chacune d'elles. Une BTS est donc identifiée par sa voie balise et son code BSIC.

Pour gérer l'interface radio, un certain nombre de fonctions ont été prévues : diffusion des informations relatives au système, fourniture des supports de transmission pour l'établissement des communications, contrôle des paramètres nécessaires pour assurer les *handovers*. Dans la norme GSM, des « canaux logiques » sont définis pour assurer ces fonctions de contrôle. On distingue quatre types de canaux logiques.

- Canaux de trafic
  - Les TCH ou *Traffic Channels* sont utilisés pour la transmission de la voix et des données.
- Canaux de contrôle dédiés
  - Le SDCCH ou *Stand-Alone Dedicated Control Channel* est le premier canal affecté à l'utilisateur. Toutes les procédures d'identification, d'authentification et de chiffrement se déroulent sur ce canal. Il est également chargé des émissions/réceptions des messages courts.

- Le SACCH ou *Slow Associated Control Channel* permet d'acheminer les mesures effectuées par le mobile dans le sens montant et de transmettre des informations à la cellule serveuse et à la zone de localisation dans le sens descendant.
- Le FACCH ou *Fast Associated Control Channel* est utilisé pour transmettre les informations nécessaires pendant la procédure du *handover*.

- Canaux de contrôle non dédiés
  - Le PCH ou *Paging Channel* est réservé à la diffusion de l'identité des mobiles en cas d'appels entrants. Il est périodiquement scruté par tous les mobiles du réseau.
  - Le RACH ou *Random Access Channel* est utilisé par le mobile afin de répondre lorsque son identité est diffusée sur le PCH.
  - Le AGCH ou *Access Grant Channel* permet au mobile de faire des requêtes pour l'utilisation du réseau (appels sortants). Quand une demande est reçue sur ce canal, le réseau alloue au mobile un SDCCH pour la procédure d'authentification puis un TCH pour la transmission des données.
  - Le CBCH ou *Cell Broadcast Channel*, qui n'est pas beaucoup utilisé, est consacré à des services de publicité, météo, informations routières…etc.
- Canaux de diffusion
  - Le FCCH ou *Frequency Correction Channel* permet le calage en fréquence du mobile grâce à une sinusoïde parfaite envoyée sur une fréquence particulière.
  - Le SCH ou *Synchronization Channel* fournit les éléments nécessaires pour une synchronisation en temps des mobiles du réseau GSM.
  - Le BCCH ou *Broadcast Channel* diffuse régulièrement des informations caractéristiques de la cellule. Le BCCH est toujours diffusé sur l'intervalle numéro 0 de la voie balise.

2.2.3. Procédures de rattachement au réseau GSM

Lors de la mise sous tension, un mobile doit être capable de fonctionner sur le réseau le plus rapidement possible afin de pouvoir recevoir et émettre des appels. Le rattachement au réseau GSM se fait par la BTS reçue qui est susceptible d'assurer la meilleure qualité de service. La sélection de la BTS de rattachement se fait au niveau du mobile. Ce dernier, en scrutant toutes les porteuses reçues de la bande GSM, constitue une liste de voies balises

candidates. Dans cette liste, le mobile choisit la voie balise d'une BTS non saturée qui présente la meilleure qualité de signal reçu, ce qui lui permet d'accéder au réseau. Lorsque le mobile trouve cette balise convenable, il se cale en fréquence et en temps sur celle-ci grâce aux canaux FCCH et SCH et lit les caractéristiques de la cellule sur le canal BCCH. Dans ce cas, le mobile est en « mode veille » et continue à faire régulièrement des mesures de puissance sur une liste de voies balises qui lui est fournie sur le BCCH afin de préparer un éventuel changement de cellule.

Pendant ce « mode veille », le mobile scrute périodiquement le canal PCH pour surveiller les messages d'appel diffusés par le réseau. Si le mobile est concerné par un appel entrant, il répond au réseau sur le canal RACH. Le réseau lui affecte donc un TCH qui lui permet d'établir cet appel et le mobile entre ainsi en « mode communication ». Pendant cette phase de communication, le mobile continue à faire des mesures de puissance sur les voies balises voisines et à les envoyer au réseau sous la forme de *Network Measurement Reports* (NMR) sur le canal SACCH. Ces informations sont indispensables en cas de *handover*, où le mobile change de cellule suite à la dégradation de la qualité de la communication. En cas d'appel sortant (le mobile souhaitant effectuer une communication), le mobile fait une requête d'utilisation du réseau sur le canal AGCH. Cette requête reçue, le réseau affecte au mobile un canal SDCCH permettant la procédure d'indentification et d'authentification. Une fois cette procédure terminée, le mobile dispose d'un canal TCH pour acheminer le trafic de voix et de données et d'un canal SACCH pour superviser la liaison radio.

Conformément au fonctionnement présenté dans le paragraphe précédant, la mesure des puissances reçues sur les porteuses GSM est donc une procédure prévue par la norme, et régulièrement effectuée par le mobile (en mode veille ou en communication). Cette information permet, au sein du réseau GSM, d'assurer une bonne qualité de service. La méthode de localisation que nous proposons, fondée sur l'analyse des puissances reçues, met donc à profit une procédure qui est effectuée normalement par le mobile, en veille ou en communication. Elle ne nécessite donc pas de modification majeure du téléphone mobile.

2.3. Dispositifs de mesure

2.3.1. Le mobile à trace TEMS

Utilisé par les opérateurs de téléphonie mobile, cet outil est conçu pour la maintenance, la gestion et le dépannage des réseaux sans fil. L'utilisation du mobile à trace TEMS [37] nécessite une liaison avec un ordinateur muni du logiciel de contrôle approprié (Figure 2.3). Ce dernier permet l'exploitation et le traitement des mesures faites par le mobile. Dans la variante du mobile à trace appelée TEMS Pocket, le logiciel de contrôle est implémenté sur le mobile [38].

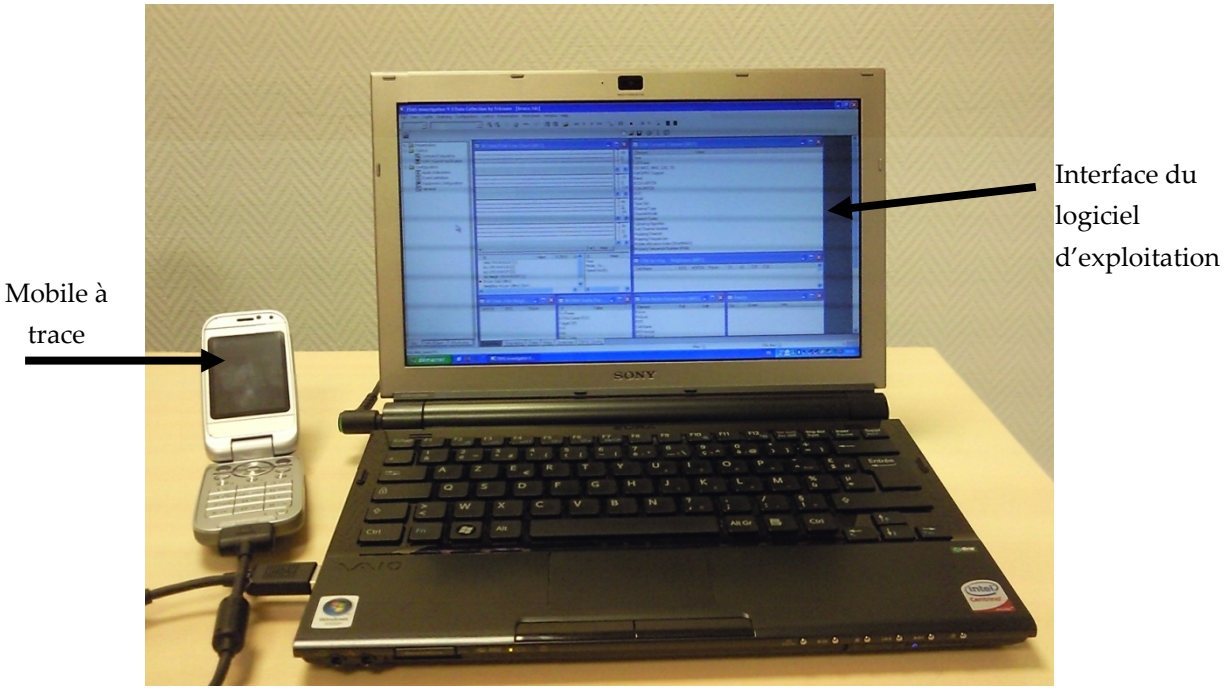

Figure 2.3. Dispositif de mesure TEMS Investigation 9.0.

Le dispositif TEMS permet, entre autres, d'accéder aux messages échangés sur le réseau GSM. Dans le cadre de ce travail, le mobile à trace est utilisé en mode *scan*, qui permet un balayage de toutes les porteuses de la bande GSM. Ainsi, on dispose du numéro *Absolute Radio-Frequency Channel Number* (ARFCN) et du *Received Signal Level* (RxLev[4]) pour chaque porteuse de largeur 200 kHz, qui pourront par la suite être incorporés dans des *fingerprints*. Pour les canaux *Broadcast Control Channel* ou BCCH (voir section 2.2.2), en plus de ces informations, le mobile à trace offre la possibilité de décoder le code BSIC. Cette dernière

[4] RxLev = Puissance reçue (dBm) + 110.

opération a malheureusement pour effet d'augmenter la durée d'un balayage, qui atteint dans ce cas une minute environ.

2.3.2. Le modem Télit

Cet outil appartient à la catégorie de dispositifs *machine to machine* (M2M) qui est à l'interface entre les domaines de l'informatique et des télécommunications. Cette technologie, qui a connu un essor considérable au cours des dernières années, permet, par exemple, la communication entre des calculateurs et un serveur central grâce aux réseaux sans fils existants. Pour cette étude, le module GM862 de Télit [39] a été choisi. Ce dernier est constitué d'une puce GSM qui permet d'accéder aux informations sur ce réseau, et d'une puce GPS. Le dispositif permet d'enregistrer les ARFCN et RxLev pour toutes les porteuses GSM reçues, et le code BSIC pour les canaux BCCH. La durée du balayage de la totalité de la bande GSM est de l'ordre de trois minutes. Les mesures ayant été effectuées à l'intérieur, la fonctionnalité GPS n'a pas été utilisée.

Le dispositif de mesure construit pour cette étude comporte, en plus du module GM862, un microcontrôleur qui sert d'interface entre l'utilisateur et le modem (Figure 2.4). Cette interface permet de faire des requêtes de balayage des porteuses GSM au module Télit. Comme c'est le cas pour tous types de modems, cette requête est implémentée via une commande AT (*ATtention Command* nécessaire pour le fonctionnement des modems). Le microcontrôleur permet d'enregistrer les mesures effectuées par le GM862 sur une carte mémoire de type *Secure Digital* (SD). Le GM862 présente l'avantage d'être moins encombrant que le TEMS, puisqu'un simple microcontrôleur permet d'interagir avec le GM862 (Figure 2.4). Cependant, la durée du balayage est trois fois plus longue que celle du TEMS. Dans le cadre de cette étude, la carte à droite de la Figure 2.4 a été conçue et réalisée au Laboratoire SIGMA. Dix dispositifs ont ainsi été construits, afin de permettre une collecte rapide des *fingerprints* en effectuant des mesures simultanées à différentes positions.

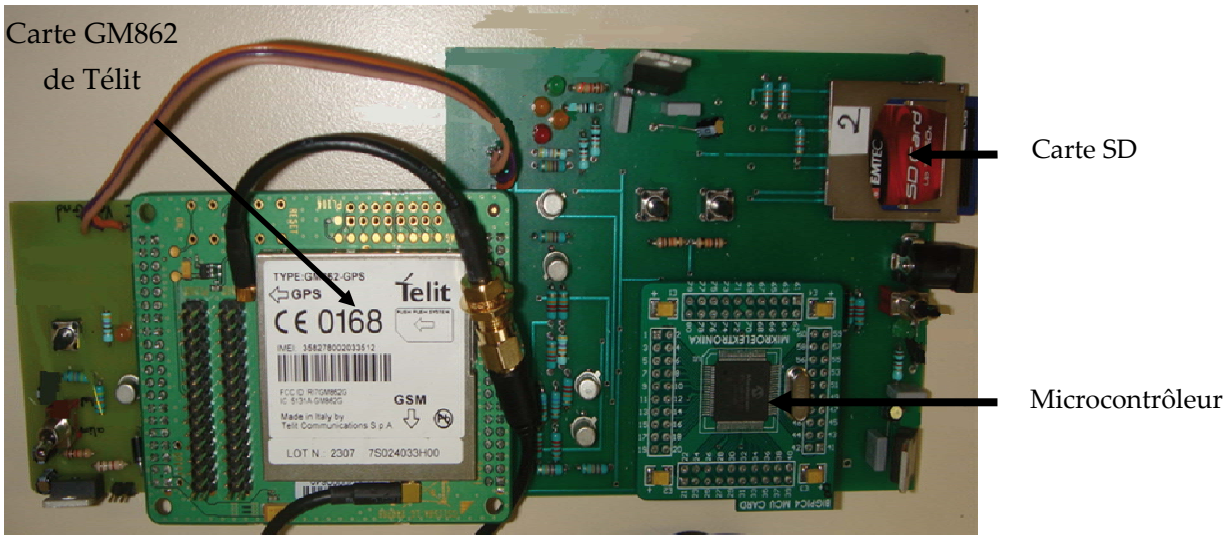

Figure 2.4. Dispositif de mesure Télit.

Étant donné les durées nécessaires à l'exécution d'un balayage, de l'ordre de la minute, les dispositifs présentés (TEMS et GM862) ne peuvent être envisagés pour une application finale de localisation, où une réponse quasiment instantanée est préconisée. Ils permettent néanmoins d'évaluer la faisabilité et les performances de la localisation par *fingerprints* avec de données acquises dans des environnements réels. Des équipements plus évolués existent ; ils sont plus compacts et permettent un balayage plus rapide (environ 300 millisecondes [40]). Ils sont beaucoup plus coûteux. Le TEMS Pocket [38], mentionné plus haut, en est un exemple.

La suite de ce chapitre est consacrée à la présentation des données mesurées et à l'étude du comportement des puissances reçues sur les porteuses GSM dans un milieu intérieur. Plusieurs expériences ont été mises en œuvre afin de comprendre les phénomènes observés sur les mesures acquises. Pour cette partie, un seul modem Télit est utilisé. Dans une deuxième phase et afin de tester les dispositifs de mesure, les mesures des dix modems ont été comparées.

2.4. Étude du spectre GSM obtenu avec un appareil de mesure

Une porteuse GSM, telle que définie par le standard, est une bande de fréquence d'une largeur de 200 kHz. Dans cette étude, chaque porteuse sera représentée, dans la suite, par un numéro entre 1 et 548. La première expérience menée a consisté à mesurer un spectre GSM à une position donnée. La Figure 2.5 montre les puissances reçues des 548 porteuses GSM.

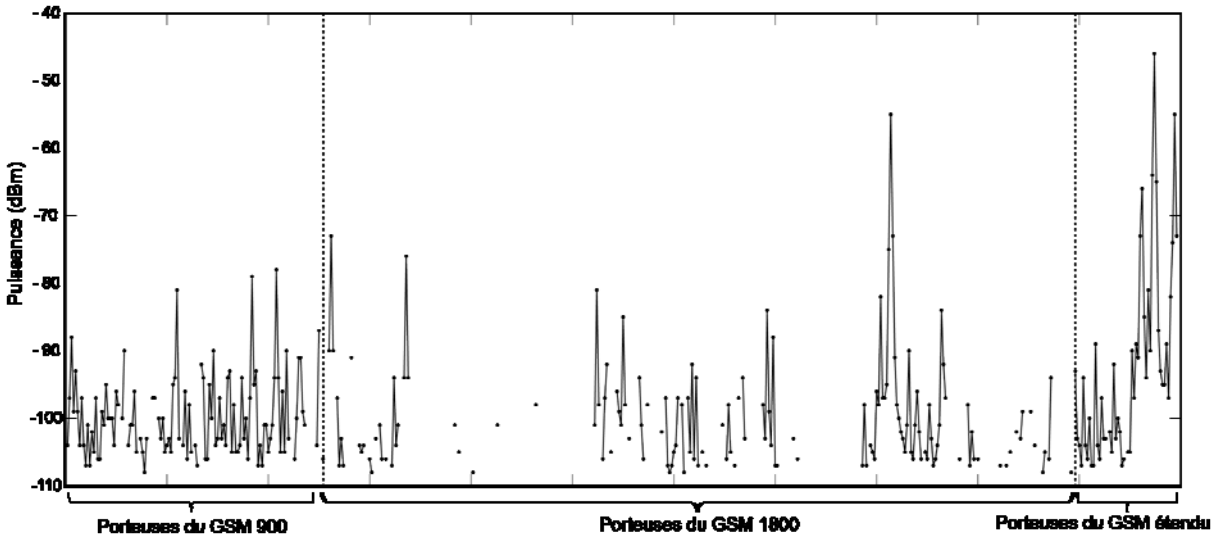

Figure 2.5. Spectre GSM mesuré. Les porteuses adjacentes dans le spectre sont reliées par des segments ; les points isolés sont les puissances des porteuses pour lesquelles les porteuses adjacentes n'ont pas été détectées.

Ce spectre ou une partie de celui-ci constituera par la suite un *fingerprint* de cette position et sera utilisé pour la localisation. On observe que les 548 porteuses ne sont pas toutes reçues ; deux explications peuvent être envisagées. La première est que certaines d'entre elles pourraient ne pas être actives sur le réseau au moment de la mesure, c'est-à-dire qu'aucun signal ne serait émis sur ces porteuses. La seconde est que leurs puissances de réception pourraient être inférieures au seuil détectable par le dispositif (-108 dBm [39]).

On observe également que ce spectre présente une forme particulière autour de certaines porteuses. En effet, si P est une puissance reçue sur une porteuse, les puissances mesurées sur les porteuses adjacentes présentent des valeurs d'environ P – 20 dB. Ce comportement est observé autour de toutes les fortes puissances ; un exemple agrandi est représenté sur la Figure 2.6. Cette forme du signal résulte du gabarit que le spectre du signal modulé doit respecter selon la norme GSM [23]. Théoriquement, si la puissance reçue sur une fréquence f est égale à P, l'interférence sur les fréquences f – 200 kHz et f + 200 kHz (porteuses adjacentes) est de P – 30 dB. Cette limite théorique est fixée pour le cas où aucun signal n'est émis sur les porteuses adjacentes ce qui n'est pas toujours vrai dans la réalité et qui explique l'interférence de P – 20 dB observée sur les mesures effectuées. Cette interférence est donc proportionnelle à la puissance, et présente sur toutes les porteuses.

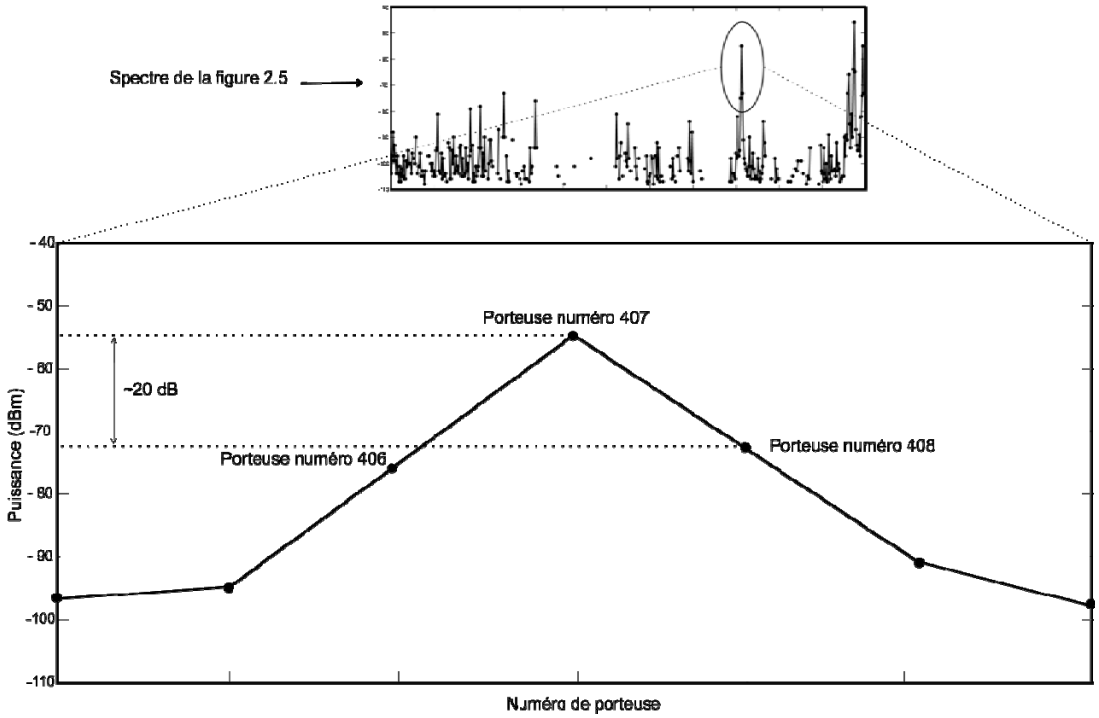

Figure 2.6. Gabarit du signal GSM reçu.

Cette observation est confirmée par les résultats de mesures de puissance effectuées, à 10 minutes d'intervalle, pendant 40 jours, par un dispositif fixe (Télit). La Figure 2.7 montre les niveaux des signaux reçus sur deux porteuses adjacentes (les porteuses 406 et 407) : les puissances reçues sur ces deux porteuses évoluent de manière identique au cours du temps.

Les puissances mesurées sur les deux porteuses étudiées présentent une variation au cours du temps qui peut atteindre 20 dB. Il était donc important d'essayer de comprendre la source de cette variabilité.

L'hypothèse selon laquelle un effet de l'électronique du dispositif serait à l'origine de cette variation a été écartée par des tests effectués dans une chambre anéchoïde. L'expérience s'est déroulée dans une chambre d'une hauteur de 5 mètres, une largeur de 3 mètres et d'une hauteur de 3 mètres. Les murs de cette chambre sont couverts de pyramides en mousse de polyuréthane chargée d'un complexe à base de carbone afin d'absorber les ondes électromagnétiques et annuler ainsi leur réflexion. Des mesures de la puissance émise par un générateur de signal Rohde & Schwarz, sur la fréquence d'une porteuse GSM ont été effectuées avec un dispositif Télit sur une durée de 90 minutes (une mesure chaque dix minutes) ; des mesures de puissance sur la même porteuse GSM et avec le même dispositif ont été enregistrés hors chambre anéchoïde.

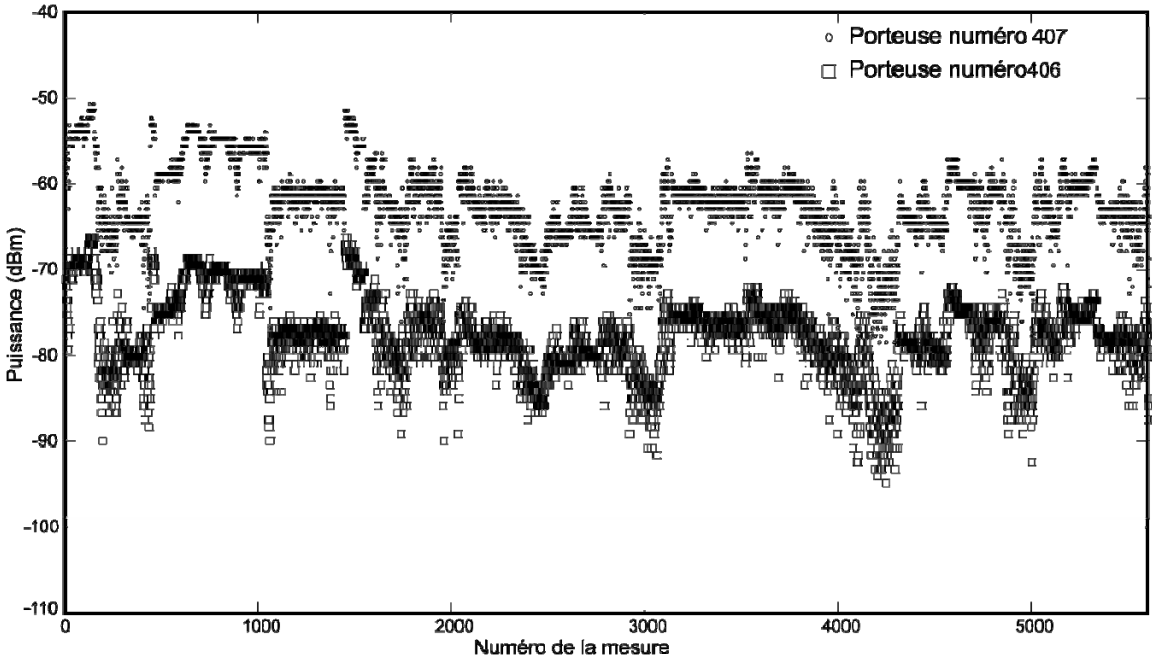

Figure 2.7. Évolution de la puissance reçue au cours du temps pour deux porteuses adjacentes.

La Figure 2.8 montre que la variation observée sur les mesures dans un milieu réel est le résultat des phénomènes de propagation (effet de masque, trajets multiples, etc.) subis par l'onde radio dans ce milieu.

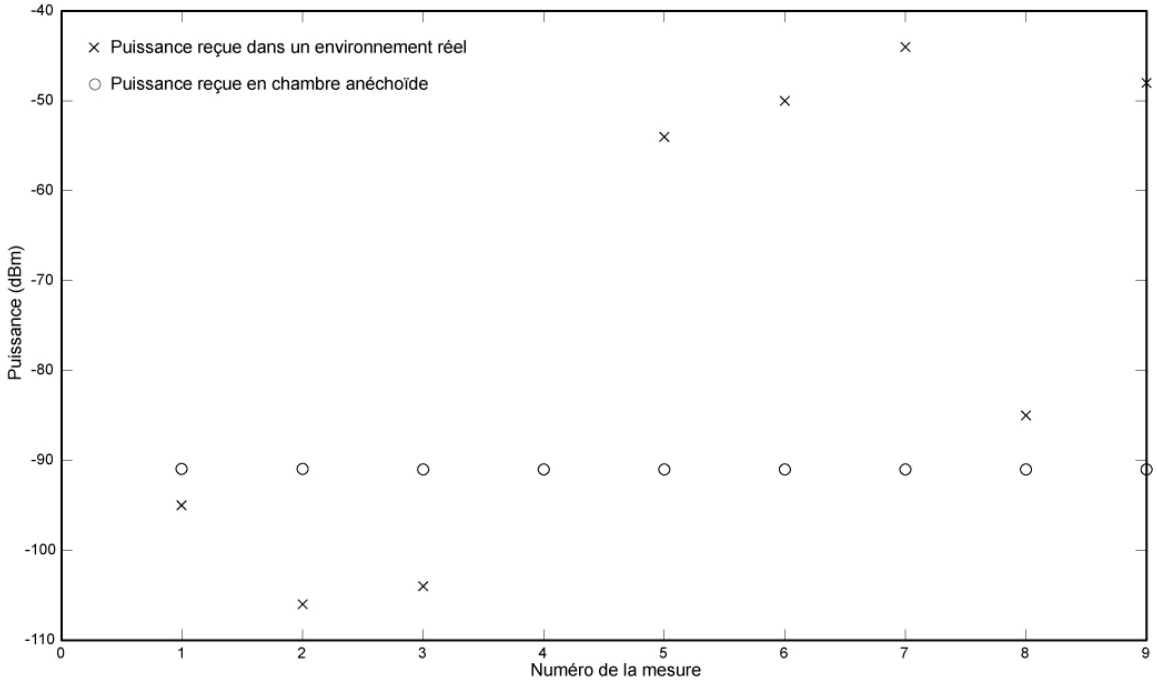

Figure 2.8. Comportement d'une porteuse GSM dans une chambre anéchoïde et dans l'environnement réel.

Afin de caractériser la variabilité des signaux reçus pendant une période de 40 jours, l'écart-type des puissances mesurées sur chaque porteuse est représenté en fonction de la puissance moyenne sur la Figure 2.9. Les porteuses pour lesquelles les puissances reçues sont supérieures à -90 dBm environ sont relativement stables (écart-type faible). Pour les porteuses reçues avec un niveau moyen inférieur à -90 dBm, on distingue deux types de comportements. Les puissances reçues de certaines porteuses ont un écart-type faible à peu près indépendant de la puissance moyenne, alors que d'autres puissances présentent un écart-type croissant avec la puissance moyenne.

Comme indiqué dans la section 2.2.2, la norme GSM a prévu que chaque cellule dispose d'une porteuse particulière, appelée voie balise, chargée de diffuser en permanence des informations qui permettent aux mobiles de se connecter au réseau GSM dans les meilleures conditions. Cette voie balise présente la particularité d'être émise à puissance constante, contrairement aux autres porteuses qui sont des canaux utilisés pour acheminer le trafic dans le réseau. Chaque voie balise s'identifie par son code BSIC.

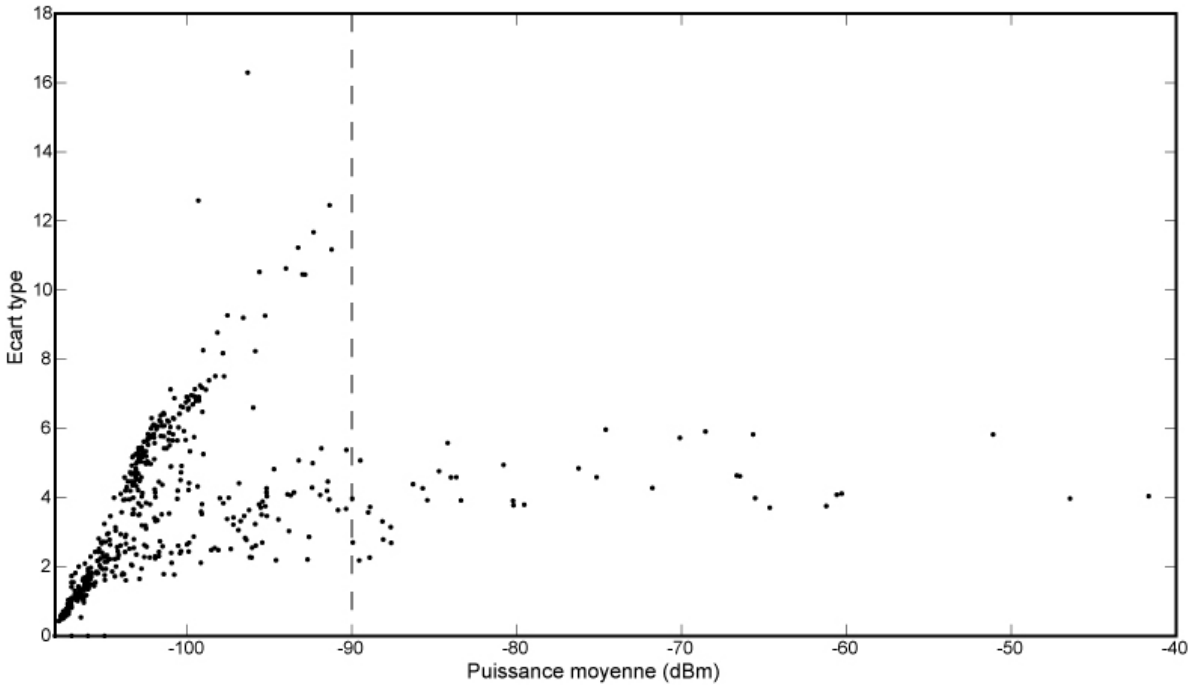

Figure 2.9. Écart-type des puissances mesurées en fonction de la puissance moyenne.

Dans les mesures que nous avons effectuées, il est souvent impossible de détecter un BSIC. La Figure 2.10, qui représente le taux de détection de BSIC en fonction de la puissance

moyenne reçue pendant une période de 40 jours, montre que l'absence d'un BSIC n'a pas de rapport particulier avec la puissance reçue sur l'ARFCN d'une porteuse. On peut même observer des porteuses de fortes puissances pour lesquelles le BSIC n'a jamais été détecté. En revanche, les porteuses pour lesquelles le BSIC a été fréquemment détecté sont certainement des voies balises. Les porteuses qui sont souvent détectées comme étant des voies balises se situent à droite de la ligne en pointillés sur la Figure 2.9. Elles présentent peu de variation de puissances reçues au cours du temps puisqu'elles sont émises à puissances constantes, et les variations observées sont uniquement dues aux phénomènes de propagation.

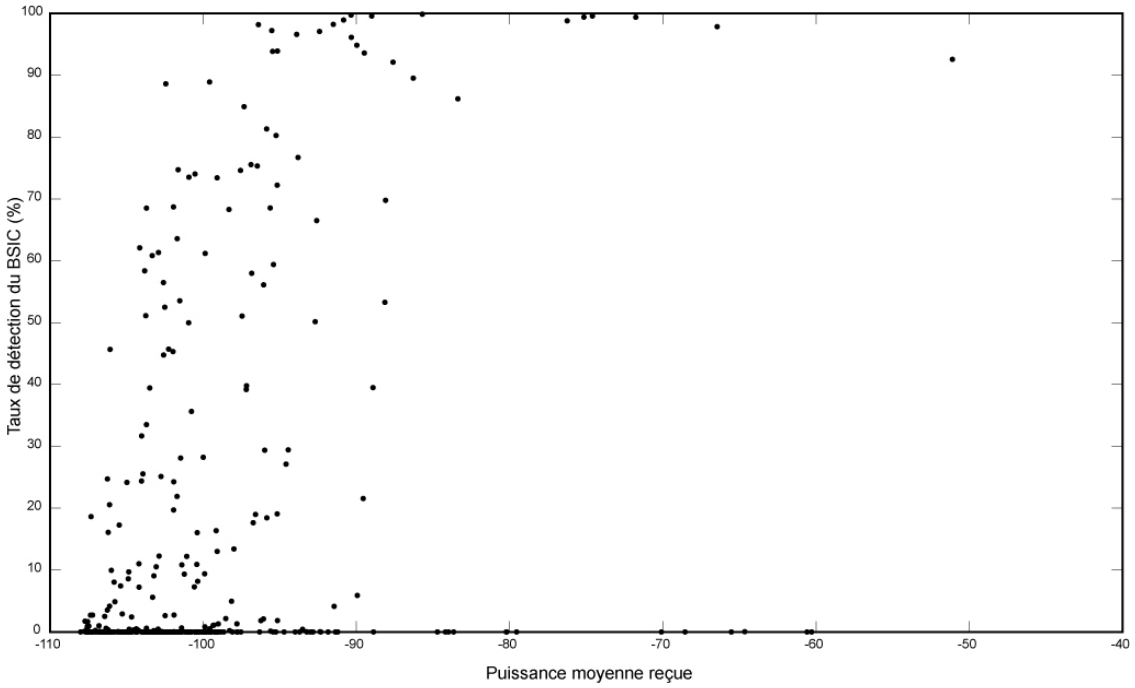

Figure 2.10. Taux de détection du BSIC en fonction de la puissance reçue.

Les mesures étudiées sur une période de 40 jours avec un dispositif Télit permettent aussi de constater que, pour une porteuse donnée, le BSIC peut ne pas être détecté pendant toute la durée de la prise de mesures. Ceci peut être expliqué par le fait que, dans la norme GSM, il est prévu que le BSIC soit diffusé sur un canal logique appelé *Broadcast Control Channel* (BCCH) émis sur la voie balise pendant les intervalles de temps (*timeslots*, voir section 2.2.2) numéros 0, 2, 4 ou 6. Sur les intervalles de temps restants (1, 3, 5, 7), ce même ARFCN peut être utilisé pour des canaux logiques de trafic [23]. Ainsi, si, au moment du balayage, la mesure est faite pendant un intervalle de temps sur lequel un canal BCCH n'est pas diffusé,

le BSIC ne peut être détecté. De plus, la lecture du BSIC nécessite la synchronisation temporelle entre l'appareil de mesure et la trame TDMA. Cette synchronisation est difficile si plusieurs répliques de l'onde radio, dues aux trajets multiples, arrivent au niveau de l'appareil de mesure en même temps, ce qui est un phénomène très courant dans un environnement intérieur.

La localisation par analyse des *fingerprints* repose sur l'hypothèse que chaque position peut être distinguée grâce au spectre GSM mesuré en cette position, compte tenu du fait que les ondes radio reçues en différentes positions ne sont pas sujettes aux mêmes perturbations pendant leur propagation. Cette hypothèse est confirmée par les résultats présentés sur la Figure 2.11, qui montre les spectres GSM mesurés dans deux positions différentes au même instant.

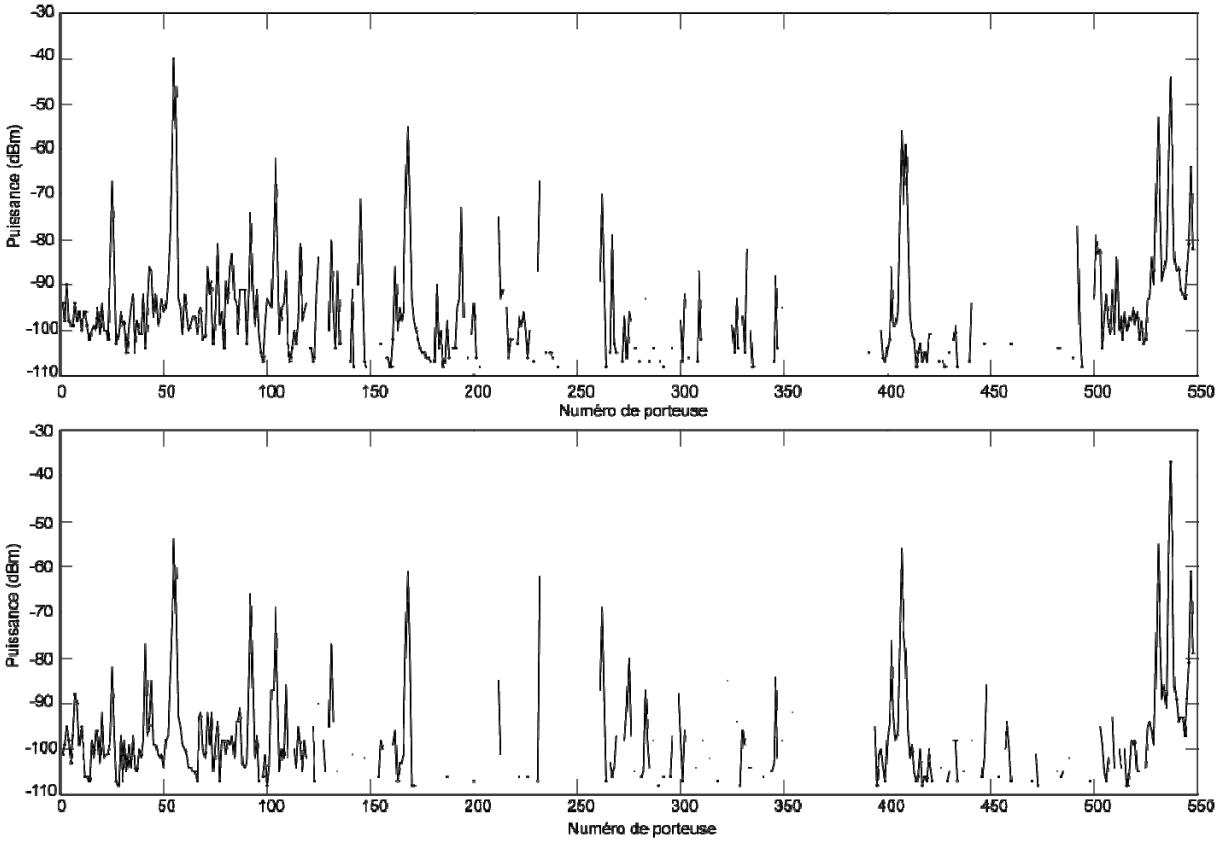

Figure 2.11. Spectres GSM mesurés à différentes positions.

2.5. Comparaison des différents appareils de mesure

Les mesures présentées précédemment ont été prises avec un seul dispositif de mesure. Dans ce qui suit, on s'intéressera à des mesures issues de différents dispositifs. Ceci

permettra de vérifier si des appareils basés sur la même plateforme matérielle et logicielle, et équipés d'antennes similaires, sont capables de mesurer le même spectre GSM à la même position. Cette étude est intéressante pour l'application finale de localisation afin de vérifier si des mesures faites avec un dispositif donné peuvent servir pour localiser d'autres appareils mobiles.

La première expérience consiste à comparer des spectres GSM mesurés au même instant et à la même position par différents dispositifs de mesure. Pour cela, quatre des dix systèmes Télit construits au laboratoire ont été utilisés. En pratique il est évidemment impossible d'effectuer des mesures du spectre GSM exactement à la même position avec quatre dispositifs en même temps. Pour cette expérience, les antennes des appareils ont été posées le plus près possible les unes des autres, mais la distance entre une paire d'antennes restait de l'ordre de 6 centimètres. Les spectres GSM obtenus sont présentés sur la Figure 2.12.

Bien que les dispositifs soient constitués de modems et d'antennes identiques et que les mesures soient effectuées à la même position, les spectres obtenus présentent des différences évidentes. Certaines porteuses ne sont pas détectées par tous les appareils ; de plus, on remarque d'importantes différences des niveaux de puissance reçue par les différents appareils sur une même porteuse. Il faut admettre que, la distance entre les antennes étant de presque une demi-longueur d'onde, les écarts de puissances observés pourraient être dus à des évanouissements de l'onde radio provoqués par l'effet des trajets multiples, voire par des interférences entre les antennes. En effet, le signal mesuré au niveau de l'antenne est le résultat des contributions des ondes reçues, qui peuvent être destructives ou constructives selon la position de l'antenne. Ce phénomène se produit à l'échelle de la longueur d'onde qui est de 30 centimètres pour le GSM 900 et de 15 centimètres pour le GSM 1800 [41].

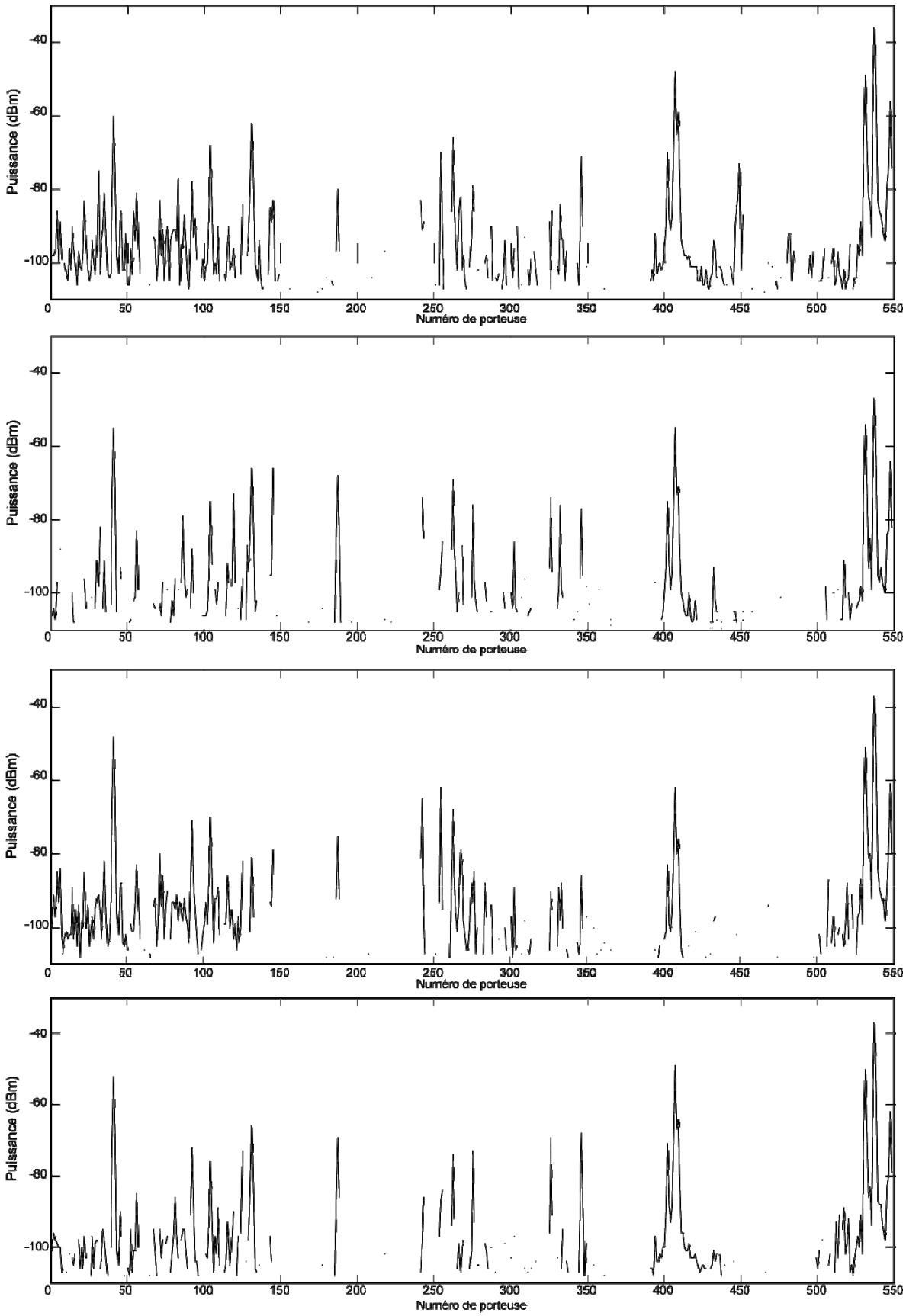

Figure 2.12. Spectres GSM mesurés par différents dispositifs à la même position.

Afin de faire une comparaison dans des conditions adéquates, huit des dix dispositifs ont été testés dans la chambre anéchoïde utilisée précédemment. Ceci a permis de comparer des mesures faites dans les mêmes conditions de propagation, uniquement sur l'onde directe. L'expérience consistait à mesurer avec chacun des dispositifs le signal émis par un générateur de signal Rohde & Schwarz, sur quatre porteuses GSM pour lesquelles les différences entre les spectres de la Figure 2.12 sont importantes.

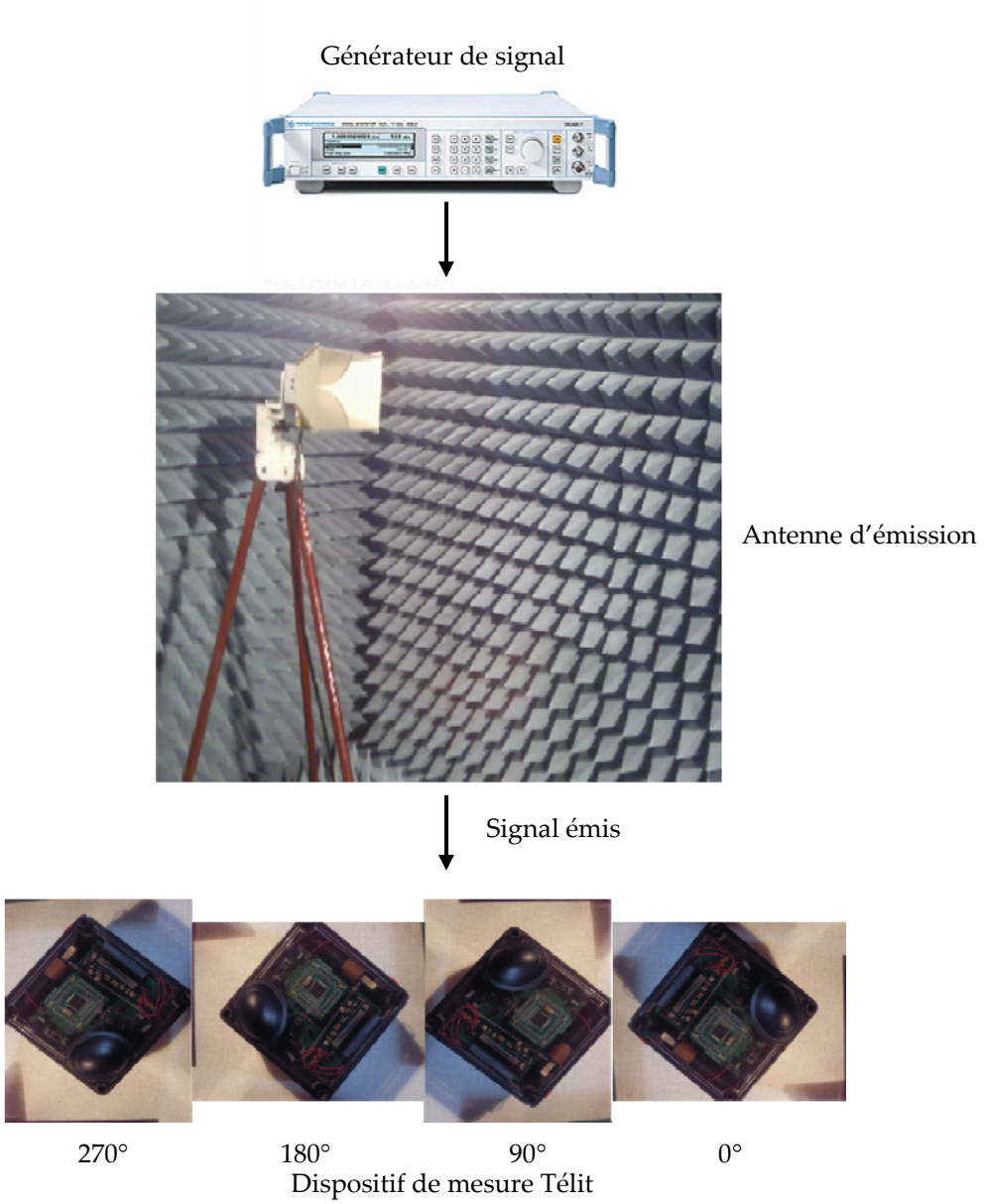

Figure 2.13. Description de l'expérience menée en chambre anéchoïde.

Les mesures ont été faites avec différentes positions de l'antenne de réception (supposée omnidirectionnelle), comme le montre la Figure 2.13, afin de tester également l'effet de l'orientation de l'antenne sur les niveaux de puissance.

Tableau 2.2. Puissances mesurées (dBm) en chambre anéchoïde.

ARFCN	Orientation	Dispositif 1	Dispositif 2	Dispositif 3	Dispositif 4	Dispositif 5	Dispositif 6	Dispositif 7	Dispositif 8
512	0°	-81	-86	-85	-92	-84	-88	-87	-82
	90°	-79	-84	-87	-89	-82	-89	-86	-87
	180°	-86	-96	-88	-94	-83	-85	-87	-89
	270°	-81	-82	-85	-89	-87	-84	-92	-93
604	0°	-82	-91	-90	-91	-90	-88	-92	-86
	90°	-80	-95	-84	-89	-93	-89	-95	-84
	180°	-78	-91	-85	-88	-90	-87	-96	-87
	270°	-82	-92	-89	-84	-94	-88	-91	-89
712	0°	-86	-82	-82	-84	-88	-88	-89	-84
	90°	-93	-81	-85	-86	-84	-92	-87	-89
	180°	-95	-92	-82	-87	-91	-87	-90	-90
	270°	-97	-89	-88	-86	-85	-93	-96	-95
729	0°	-88	-82	-81	-91	-92	-89	-88	-78
	90°	-91	-87	-83	-84	-84	-86	-96	-85
	180°	-96	-89	-83	-92	-84	-90	-97	-85
	270°	-85	-88	-84	-93	-86	-87	-94	-88

Les résultats obtenus en chambre anéchoïde sont résumés dans le Tableau 2.2. Chaque ligne de ce tableau représente les mesures faites par tous les dispositifs avec la même orientation de l'antenne et sur la même porteuse. Les écarts observés sur les puissances

(pouvant atteindre 15 dB) présentées sur une ligne du tableau confirment la différence entre les dispositifs testés. D'après les mesures présentées sur le Tableau 2.2, il est évident qu'il existe une directivité de l'antenne (supposée omnidirectionnelle) qui se traduit par une variabilité des niveaux de puissances reçus sur une porteuse GSM en différentes orientations.

2.6. Conclusion

Cette étude des signaux GSM nous a permis de mettre en évidence des caractéristiques du comportement des porteuses de la bande GSM qui constituent un élément majeur dans la méthode de localisation proposée dans ce travail. En résumé :

- Les spectres des puissances GSM reçues varient en fonction du lieu à l'intérieur d'un bâtiment, ce qui justifie l'idée d'effectuer une localisation à partir de ces données.
- Il existe une variabilité au cours du temps des puissances reçues sur les porteuses GSM, due aux fluctuations de l'environnement de propagation.
- Il existe une corrélation, entre les porteuses adjacentes, proportionnelle à la puissance reçue.
- Il existe une variabilité entre les différents dispositifs de mesure bien qu'ils soient constitués du même modem et d'antennes identiques.

Ces résultats ont conditionné en partie les objectifs et méthodes mis en œuvre dans le présent travail ; ils sont décrits dans le chapitre suivant.

Chapitre 3. Objectifs et méthodes

3.1. Introduction

Le problème de la localisation en intérieur peut être abordé de différentes manières. Comme indiqué dans la section 1.1, on peut définir la localisation d'un mobile de deux manières :

- par ses coordonnées dans un espace donné,
- par sa position par rapport à son environnement.

Cette étude se place dans le second cas : il s'agit de définir la pièce, d'un environnement intérieur connu, dans laquelle se trouve un utilisateur mobile, à partir de l'enregistrement du spectre GSM. Nous avons donc abordé ce problème comme un problème de classification : tous les enregistrements de spectres GSM effectués dans une même pièce sont affectés à la même classe ; un enregistrement réalisé par un mobile à localiser doit être affecté par le dispositif à une des classes ainsi définies. Compte tenu des variabilités mises en évidence par les résultats décrits dans le chapitre précédent, nous avons eu recours à des méthodes de classification qui entrent dans le cadre de l'apprentissage statistique.

Un système de localisation efficace doit obéir à des contraintes de simplicité et de performance (en termes de taux de localisation de la pièce). Celui-ci doit aussi tenir compte des variabilités, dans le temps et entre les dispositifs de mesure, observées lors des mesures présentées dans le chapitre 2. La solution proposée dans le cadre de ce travail comporte quatre étapes, illustrées par le schéma de la Figure 3.1. Chacune de ces étapes fera l'objet d'une section de ce chapitre.

L'acquisition des données utilisées pour la localisation constitue la première phase du processus proposé. Ces données sont les puissances reçues sur les porteuses GSM qui constitueront des *fingerprints* pour des positions connues. Trois campagnes de mesures, avec les dispositifs présentés en section 2.3, sont décrites dans la section suivante. Une étape de traitement des *fingerprints* utilisés peut s'avérer nécessaire (cela dépend de la méthode d'apprentissage statistique utilisée) avant la construction du modèle. En effet, disposant des puissances reçues sur toutes les porteuses GSM, il s'agit de savoir si toutes, ou seulement certaines d'entre elles, apportent de l'information sur la position. Nous présenterons, dans la section 3.3, des critères et méthodes qui permettent de sélectionner les porteuses qui sont

réellement pertinentes afin de réduire la dimension du problème, et donc de construire un modèle de localisation simple et fiable.

Un bâtiment comportant généralement plus de deux pièces, le problème de classification est dit « multi-classes ». Il est alors nécessaire de définir une stratégie de classification. Dans cette étude, une stratégie mettant en œuvre plusieurs classifieurs est utilisée. Une règle de décision doit alors être mise en œuvre pour la prédiction de la classe : ceci constitue la dernière étape du processus proposé.

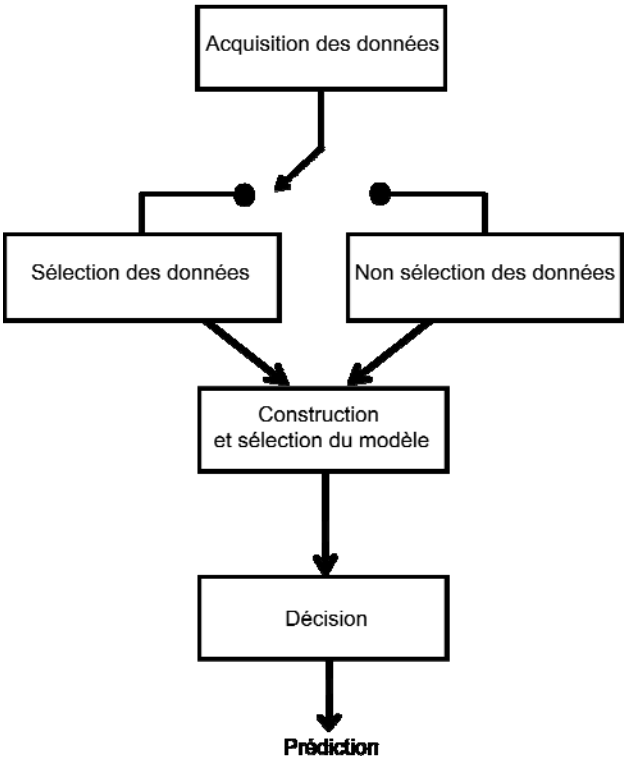

Figure 3.1. Schéma du processus de localisation.

3.2. Acquisition des données

La première étape du processus de localisation proposé est l'acquisition de bases de *fingerprints* décrivant l'environnement intérieur étudié. Ces mesures serviront à construire le modèle de correspondance entre les *fingerprints* et les positions. Dans le cadre de cette thèse, trois bases de données ont été collectées. Elles seront désignées dans la suite comme base *Home*, base *Lab* et base *Minipegs*.

3.2.1. La base *Home*

C'est la première base de *fingerprints* qui a été construite pour cette étude. Les mesures ont été collectées dans un appartement situé au cinquième et dernier étage d'un immeuble situé à Paris. Les mesures de puissance reçue sur les porteuses GSM ont été effectuées avec la version 1998 du mobile à trace TEMS. Cette version permet d'accéder aux porteuses des bandes de GSM 900 et 1800 ; la bande GSM étendue n'est pas accessible par cette version.

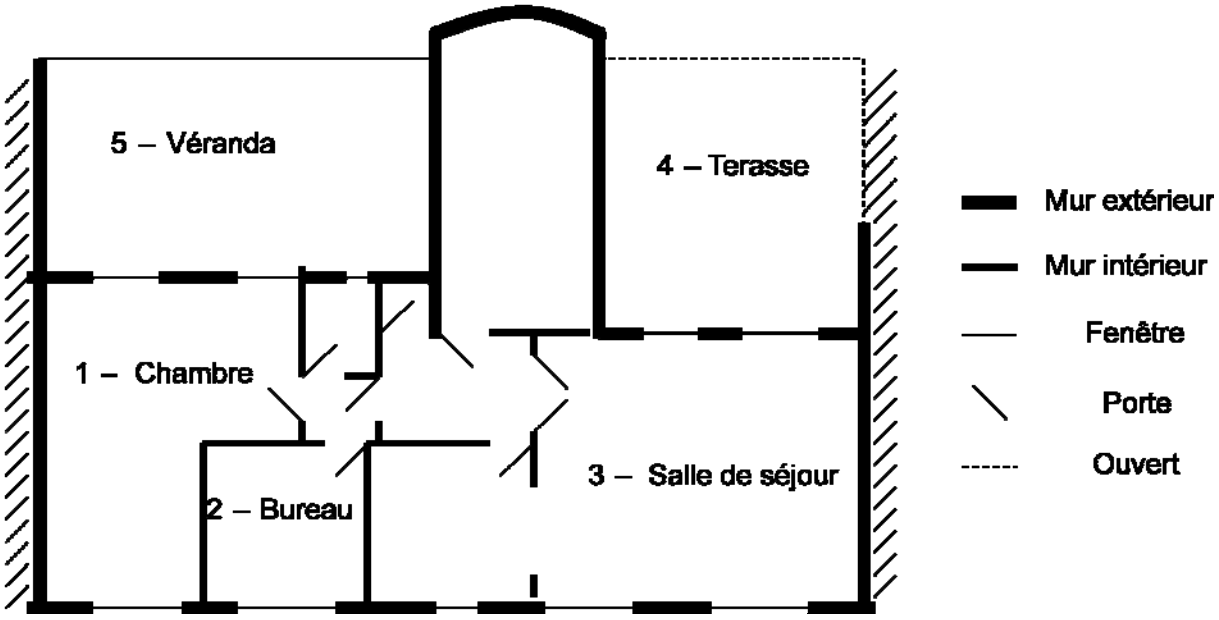

Figure 3.2. Schéma du plan de l'appartement de la base *Home*.

Sur une durée d'un mois, 241 balayages de la totalité des porteuses de la bande GSM (498 porteuses), ont été effectués à raison de deux par jour dans les cinq pièces de l'appartement soit environ 48 *fingerprints* par pièce. Le plan de l'appartement ainsi que les pièces concernées par cette étude sont indiqués sur la Figure 3.2. Les endroits dans lesquels les mesures ont été effectuées ont été choisis pour pouvoir poser convenablement le dispositif de mesure et l'ordinateur auquel il est connecté. À chaque balayage a été affecté le numéro de la pièce dans laquelle il a été effectué. Cette dernière est l'étiquette de la position que le système de localisation doit être capable de prédire.

3.2.2. La base *Lab*

Cette base a été établie en utilisant le module GM-862 de Télit présenté dans la section 2.3. Environ 600 balayages ont été effectués sur une période d'un mois dans un laboratoire de

recherche situé au deuxième et dernier étage d'un bâtiment surmonté d'un grenier dans la ville de Paris. Parmi les huit pièces constituant le laboratoire, cinq ont été choisies pour y effectuer les balayages. La Figure 3.2 illustre le plan du laboratoire et précise les lieux exacts où a été disposé l'appareil lors des mesures. Le module Télit prenant en considération également la bande du GSM étendu, cette base comporte les puissances de 534 porteuses par *fingerprint* ; comme pour la base *Home*, chaque *fingerprint* est étiqueté avec le numéro de la pièce correspondante.

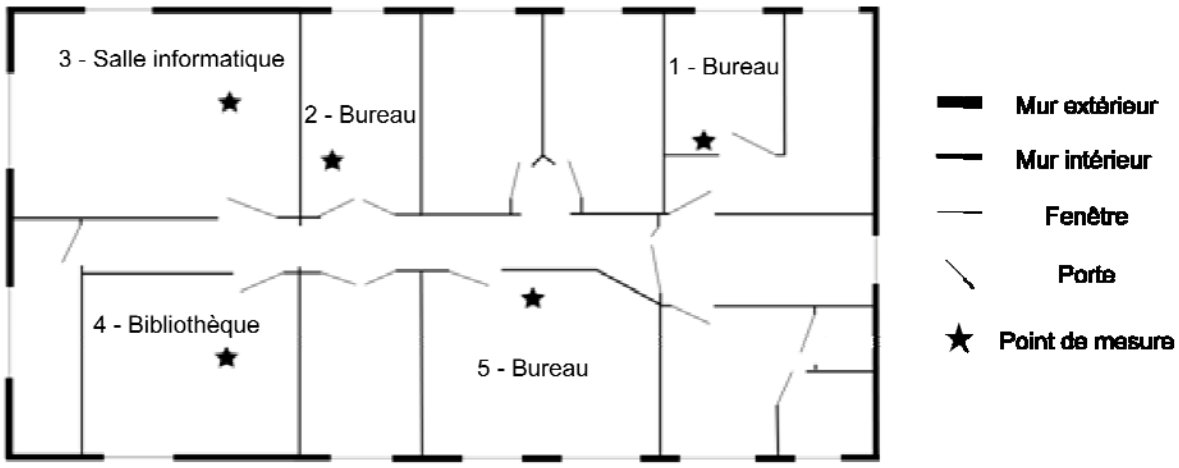

Figure 3.3. Schéma du plan du laboratoire de la base *Lab*.

Pour la collecte de chacune des deux bases de *fingerprints* présentées, un même dispositif a été utilisé (Tems pour la base *Home* et Télit pour la base *Lab*) durant toute la période de mesure.

3.2.3. La base *Minipegs*

Comme indiqué dans la section 2.5, les mesures de puissance effectuées par des dispositifs distincts, présentent des différences. Afin de tester l'influence de ces dernières sur la performance de localisation de la technique étudiée, une base de données a été construite avec plusieurs dispositifs Télit. Les mesures ont été effectuées dans les pièces numéros 3 et 4 du laboratoire de recherche de la Figure 3.3. La collecte de cette base de données s'est déroulée en deux étapes. Pendant le jour *J*, quatre appareils de mesure ont été disposés au même endroit dans chacune des pièces comme le montre la figure 3.4. Le Jour *J*+1, les

appareils de mesure ont été intervertis. Ainsi, chaque dispositif a pu effectuer environ 140 balayages des fréquences GSM pendant 24 h.

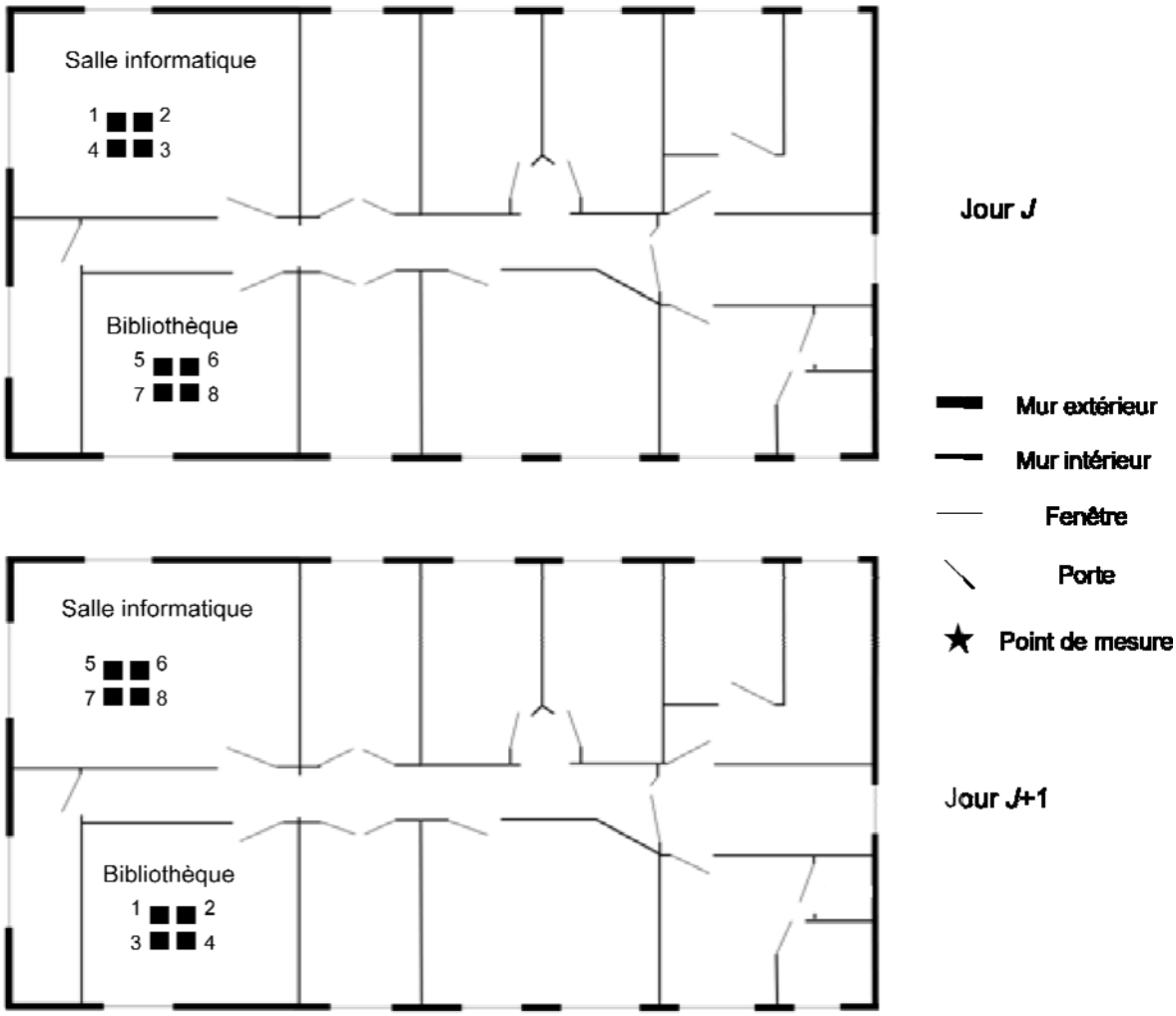

Figure 3.4. Collecte de la base *Minipegs*.

Pour la suite des étapes du processus de localisation, chacune des bases de données a été divisée en deux sous-ensembles : un ensemble d'apprentissage permettant la construction et la sélection du modèle, et un ensemble de test qui permet de simuler une situation réelle où les *fingerprints* à localiser n'ont pas été utilisés pendant la construction du système. On peut ainsi estimer la capacité de généralisation du modèle construit. Dans ce qui suit, les termes *fingerprint* et exemple seront utilisés indifféremment pour désigner une mesure de la base de données.

3.3. Sélection de porteuses GSM

Disposant des mesures de puissance sur toutes les porteuses GSM, il faut choisir le sous-ensemble de porteuses qui décrit le mieux la position, si celui-ci existe. Comme on peut le voir sur la Figure 3.1, cette étape est facultative, car toutes les porteuses mesurées peuvent être prises en considération lors de la construction du modèle de localisation. Deux critères de sélection ont été abordés lors de cette étude. Le premier repose sur le standard GSM, qui considère que, dans le cas d'une communication, les porteuses reçues avec de fortes puissances sont les plus intéressantes. L'idée est donc de vérifier si les porteuses adaptées à l'établissement d'un appel sur le réseau GSM le sont aussi pour la description de la position du mobile. Le second critère de sélection, relevant du domaine de l'apprentissage statistique, consiste à retenir les porteuses qui dotent le modèle de localisation de la meilleure capacité de discrimination.

3.3.1. Sélection sur le critère de puissance

Habituellement, la localisation utilisant des enregistrements des réseaux GSM s'appuie sur les NMR. Ces derniers incluent les puissances mesurées sur la voie balise de rattachement et les six porteuses balises les plus puissantes. En ce qui nous concerne, les deux modules Telit et TEMS enregistrent les puissances de toutes les porteuses détectables. Lors de l'analyse des mesures présentée dans la section 2.4, nous avons observé que le BSIC n'est pas décodé sur toute la durée des mesures : le statut de certaines porteuses reste donc inconnu. Afin de reproduire un NMR à partir des mesures enregistrées, et à défaut de ne garder que les voies balises identifiées, les sept porteuses les plus puissantes disponibles sont prises en considération. Cette sélection est faite sur la base d'apprentissage et permet de comparer l'approche proposée avec ce qui est habituellement adopté dans la littérature de la localisation par GSM.

Les méthodes d'apprentissage statistique que nous avons mises en œuvre (décrites dans la section 3.4), traitent des vecteurs de données de longueur fixe. Or, les 7 porteuses les plus puissantes ne sont pas les mêmes pour tous les balayages. La construction des *fingerprints* a donc été effectuée de la manière suivante : on a cherché les porteuses qui apparaissent au moins une fois parmi les 7 plus puissantes d'un balayage. Soit n_p le nombre de ces porteuses ; chaque *fingerprint* est un vecteur de dimension n_p, dont 7 composantes sont les puissances

des 7 porteuses les plus puissantes pour ce balayage, et dont les $n_p - 7$ autres composantes sont nulles. Cette procédure de construction de *fingerprints* est illustrée par la Figure 3.5.

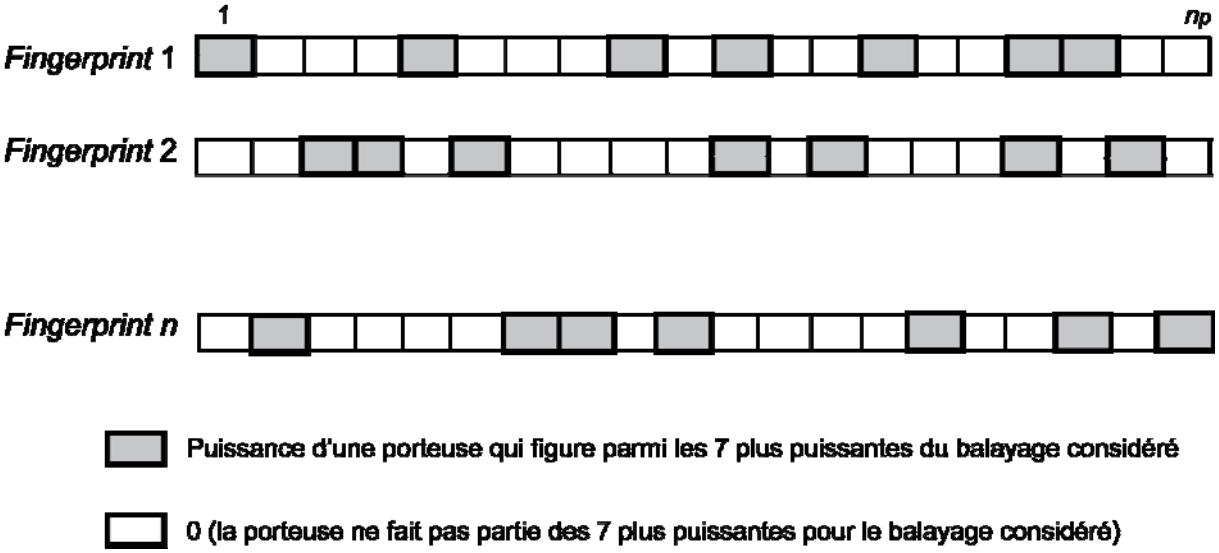

Figure 3.5. Construction des *fingerprints* des sept plus fortes porteuses.

3.3.2. Sélection sur le critère de pertinence

L'objectif de ce travail étant la localisation de la pièce où se trouve le mobile, cette démarche vise à utiliser les porteuses les plus discriminantes pour la construction du modèle de localisation. Pour cela, il faut classer les porteuses disponibles par ordre de pertinence ; nous avons procédé à ce classement par deux méthodes : l'algorithme *Orthogonal Forward Regression* (OFR) [42], et l'algorithme *Recursive Formard Elimination* (RFE).

Le premier est d'usage général, en ce sens qu'il est indépendant de la méthode de classification utilisée. C'est un pré-traitement des données, qui est réalisé avant la mise en œuvre d'algorithme d'apprentissage ; dans le jargon de la sélection de variables, cet algorithme entre dans la catégorie des *filtres.* L'algorithme RFE, en revanche, est utilisable uniquement lorsque l'algorithme de classification est une machine à vecteurs supports (SVM) ; il met en œuvre directement l'algorithme de classification lui-même, sous la forme d'une boucle de calcul « enroulée » autour de l'algorithme de classification, ce qui justifie le terme de *wrappers* utilisé pour désigner cette catégorie d'algorithmes de sélection de variables.

En raison de son caractère plus général, et aussi parce qu'il nous a permis d'obtenir de meilleurs résultats, nous présentons ici l'algorithme OFR uniquement. L'algorithme RFE est présenté succinctement dans l'annexe 2.

L'algorithme OFR est une méthode *constructive* : la liste des variables pertinentes est constituée à partir d'une liste vide, en ajoutant les variables une à une. Elle permet de sélectionner les porteuses les plus corrélées avec la position (la pièce dans ce cas) en évitant d'inclure, dans la liste des variables, les puissances des porteuses mutuellement corrélées.

Soit $\mathbf{v}_i$ le vecteur dont les composantes sont les mesures des puissances de la porteuse *i*, et soit $\mathbf{y}$ le vecteur dont les composantes sont les étiquettes (-1 et +1) des deux classes à discriminer. Si les composantes de $\mathbf{v}_i$ sont centrées (de moyenne nulle), le coefficient de corrélation entre $\mathbf{v}_i$ et $\mathbf{y}$ est le carré du cosinus de l'angle entre $\mathbf{v}_i$ et $\mathbf{y}$:

$$\cos^2\left(\mathbf{v}_i, \mathbf{y}\right) = \left(\mathbf{v}_i \cdot \mathbf{y}\right)^2 \Big/ \left\|\mathbf{v}_i\right\|^2 \left\|\mathbf{y}\right\|^2 \tag{3.1}$$

Celui-ci vaut 1 si $\mathbf{v}_i$ et $\mathbf{y}$ sont colinéaires, c'est-à-dire si $\mathbf{v}_i$ explique complètement $\mathbf{y}$, et il vaut 0 si $\mathbf{v}_i$ et $\mathbf{y}$ ne sont pas corrélés.

Le classement des porteuses par ordre de pertinence se fait comme suit. La première porteuse sélectionnée est celle dont puissances mesurées sont les plus corrélées aux étiquettes des classes. Le vecteur des puissances des autres porteuses, et le vecteur $\mathbf{y}$, sont alors projetés dans le sous-espace orthogonal au vecteur des puissances de la première porteuse sélectionnée. La procédure est itérée dans ce sous-espace, et ainsi de suite jusqu'à ce que toutes les porteuses soient classées, ou qu'un un critère d'arrêt soit satisfait.

Une fois le classement terminé, il faut sélectionner le nombre de porteuses nécessaires pour la construction d'un modèle de localisation obéissant aux critères cités dans la section 3.1. Ce nombre est fixé par la procédure de validation croisée [43], décrite dans la section suivante, qui tient compte de la méthode d'apprentissage statistique utilisée.

3.4. Construction et sélection du modèle

C'est durant cette étape qu'est construit le modèle qui décrit les positions à partir des *fingerprints*. Certaines des méthodes d'apprentissage utilisées dans ce travail (décrites dans cette section), mettent en œuvre des hyperparamètres dont les valeurs doivent être choisies de manière à optimiser les performances du classifieur. Ce choix est effectué par une procédure rigoureuse de validation croisée, en même temps que la sélection des variables qui

ont été préalablement classées par ordre de pertinence comme indiqué dans la section précédente.

La validation croisée

L'ensemble des mesures de chaque base d'apprentissage est partitionné en plusieurs sous-ensembles disjoints ou *plis* (6 pour la base *Home* et 10 pour la base *Lab* qui comporte un plus grand nombre de mesures). La procédure de validation croisée, illustrée par la Figure 3.6, consiste à effectuer l'apprentissage d'un modèle, muni d'un ensemble d'hyperparamètres, en utilisant tous les plis sauf un (*pli de validation*), et à calculer l'erreur de classification commise par le modèle sur les exemples de ce dernier. On itère autant de fois qu'il y a de plis, de telle sorte que chaque exemple se trouve une fois et une seule dans un pli de validation. Une fois la procédure terminée, on calcule l'erreur de classification commise par les modèles sur les exemples lorsque ceux-ci étaient dans un pli de validation, ce qui constitue une estimation de la capacité de généralisation du modèle muni de l'ensemble d'hyperparamètres considéré. On réalise cette procédure pour plusieurs ensembles d'hyperparamètres, afin de déterminer celui qui confère au modèle la meilleure capacité de généralisation. Plusieurs tirages des plis de validation ont été effectués permettant ainsi d'obtenir une moyenne et un écart-type sur les performances de validation croisée.

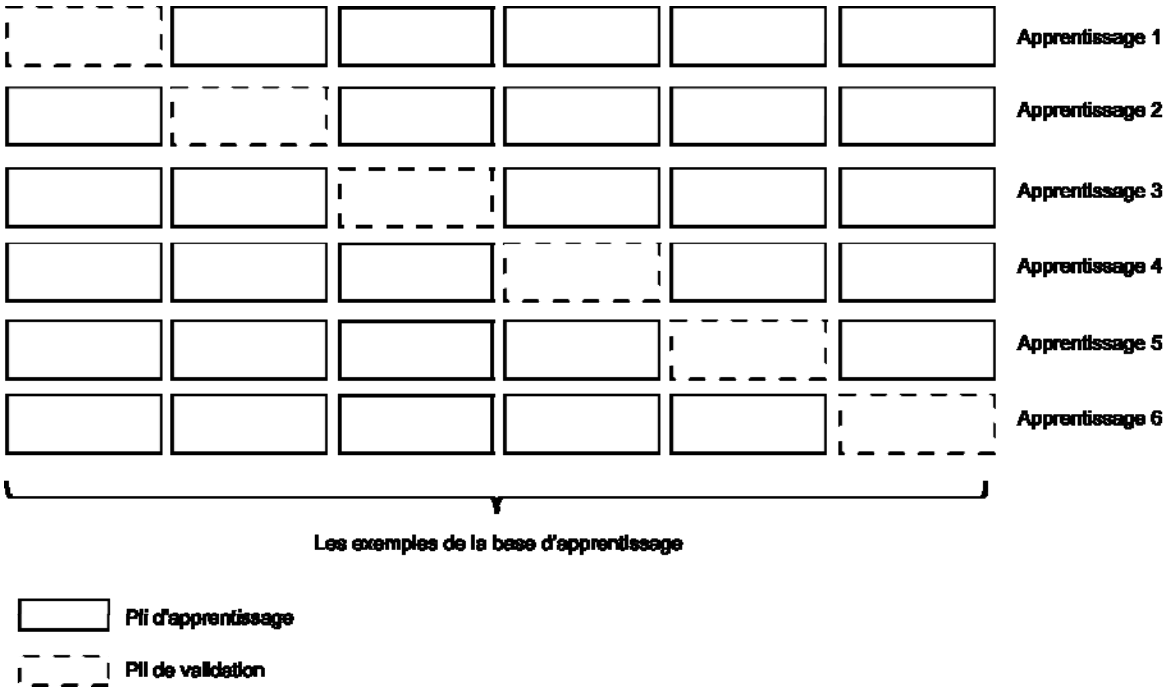

Figure 3.6. Exemple de la procédure de validation croisée avec six plis.

La stratégie multi-classes

Pour résoudre un problème de discrimination entre plusieurs classes, ce qui est le cas dans la présente étude, on met en œuvre habituellement un ensemble de classifieurs chargés de discriminer deux classes. Deux stratégies peuvent être considérées : *un contre tous* et *un contre un.*

La stratégie *un contre tous* consiste à utiliser autant de classifieurs que de classes à séparer, chaque classifieur ayant pour tâche de séparer les exemples d'une classe de ceux de toutes les autres classes. Dans la stratégie *un contre un*, chaque classifieur sépare les exemples d'une classe de ceux d'une autre classe ; pour un problème à c classes, cette stratégie nécessite $c(c-1)/2$ classifieurs.

Quelle que soit la stratégie utilisée, il est nécessaire de mettre en œuvre un algorithme de prise de décision finale, qui effectue une synthèse des décisions individuelles des classifieurs à deux classes. Cette règle de décision est décrite dans la section 3.5.

3.4.1. Les *K*-plus proches voisins (*K*-PPV)

La méthode des *K*-PPV, introduite dans [44], est la plus simple des méthodes de classification. Elle consiste à constituer une base de référence associant, dans notre cas, chaque balayage de fréquences au numéro de la pièce dans laquelle il a été mesuré, et à affecter le balayage à localiser à la classe qui est la plus représentée dans ses K plus proches voisins dans la base de référence. Le plus souvent, la distance utilisée dans l'espace des variables (ici des puissances reçues) est la distance euclidienne. La valeur de l'hyperparamètre K est déterminée par validation croisée.

Cette méthode présente l'avantage de ne pas nécessiter une phase d'apprentissage et d'aborder le problème de localisation comme une classification multi-classes sans avoir recours à la décomposition en sous-problèmes à deux classes. En revanche, prédire la pièce dans laquelle a été mesuré un *fingerprint* est numériquement coûteux, en particulier si le nombre de *fingerprints* de la base de données est important, puisqu'il est nécessaire de calculer la distance entre le *fingerprint* du point à localiser à tous les *fingerprints* de la base de données.

Dans cette étude, la classification par les *K*-PPV a été mise en œuvre essentiellement à des fins de comparaison, cette méthode étant très fréquemment mise en œuvre dans la littérature de la localisation par *fingerprints* [29, 32].

3.4.2. L'analyse discriminante linéaire

L'analyse discriminante linéaire ou *Linear Discriminant Analysis* (LDA) est une méthode de classification qui a pour objectif de trouver surface de séparation linéaire qui minimise le taux d'erreur de classification, c'est-à-dire la direction dans l'espace des variables (les puissances reçues) où les densités de probabilité des exemples des différentes classes sont le mieux séparées. Ainsi la LDA, comme toutes les méthodes de classification fondées sur l'estimation des probabilités d'appartenance des objets à classer aux classes, minimise l'erreur de classification en attribuant à chaque *fingerprint* **x** la pièce *i* pour laquelle la probabilité d'appartenance est la plus grande. En supposant un problème à deux classes (pièces) c_1 et c_2 cela veut dire que

$$P\left(c_1\middle|\mathbf{x}\right) > P\left(c_2\middle|\mathbf{x}\right) \tag{3.3}$$

$P(c_1|\mathbf{x})$ et $P(c_2|\mathbf{x})$ sont les probabilité à posteriori des classes. Elles sont reliées aux probabilités a priori des classes $P(c_k)$ et aux densités de probabilités des variables dans chaque classe $P(\mathbf{x}|c_k)$ par la relation de Bayes

$$P\left(c_1\middle|\mathbf{x}\right) = \frac{P\left(\mathbf{x}\middle|c_1\right)P\left(c_1\right)}{\sum_{\forall k} P\left(\mathbf{x}\middle|c_k\right)P\left(c_k\right)} \tag{3.4}$$

En substituant cette valeur dans l'expression (3.3) on obtient :

$$\frac{P\left(\mathbf{x}\middle|c_1\right)P\left(c_1\right)}{\sum_{\forall k} P\left(\mathbf{x}\middle|c_k\right)P\left(c_k\right)} > \frac{P\left(\mathbf{x}\middle|c_2\right)P\left(c_2\right)}{\sum_{\forall k} P\left(\mathbf{x}\middle|c_k\right)P\left(c_k\right)} \tag{3.5}$$

Puisque les dénominateurs sont identiques, ceci peut s'écrire sous la forme :

$$P\left(\mathbf{x}\middle|c_1\right)P\left(c_1\right) > P\left(\mathbf{x}\middle|c_2\right)P\left(c_2\right) \tag{3.6}$$

La densité de probabilité des variables **x** dans la classe *c* est inconnue, et son estimation est soumise à la « malédiction de la dimensionnalité » : le nombre d'exemples nécessaires croît exponentiellement avec le nombre de variables. Le problème que nous nous posons étant de grande dimension (une ou plusieurs centaines), l'estimation des densités de

probabilités des variables est impossible. On peut alors avoir recours à des hypothèses sur la forme de ces dernières.

L'analyse discriminante linéaire repose sur l'hypothèse que les variables suivent des lois de probabilité gaussiennes dont les matrices de covariance sont identiques :

$$P\left(\mathbf{x}\middle|c\right)=\frac{1}{\left(2\pi\right)^{n/2}\sqrt{\left|\mathbf{Cov}\right|}}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_c\right)^T\mathbf{Cov}\left(\mathbf{x}-\boldsymbol{\mu}_c\right)\right) \tag{3.7}$$

où n est la dimension du *fingerprint* $\mathbf{x}$ (le nombre de porteuses considérées dans ce cas), $\boldsymbol{\mu}_c$ est le vecteur des centres des gaussiennes pour la classe c, et **Cov** est la matrice de covariance des exemples.

L'équation (3.6) s'écrit alors :

$$\ln\left(P\left(c_1\right)\right)+\boldsymbol{\mu}_{c_1}\mathbf{Cov}^{-1}\mathbf{x}^T-\frac{1}{2}\boldsymbol{\mu}_{c_1}\mathbf{Cov}^{-1}\boldsymbol{\mu}_{c_1}^T>\ln\left(P\left(c_2\right)\right)+\boldsymbol{\mu}_{c_2}\mathbf{Cov}^{-1}\mathbf{x}^T-\frac{1}{2}\boldsymbol{\mu}_{c_2}\mathbf{Cov}^{-1}\boldsymbol{\mu}_{c_2}^T \tag{3.8}$$

Ainsi, la règle de décision dans le cas de la LDA utilise la fonction :

$$f_c\left(\mathbf{x}\right)=\boldsymbol{\mu}_c\mathbf{Cov}^{-1}\mathbf{x}^T-\frac{1}{2}\boldsymbol{\mu}_c\mathbf{Cov}^{-1}\boldsymbol{\mu}_c^T+\ln\left(P\left(c\right)\right) \tag{3.9}$$

Un exemple $\mathbf{x}$ est attribué à la classe c pour laquelle $f_c(\mathbf{x})$ est la plus grande. La frontière séparatrice entre les classes c_1 et c_2 dans le cas de la LDA est le lieu des points, dans l'espace des variables, pour lesquels $f_{c1}=f_{c2}$. Une classification par LDA nécessite donc peu de calculs, mais les hypothèses sur lesquelles repose cette classification sont rarement vérifiées en pratique. De plus, dans le cas où le nombre de variables (porteuses dans ce cas) est supérieur au nombre d'exemples dans une classe, la matrice de covariance n'est pas de rang plein. L'étape de sélection de porteuses est alors indispensable. Sachant qu'on dispose de bases de données de grande dimensionnalité (498 pour la base *Home* et 534 pour la base *Lab*), nous avons également mis en œuvre une autre méthode d'apprentissage statistique, plus coûteuse en temps de calcul mais qui offre de meilleures garanties théoriques et conserve toute son efficacité pour traiter des problèmes de grande dimension.

3.4.3. Les machines à vecteurs supports

L'objectif principal visé par ce travail est de trouver une frontière permettant de séparer, dans l'espace de puissances reçues, les *fingerprints* mesurés dans différentes pièces. Les machines à vecteurs supports ou *Support vector machines* (SVM) pour la classification [45] possèdent un mécanisme qui permet le contrôle de la complexité du modèle, et donc évite le

phénomène de surajustement du modèle aux données qui apparaît dans toute méthode de classification lorsque la complexité du modèle est trop grande.

Lorsque les exemples dont on dispose pour les deux classes sont linéairement séparables, c'est-à-dire lorsqu'il existe au moins un hyperplan qui sépare ces exemples sans erreur, une SVM permet de trouver l'hyperplan séparateur optimal, c'est-à-dire l'hyperplan qui classe sans erreur tous les exemples de la base d'apprentissage, et dont la distance aux exemples d'apprentissage les plus proches est maximale.

Supposons la base d'apprentissage constituée de $\{\mathbf{x}_i, y_i\}$, $i = 1\ldots l$, et $y_i \in \{-1, +1\}$, où $\mathbf{x}_i$ est i ème le vecteur de puissances reçues, y_i est l'étiquette du *fingerprint* (-1 si le *fingerprint* i appartient à une des deux classes, et +1 s'il appartient à l'autre), et l est le nombre d'exemples de la base.

Soit :

$$\mathbf{x} \cdot \mathbf{w} + b = 0 \tag{3.10}$$

l'équation recherchée, où $\mathbf{w}$ et b doivent être estimés par apprentissage. Cet hyperplan devant classer sans erreur tous les exemples de l'ensemble d'apprentissage, ses paramètres doivent obéir à la contrainte :

$$y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) > 0 \,. \tag{3.11}$$

La *marge géométrique* du classifieur est la distance entre l'hyperplan séparateur et l'exemple qui en est le plus proche. Elle a pour expression :

$$\frac{1}{\|\mathbf{w}\|} \min_i \left(y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \right) \tag{3.12}$$

L'équation de l'hyperplan étant définie à un facteur multiplicatif près, on définit *l'hyperplan canonique*, tel que :

$$\min_i \left(y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \right) = 1 \,, \tag{3.13}$$

de sorte que la marge géométrique vaut $\frac{1}{\|\mathbf{w}\|}$.

L'hyperplan séparateur optimal est donc l'hyperplan canonique de marge géométrique maximale : on peut l'obtenir en résolvant numériquement le problème d'optimisation quadratique sous contraintes suivant :

$$\text{minimiser } \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.14}$$

$$\text{sous la contrainte } y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 \ \forall \text{i} \,. \tag{3.15}$$

Un tel classifieur est appelé SVM « à marge dure », illustré par la Figure 3.7. La minimisation de $\|\mathbf{w}\|^2$ permet de contrôler la complexité du modèle, et les contraintes d'inégalité garantissent que tous les exemples sont bien classés, et sont à une distance de l'hyperplan qui est supérieure ou égale à la marge géométrique. Les exemples qui sont à une distance juste égale à la marge géométrique sont appelés *vecteurs supports*, ce qui justifie le nom de la méthode.

Pour éviter que la position de l'hyperplan séparateur ne dépende trop fortement des vecteurs supports, on peut admettre que, à la fin de l'apprentissage, la distance de l'hyperplan à certains des exemples soit inférieure à $1/\|\mathbf{w}\|$, voire mal classés. Les contraintes deviennent alors

$$y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 - \xi_i \quad \text{avec} \ \ \xi_i \geq 0 \ \ \forall \text{i} \,, \tag{3.16}$$

et, pour limiter le nombre d'exemples trop proches de l'hyperplan séparateur ou mal classés, on minimise la fonction

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i^2 \tag{3.17}$$

où $C \geq 0$ est un hyperparamètre, appelé *constante de régularisation*. Ce dernier détermine un compromis entre la qualité du classifieur et son ajustement aux vecteurs supports. Si la constante C est très grande, l'optimisation fournit des ξ_i très voisins de zéro, de sorte que l'on est ramené à une machine à vecteurs supports à marge dure.

Une fois les paramètres $\mathbf{w}$, b, ξ_i et l'hyperparamètre C fixés pendant les phases d'apprentissage et de validation croisée, l'étiquette y attribuée au *fingerprint* $\mathbf{x}$ a pour expression

$$y = \text{sgn}\left(\mathbf{x} \cdot \mathbf{w} + b\right). \tag{3.18}$$

On peut reformuler le problème d'optimisation sous contrainte dont les paramètres de l'hyperplan séparateur optimal sont les solutions de la manière suivante (forme *duale*) : minimiser

$$F(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right) \tag{3.19}$$

sous les contraintes

$$\sum_{i=1}^{N} \alpha_i \, y_i = 0 \quad , \quad \alpha_i \geq 0 \, . \tag{3.20}$$

Une fois le vecteur optimal **α*** déterminé, le vecteur des paramètres **w** est donné par l'expression

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i^* \, y_i \mathbf{x}_i \, . \tag{3.21}$$

Cette formulation montre que la position de l'hyperplan séparateur optimal ne dépend que des produits scalaires entre les vecteurs qui décrivent les exemples. Cette remarque est mise à profit lorsque l'on cherche à séparer de manière optimale des exemples qui ne sont pas linéairement séparables : on cherche alors une transformation $\mathbf{z} = \varphi(\mathbf{x})$, si elle existe, telle que, dans l'espace des nouvelles variables **z**, les exemples d'apprentissage soient linéairement séparables. L'hyperplan séparateur optimal, s'il existe, ne dépend donc que des produit scalaire $\mathbf{z}_i \cdot \mathbf{z}_j = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$. La fonction $K(\mathbf{x}, \mathbf{y})$ est appelée *fonction noyau*. La fonction noyau la plus fréquemment utilisée est le noyau gaussien, qui contient un hyperparamètre σ à déterminer par validation croisée:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left[-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2} \right] \tag{3.22}$$

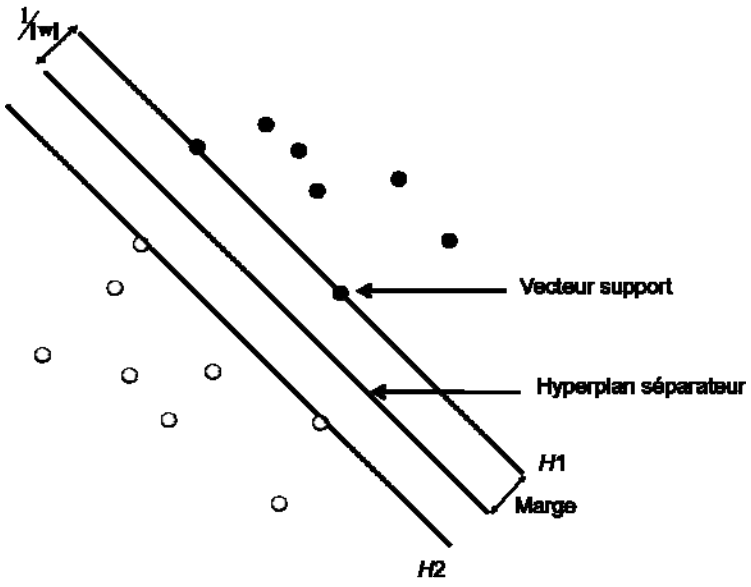

Figure 3.7. Hyperplan séparateur SVM dans le cas d'exemples linéairement séparables.

Dans le présent travail, les classifieurs SVM utilisés ont été implémentés à l'aide de la boîte à outils *The Spider* [46].

3.4.4. Les machines à vecteurs supports *transductives*

Pendant la collecte des données présentées dans la section 3.2, de manière systématique, chaque fois qu'un balayage est effectué, la pièce correspondante est notée et associée à celui-ci. Cette opération d'étiquetage des exemples allonge le temps nécessaire pour constituer un ensemble de mesures et nécessite une intervention humaine. Il est donc probable que l'on soit confronté à des ensembles partiellement étiquetés. Afin de faire face à cette situation, on peut avoir recours à des méthodes d'apprentissage semi-supervisé permettant de tirer profit de ces mesures non étiquetées. Des simulations de cette situation ont été créées à partir des bases *Home* et *Lab* en gardant qu'une partie des étiquettes collectées.

Plusieurs méthodes de classification semi-supervisée sont décrites dans la littérature [47]. Encouragés par des niveaux de performances satisfaisants obtenus avec des classifieurs utilisant des machines à vecteurs supports (voir chapitre 4), nous proposons de mettre en œuvre une méthode à noyaux appelée « machines à vecteurs supports *transductives* » ou *Transductive Support Vector Machines* (TSVM) [48].

La méthode des TSVM a été appliquée avec succès à la reconnaissance de texte [48] et au traitement d'images [49]. Le principe de la méthode des TSVM est semblable à celui des SVM classiques. En effet, il s'agit de trouver l'hyperplan séparateur optimal dont la distance aux exemples étiquetés et non étiquetés est maximale. L'algorithme d'apprentissage TSVM, schématisé sur la Figure 3.8, est constitué de deux étapes.

La première étape consiste à classer les exemples étiquetés à l'aide de SVM. Un modèle de la forme de l'expression (3.9) est obtenu et utilisé pour attribuer des étiquettes aux exemples non étiquetés de l'ensemble d'apprentissage. Une fois tous les exemples de la base d'apprentissage étiquetés, la seconde étape consiste à trouver l'hyperplan séparateur situé le plus loin possible des exemples étiquetés et non étiquetés. L'équation de ce nouvel hyperplan est :

$$\mathbf{x} \cdot \mathbf{w}_{TSVM} + b_{TSVM} = 0 \tag{3.23}$$

Ceci revient à :

$$\text{minimiser } \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{u} L\left(y_i\left(\mathbf{x}_i \cdot \mathbf{w}_{TSVM} + b_{TSVM}\right)\right) + C_2 \sum_{j=u+1}^{l} L\left(y_j\left(\mathbf{x}_j \cdot \mathbf{w}_{TSVM} + b_{TSVM}\right)\right)$$

où $L(z) = (1\text{-}z)_+ = \max(1\text{-}z,0)$, $\{\mathbf{x}_i, y_i\}$ sont les exemples d'apprentissage non étiquetés et les étiquettes qui leur sont attribuées pendant la première phase, $\{\mathbf{x}_i, y_i\}$ sont les exemples étiquetés, C_1 et C_2 sont les paramètres de régularisation, u est le nombre d'exemples non étiquetés et l le nombre total d'exemples d'apprentissage.

L'optimisation procède par permutation des étiquettes prédites. Une condition permet de déterminer si deux exemples ont des étiquettes permutables. Si une permutation conduit à une diminution du coût à optimiser, elle est acceptée. Sinon, elle est rejetée. L'optimisation s'arrête lorsque qu'il n'y a plus d'exemples dont les étiquettes sont permutables [48].

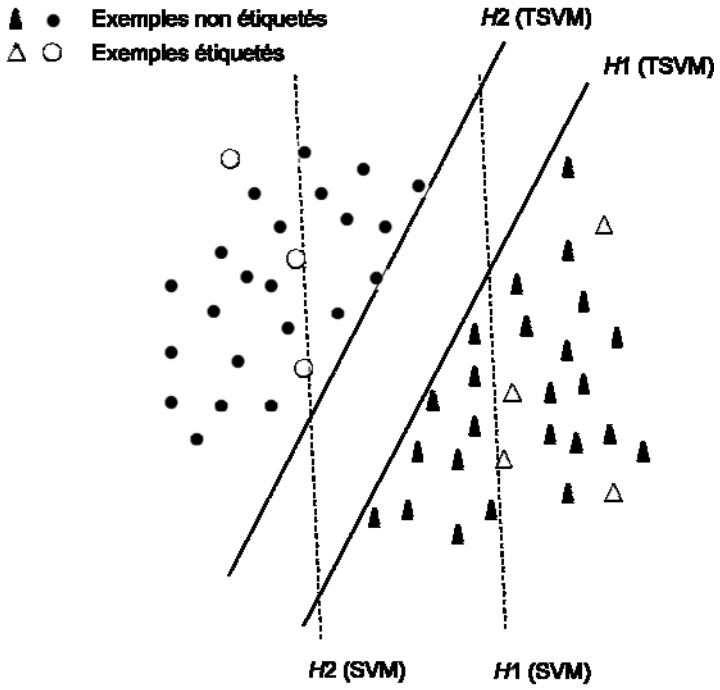

Figure 3.8. Principe de fonctionnement de l'algorithme TSVM.

Comme les SVM classiques, les TSVM peuvent avoir recours à des fonctions noyaux non linéaires permettant de passer dans un espace où les exemples sont séparables. Ceci introduit des hyperparamètres dont les valeurs sont estimées par la procédure de validation croisée. Dans notre étude, les classifieurs TSVM ont été mis en œuvre à l'aide de la boîte à outils SVM^{light} [50].

3.5. Règle de décision

Afin de gérer le problème de localisation abordé, qui consiste en la séparation des cinq pièces (voir section 3.2) à partir de *fingerprints*, dix classifieurs *un contre un* on été mis en œuvre. Ces derniers peuvent être de type KPPV, LDA, SVM ou TSVM et permettent de séparer chaque couple de pièces. Le principe de cette stratégie est illustré par la Figure 3.9.

Pendant la phase de localisation, un *fingerprint* de test doit être présenté à chacun de ces classifieurs, qui attribue à cet exemple le numéro d'une pièce. Ainsi, à chaque exemple seront associées plusieurs pièces. Une décision doit donc être prise en fonction de ces résultats. La règle de décision adoptée dans cette étude est fondée sur un principe de vote. Il suffit d'observer les sorties de tous les classifieurs. Si la sortie du classifieur séparant la pièce c_1 de la pièce c_2 vaut 1, la note de la pièce c_1 est incrémentée de 1. Sinon, c'est la note de la pièce c_2 qui est incrémentée. La pièce associée à l'exemple de test est celle qui possède la note la plus élevée.

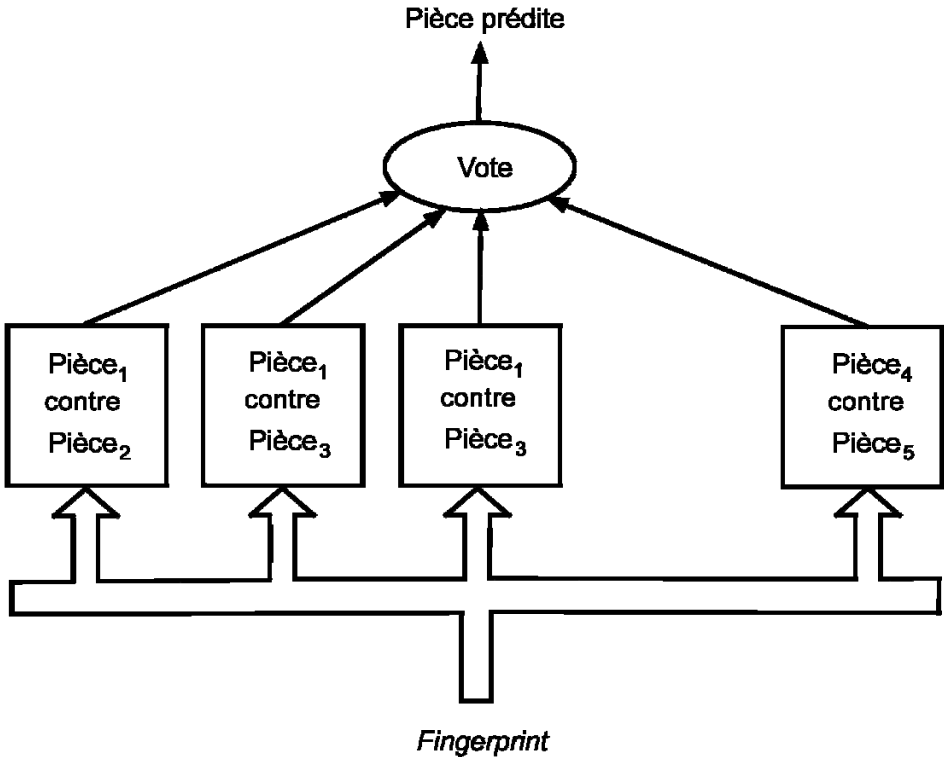

Figure 3.9. Localisation par classifieurs *un contre un* et système de vote.

Le nombre de classifieurs nécessaires pour une stratégie *un contre un* varie comme le carré du nombre de classes ; si celui-ci est grand, une méthode reposant sur un arbre de

décision [51] peut nécessiter moins de calculs. Dans ce cas, un *fingerprint* de test n'est présenté à un classifieur *un contre un* que si cette opération est nécessaire. Pendant la phase de test, les classifieurs ne sont donc pas tous sollicités. Dans la présente étude, puisque le nombre de classes est raisonnable et que la décision par vote a conduit à de bons résultats (voir chapitre 4), la méthode de l'arbre de décision n'a pas été implémentée.

Quand la stratégie *un contre tous* est adoptée, le nombre de classifieurs nécessaires est égal au nombre de classes à séparer. Chaque classifieur i est chargé de séparer les *fingerprints* d'une classe c_i du reste des *fingerprnits*. Dans ce cas, et pendant la phase de localisation, l'ensemble de test est également présenté à chaque classifieur. Un classifieur est dit actif si sa sortie vaut 1. Sinon, sa sortie vaut -1 et il est dit inactif. La règle de décision pour attribuer une classe à chaque exemple est appelée le *conventional recipe* :

1. Si un unique classifieur i est actif, l'exemple est affecté à la classe c_i.
2. Si plusieurs classifieurs ou, au contraire, aucun n'est actif, l'exemple est affecté à la classe c_i où i est l'indice du classifieur dont la valeur de sortie, avant seuillage, est la plus grande.

Plus de détails sur cette approche ainsi que les résultats obtenus, peuvent être trouvés dans l'annexe 3.

3.6. Conclusion

Cette partie du manuscrit a été consacrée à la présentation du processus de localisation proposé. Les bases de données collectées pour cette étude, leur prétraitement, la construction du modèle ainsi que la règle de décision, qui constituent les différentes étapes de processus, ont été présentées. Pendant l'étape de construction des bases de données, les mesures de puissances ont été collectées afin de tester l'influence des effets, observés lors de l'étude des signaux GSM, sur les performances de localisation. Pour le prétraitement de ces données, des méthodes issues des domaines du GSM et de l'apprentissage statistique ont été proposées. Des méthodes d'apprentissage adaptées aux données ont ensuite été utilisées pour la construction des modèles de localisation. Enfin, une stratégie permettant de gérer le problème multi-classes et la prise de décision a été adoptée. Cette démarche est illustrée par la Figure 3.10.

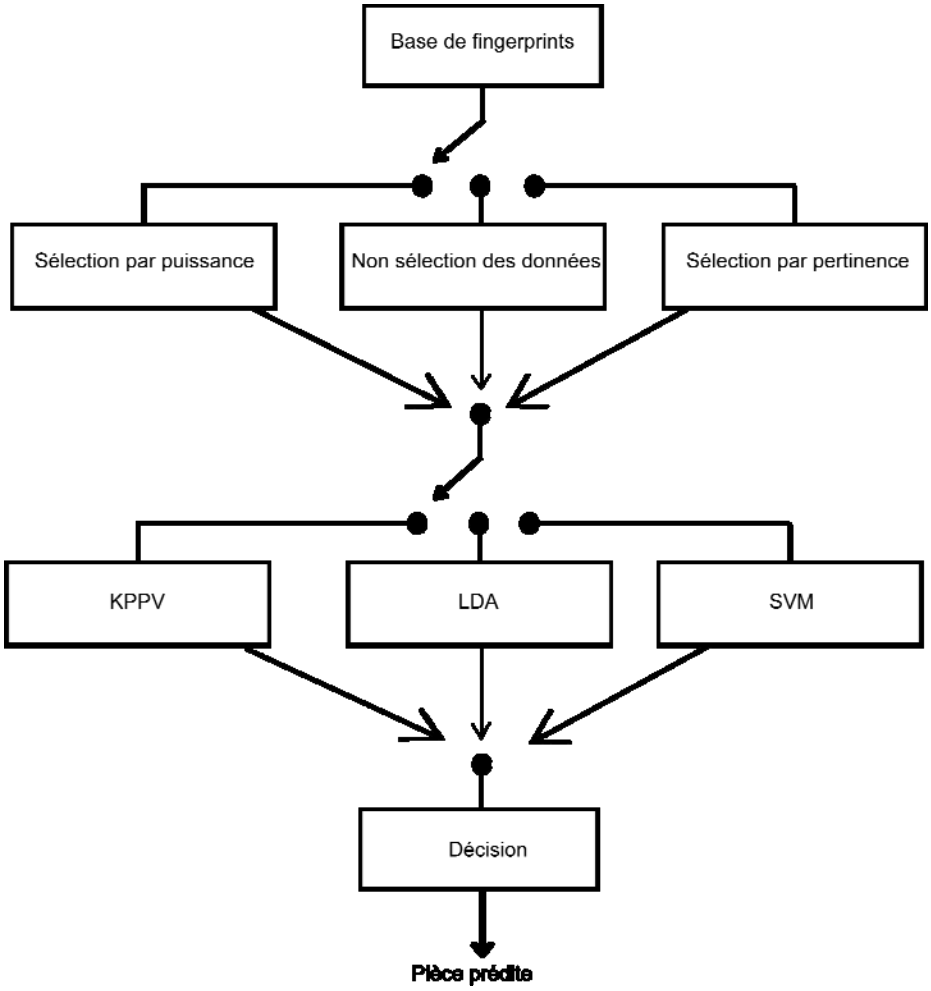

Figure 3.10. Processus de localisation détaillé.

Dans le chapitre suivant, les résultats obtenus par la mise en œuvre de chaque branche du schéma de la Figure 3.10 seront présentés et discutés. Ces résultats sont les performances de localisation du système, avec chacune des méthodes d'apprentissage statistique proposée, sur un ensemble de test. Ceci permettra de déterminer l'apport de chaque méthode au problème de localisation.

Chapitre 4. Résultats et discussions

4.1. Introduction

Ce chapitre est organisé en quatre sections. Dans la première partie, les résultats de l'application des *K*-PPV, LDA et SVM sur les bases de données *Home* et *Lab* sont présentés. Une étude des porteuses sélectionnées pour la localisation est également abordée dans cette partie. La deuxième section est consacrée à l'étude de l'influence de la variabilité temporelle des mesures de puissance sur la performance des classifieurs. Les résultats de localisation sur la base *Minipegs* font l'objet de la troisième partie afin de tester l'influence la différence entre les dispositifs de mesure sur la performance de localisation. Enfin, la dernière partie présente les performances des classifieurs TSVM sur les bases de données *Home* et *Lab*.

4.2. Résultats de localisation sur les bases *Home* et *Lab*

Comme indiqué dans la section 3.1, on se place, dans cette étude, dans le cadre de la localisation en intérieur. En particulier, il s'agit de localiser la pièce dans laquelle se trouve un objet ou un appareil mobile à partir des mesures des puissances reçues sur les porteuses de la bande GSM. Dans ce cas, la performance de localisation représente le pourcentage de *fingerprints* que le système localise dans la bonne pièce.

Dans le cas d'une stratégie *un contre tous*, l'application de l'algorithme de Ho & Kashyap [52] nous a montré que les exemples d'apprentissage ne sont pas linéairement séparables. La mise en œuvre de classifieurs non linéaires est alors indispensable. En revanche, dans le cas de la stratégie *un contre un*, où chaque couple de classes est traité indépendamment des autres, les frontières de séparations sont alors linéaires, ce qui simplifie l'implémentation ; le prix à payer réside dans le fait que, pour un problème à c classes, $c(c\text{-}1)/2$ classifieurs doivent alors être considérés. Dans le cadre de cette étude (c = 5), le nombre de classifieurs à construire reste raisonnable, donc cette stratégie est applicable. Pour chaque classifieur, les variables et les hyperparamètres des modèles sont sélectionnés par validation croisée, puis le classifieur est appliqué à une base de test, comme indiqué dans le chapitre précédent. Nous détaillerons dans ce mémoire les résultats obtenus en mettant en œuvre la stratégie des

classifieurs *un contre un*. Les résultats de l'implémentation des classifieurs *un contre tous* sont décrits dans l'annexe 3.

Dans les travaux décrits dans cette partie du chapitre, un ensemble de test est tiré aléatoirement de chacune des bases de données étudiées. Le reste des données est utilisé pour la procédure de validation croisée décrite dans la section 3.4. Ce tirage est fait en s'assurant d'avoir la même proportion d'exemples de chaque classe dans les deux ensembles (apprentissage et test). Environ 60 exemples parmi les 241 qui constituent la base *Home* forment l'ensemble de test. La base *Lab*, qui est plus importante a permis de construire un ensemble de test d'environ 100 exemples. Pour chaque méthode de classification, trois types de *fingerprints* sont testés. En ce qui concerne les *fingerprints* constitués de la totalité des porteuses GSM ou des porteuses sélectionnées sur le critère de puissance, seuls les hyperparamètres du modèle (*K*-PPV ou SVM) sont à déterminer par validation croisée. Dans le cas des *fingerprints* constitués des porteuses sélectionnées sur le critère de pertinence, en plus des hyperparamètres, le nombre de porteuses retenues est également déterminé par la procédure de validation croisée.

4.2.1. Résultats de la validation croisée

Nous présentons ici la performance de validation croisée dans le cas des *fingerprints* classés par pertinence à l'aide de la méthode OFR. Cette procédure, décrite dans la section 3.3.2, permet de déterminer le nombre de porteuses nécessaire et suffisant pour chaque classifieur *un contre un* afin de fournir une bonne performance de localisation. Lors de ce travail de thèse, une seconde méthode de classement des variables par ordre de pertinence a été étudiée : la méthode *Support Vector Machine - Recursive Feature Elimination* (SVM - RFE) [53], qui est utilisable uniquement pour les classifieurs SVM. Des classifieurs *un contre un* ont été mis en œuvre, chaque classifieur utilisant un ensemble de porteuses constitué de l'union des porteuses pertinentes de tous les classifieurs du système. Afin de minimiser le nombre de porteuses utilisé par les classifieurs et ainsi simplifier leur implémentation, une stratégie d'optimisation de chaque classifieur indépendamment des autres a été adoptée. De bonnes performances de localisation ont été obtenues, comme indiqué en détail dans l'annexe 2.

À titre d'exemple, nous considérons ici la mise en œuvre de l'algorithme LDA pour séparer chaque couple de pièces de la base *Lab*. La Figure 4.1 présente les performances de

validation croisée en fonction du nombre de porteuses pertinentes utilisées par chacun des dix classifieurs. Des barres d'erreur montrent les écarts-types des performances obtenues sur les différents ensembles de validation (voir section 3.4). Afin de tester chaque classifieur indépendamment des autres (dans un premier temps), un ensemble constitué de *fingerprints* appartenant aux deux classes concernées par chaque classifieur est utilisé. Les performances de chaque classifieur sur de tels ensembles sont représentées en fonction du nombre de porteuses retenues.

Cette procédure a pour objectif de déterminer le nombre de porteuses à conserver et à utiliser lors de la phase d'exploitation (la phase où le système est utilisé pour localiser des *fingerprints*). Le nombre retenu est celui qui permet d'atteindre le maximum de performance pendant la phase de validation croisée (ou le premier maximum s'il y a saturation de la performance comme c'est le cas pour le classifieur qui sépare la pièce 1 de la pièce 4 sur la Figure 4.1). On observe que le maximum de la performance de validation est atteint, pour chaque classifieur, avec moins de dix porteuses pertinentes.

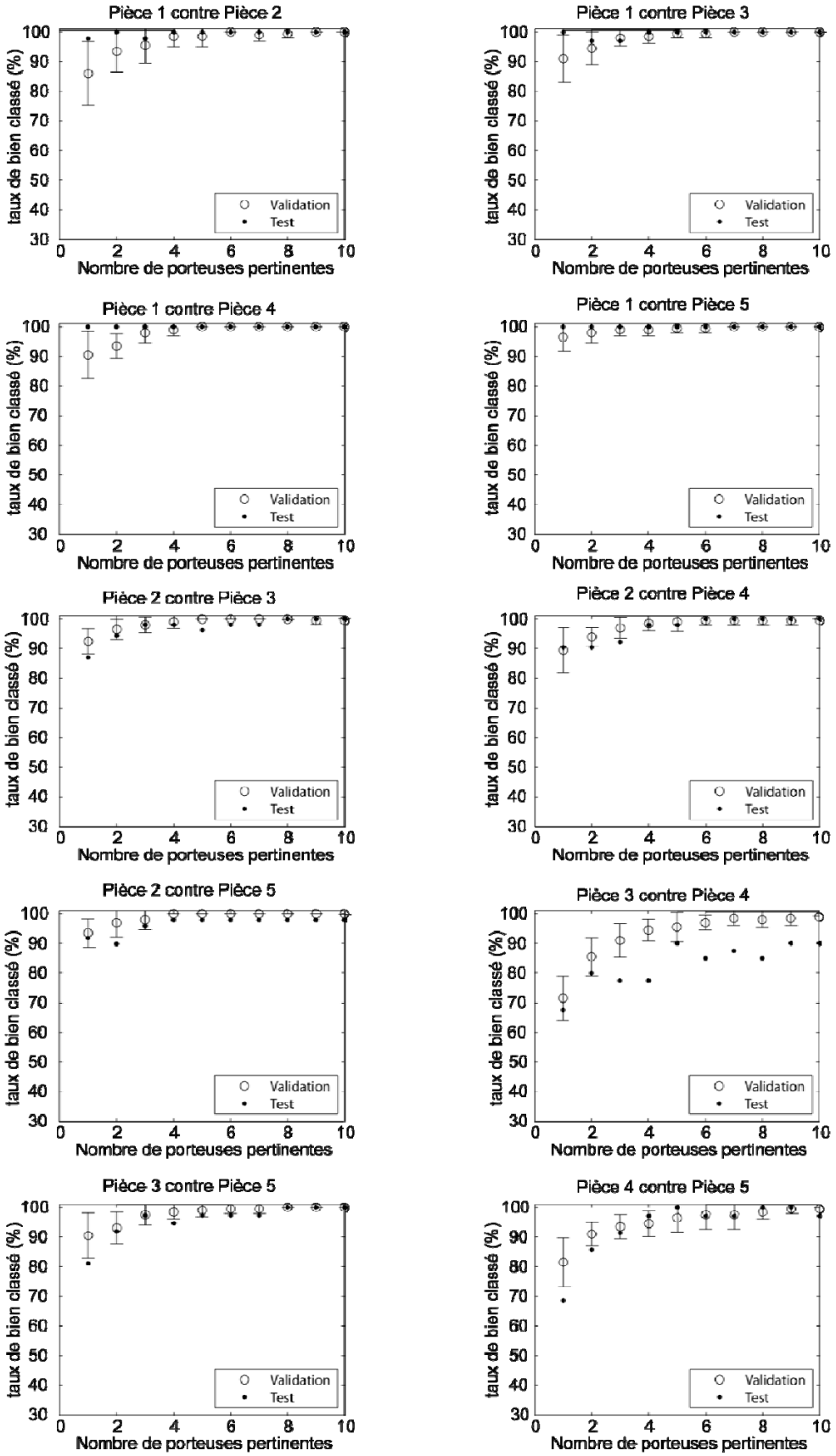

Figure 4.1. Performances de validation et de test des dix classifieurs LDA (Base *Lab*).

La validation croisée a été appliquée pour chacune des méthodes de classification étudiées (*K*-PPV, LDA et SVM). La Figure 4.2 montre le nombre de porteuses retenues pour chaque classifieur dans chaque cas (pour la base *Lab*).

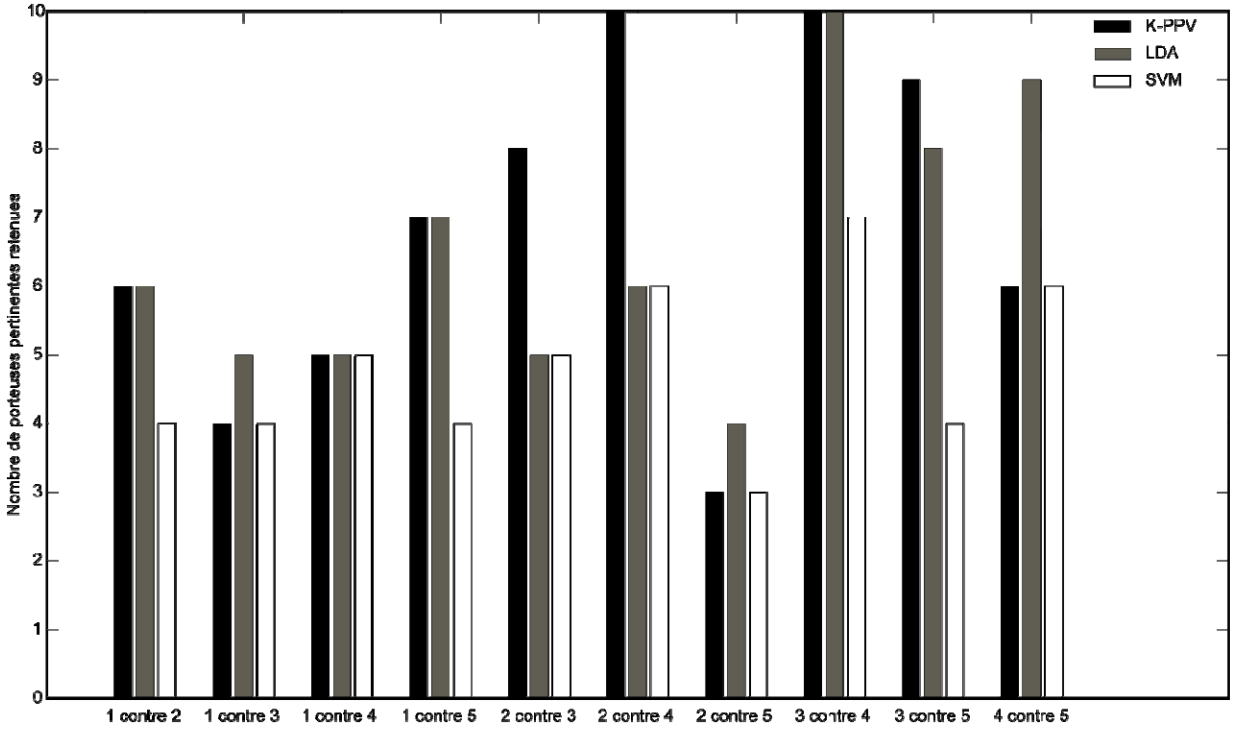

Figure 4.2. Nombre de porteuses pertinentes retenues pour chaque classifieur *un contre un* (base *Lab*).

Sachant qu'on dispose dans la base *Lab* d'environ 500 porteuses, on remarque qu'un nombre réduit de porteuses est suffisant pour atteindre une bonne performance de localisation. On observe également que ce nombre dépend, non seulement des classifieurs, mais aussi de la méthode de classification utilisée. Globalement, et comme on pouvait s'y attendre, les classifieurs SVM nécessitent moins de porteuses que les autres méthodes pour obtenir des résultats comparables (voir Tableau 4.1).

4.2.2. Étude des porteuses pertinentes retenues

Comme nous l'avons souligné dans les chapitres précédents, on pourrait s'attendre à ce que les voies balises suffisent à la localisation. On observe sur le Tableau 4.1 que la plupart des porteuses sélectionnées sont en effet des voies balises, mais que chaque classifieur *un contre un* utilise au moins une porteuse qui n'a jamais été détectée comme balise par le dispositif de mesure. Les mesures étant effectuées sur une période d'un mois, ceci suggère

que, parmi les plus pertinentes pour la séparation entre pièces, il existe des porteuses qui ne sont effectivement pas des balises.

Tableau 4.1. Nombre des voies balises parmi les porteuses sélectionnées.

		Classifieurs									
		1 contre 2	1 contre 3	1 contre 4	1 contre 5	2 contre 3	2 contre 4	2 contre 5	3 contre 4	3 contre 5	4 contre 5
SVM	Nombre de porteuses	4	4	5	4	5	6	3	7	4	6
	Nombre de balises	3	3	4	2	4	6	3	6	2	6
LDA	Nombre de porteuses	6	5	5	7	5	6	4	10	8	9
	Nombre de balises	5	4	4	4	4	6	4	9	6	8
K-PPV	Nombre de porteuses	6	4	5	7	8	10	3	10	9	6
	Nombre de balises	5	3	4	4	6	9	3	9	7	6

Il est également intéressant de savoir combien de porteuses sont nécessaires pour une localisation avec le système constitué des dix classifieurs. Ceci permet de savoir combien de porteuses le mobile à localiser doit mesurer lors de la phase d'utilisation du système. Le Tableau 4.2 montre, pour chaque méthode de classification, le nombre de porteuses utilisées par le système ainsi que le nombre de balises parmi ces porteuses sélectionnées.

Tableau 4.2. Nombre total des porteuses utilisées par le système de localisation.

	Nombre de porteuses	Nombre de balises
SVM	36	29
LDA	48	41
K-PPV	54	45

La Figure 4.3 représente l'écart-type des puissances en fonction des puissances moyennes mesurées dans la pièce 1, et met en évidence les porteuses retenues pour séparer la pièce 1 de la pièce 2. On constate que celles-ci ont un écart-type faible : ce ne sont donc pas les plus puissantes, mais celles qui sont reçues avec une puissance stable. Même si toutes ces

porteuses ne sont pas des balises, le comportement décrit est observé sur l'ensemble des porteuses sélectionnées.

L'utilisation des porteuses balises les plus fortes (ou des NMR) pour la localisation n'est donc pas toujours optimale : des porteuses non balises reçues avec des puissances faibles mais stables sont discriminantes. Les performances de localisation qui seront présentées dans la suite de ce chapitre confirmeront cette hypothèse.

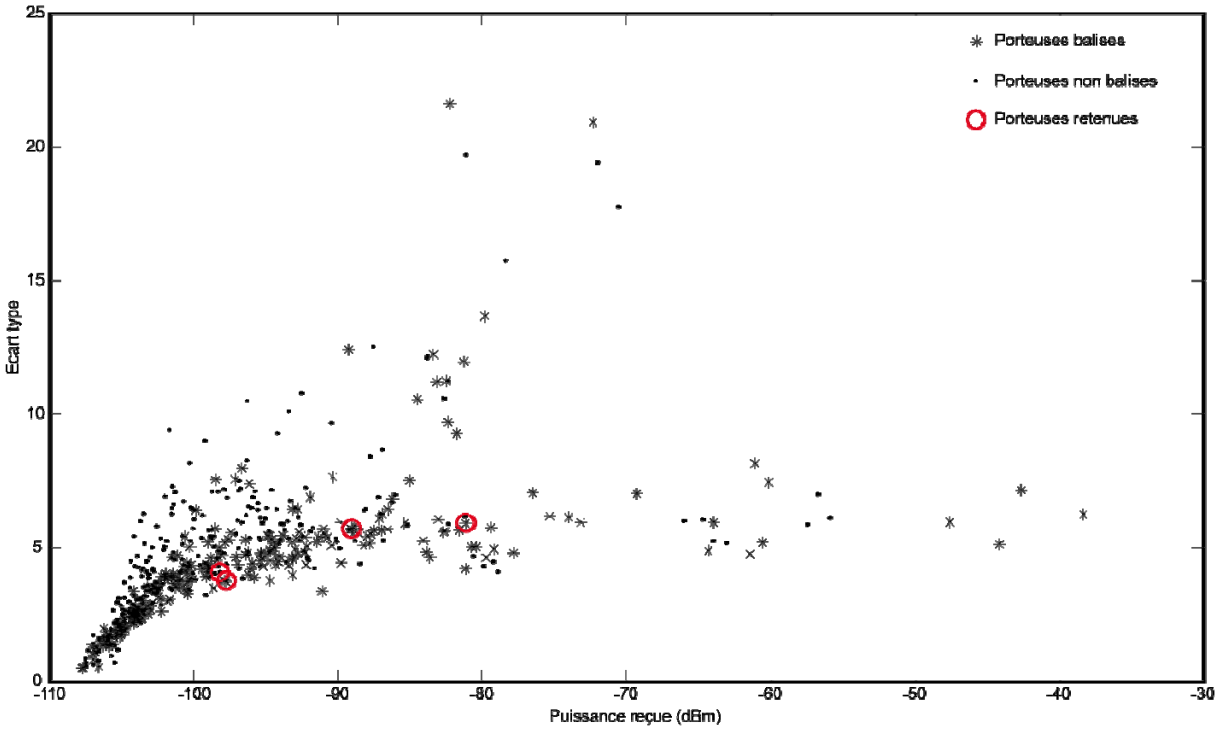

Figure 4.3. Comportement des porteuses retenues pour discriminer la pièce 1 de la pièce 2.

Le phénomène de corrélation des porteuses puissantes (qui sont souvent des balises) avec celles qui leur sont adjacentes, décrit dans le chapitre 2, est susceptible d'expliquer la sélection de porteuses non balises pour la localisation des pièces. La Figure 4.4 présente un balayage moyen des *fingerprints* mesurés dans la pièce 1. Les porteuses détectées, au moins une fois pendant la durée des mesures, comme balises sont indiquées par des astérisques. Les porteuses à l'intérieur des cercles sont celles qui sont retenues pour le classifieur qui sépare la pièce 1 de la pièce 2. Comme on peut le constater, trois des quatre porteuses désignées par des cercles sont des balises (voir aussi le Tableau 4.1). La porteuse restante est adjacente à un voie balise et par conséquent corrélée à celle-ci. L'absence du BSIC sur ce

canal adjacent peut être due à une difficulté de synchronisation empêchant le décodage de cette information.

Le processus de sélection adopté dans cette étude, ayant comme point de départ toutes les porteuses de la bande GSM, a permis d'aboutir à un ensemble de porteuses pertinentes pour la localisation qui ne sont pas nécessairement des voies balises. Pour tous les classifieurs étudiés, les porteuses non balises sélectionnées sont adjacentes à des voies balises non sélectionnées. Elles présentent donc un comportement analogue à celui d'une voie balise, comme indiqué dans le chapitre 2, ce qui peut expliquer leur présence parmi les porteuses pertinentes sélectionnées.

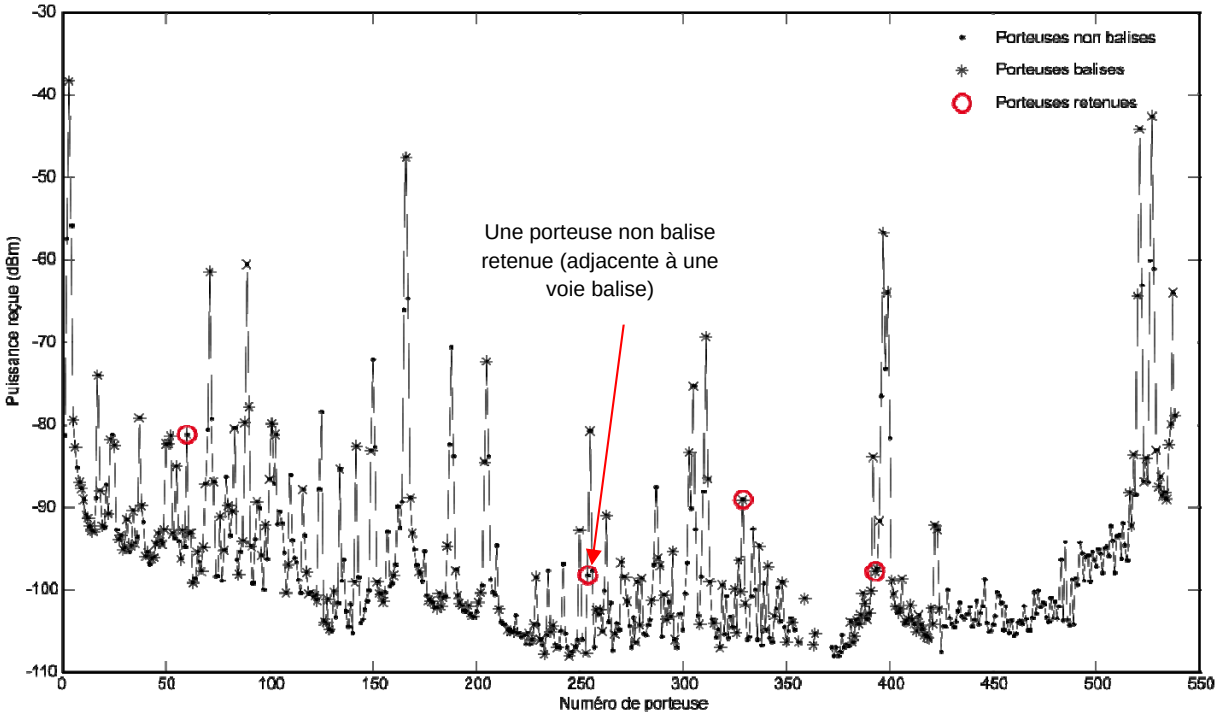

Figure 4.4. Position de porteuses retenues sur un balayage moyen.

Une fois le nombre de porteuses et les hyperparamètres des classifieurs déterminés par la procédure de validation croisée, un ensemble de *fingerprints* est testé afin d'évaluer la performance du système construit.

4.2.3. Résultats de test global

Dans cette étape du processus, les ensembles de test, tirés aléatoirement de chacun des deux bases de données étudiées, sont présentés à chacun des classifieurs *un contre un*. La

procédure de vote décrite dans la section 3.5 est ensuite appliquée pour attribuer à chaque *fingerprint* de test la pièce dans laquelle il a été mesuré. Les performances de localisation obtenues pour chaque type de *fingerprint* et pour chaque méthode de classification sont résumées dans le Tableau 4.3. Les chiffres représentent le pourcentage de cas où la pièce a été correctement identifiée. On trouvera plus de détails sur les résultats dans les annexes 1, 2 et 3.

Tableau 4.3. Performance de localisation sur des ensembles de test.

	Sélection par puissance			Sélection par pertinence			Pas de sélection		
	K-PPV	LDA	SVM	*K*-PPV	LDA	SVM	*K*-PPV	LDA	SVM
Base *Home*	54.1%	(1)	68.9%	90.2%	91.8%	93.4%	93.4%	(2)	96.7%
Base *Lab*	52.5%	(1)	59%	95%	95%	95%	86.1%	(2)	99%

(1) La matrice de covariance n'est pas inversible en raison de la constitution des *fingerprints* (voir section 3.1.1), donc l'équation de la surface de séparation ne peut pas être déterminée.

(2) De nombreuses porteuses étant corrélées, la matrice de covariance n'est pas inversible.

Les résultats présentés dans le Tableau 4.3 montrent clairement que le choix de la sélection par critère de puissance ne permet pas une bonne performance de localisation. En effet, quelle que soit la méthode de classification mise en œuvre, la performance avec ce type de *fingerprint* ne dépasse pas 69%. Nous avons étudié en détail le classement des porteuses par ordre de puissance et la sélection des plus fortes pour la localisation [54-56]. Nous avons notamment montré (annexe 2) que les porteuses les plus puissantes contiennent une faible proportion de porteuses pertinentes : par exemple, les 100 porteuses les plus puissantes ne comprennent que 40 % des porteuses les plus pertinentes.

La prise en compte de la totalité de la bande GSM comme *fingerprint* aboutit à d'excellentes performances de localisation. Sur la base *Lab*, 99% des exemples de test ont été localisés dans la bonne pièce en utilisant des classifieurs SVM. Ce résultat confirme la capacité des SVM à gérer les problèmes de grande dimension [45]. En revanche, l'application de l'algorithme des *K*-PPV à la localisation par *fingerprints* de grande dimension n'est pas satisfaisante. Pour la base *Lab*, sa performance est même inférieure à celle qui est obtenue avec un nombre réduit de porteuses pertinentes.

En sélectionnant un petit nombre de porteuses pertinentes (3 à 10), on peut obtenir des performances comparables (quoiqu'un peu inférieures) à celles que l'on obtient en conservant toutes les porteuses. En effet, le fait de considérer, pour chaque classifieur *un contre un*, les porteuses qui sont pertinentes pour sa propre tâche permet d'optimiser le système de localisation : il n'utilise que des classifieurs parcimonieux en nombre de variables, au prix d'une petite perte de performance. Ceci pourrait apporter un gain de temps pour l'utilisation finale, où l'appareil à localiser serait chargé de faire des mesures de puissance uniquement sur les porteuses pertinentes.

Afin de mettre en évidence les pièces confondues par le classifieur, une matrice de confusion est présentée sur le Tableau 4.4, où est considéré, à titre d'exemple, le cas de l'analyse discriminante linéaire appliquée aux *fingerprints* des porteuses pertinentes.

Tableau 4.4. Matrice de confusion de l'analyse discriminante linéaire utilisant les porteuses pertinentes sur la base *Lab*.

	Pièces prédites				
	Pièce 1	Pièce 2	Pièce 3	Pièce 4	Pièce 5
Pièce 1	100	0	0	0	0
Pièce 2	0	97	0	0	3
Pièce 3	0	0	95.2	4,8	0
Pièce 4	0	0	5.3	84.2	10.5
Pièce 5	0	0	0	0	100

On observe que les exemples les plus fréquemment confondus sont ceux des pièces 3-4 et 4-5 : les *fingerprints* de ces couples de pièces sont difficiles à séparer par la méthode choisie. Ceci est confirmé par les résultats de validation croisée de la Figure 4.1, où les classifieurs 3 contre 4 et 4 contre 5 présentent des performances de validation inférieures à celles qui sont obtenues sur les autres classifieurs. La Figure 4.2, sur laquelle on peut voir le nombre de porteuses utilisées par chaque classifieur, permet également de constater que les classifieurs 3 contre 4 et 4 contre 5 sont plus exigeants que les autres en termes de nombres de porteuses.

En ce qui concerne les méthodes de classification, LDA et SVM, qui nécessitent un apprentissage, s'obtiennent, avec tous les types de *fingerprints*, de meilleures performances que la méthode des *K*-PPV traditionnellement utilisée pour la localisation par *fingerprints*. Les classifieurs SVM, bien adaptés au traitement des problèmes de classification dans des espaces de grande dimension, permettent effectivement d'atteindre la performance maximale de 99% avec des *fingerprints* de grande taille (488 pour la base *Home* et 534 pour la base *Lab*).

Pour des applications exigeantes en précision, les *fingerprints* de grande dimension peuvent être envisagés avec une méthode de classification qui permet de les gérer telle que les SVM. Dans le cas des applications de localisation d'urgence, où le temps est un élément important, des *fingerprints* mettant en œuvre la sélection par pertinence pourraient se révéler mieux adaptés en permettant un compromis entre la précision de localisation et le temps de réponse.

Dans ce mémoire, nous avons décrit en détail la sélection des porteuses (sur critère de pertinence ou de puissance) pour réduire la taille des *fingerprints*. Nous avons mis en œuvre l'Analyse en Composantes Principales (ACP), qui permet d'exprimer les données comme combinaisons linéaires de composantes, dites *composantes principales*, en nombre inférieur au nombre de porteuses disponibles. Le nombre de composantes principales nécessaire a été déterminé par validation croisée. Nous avons montré (annexe 3) que la mise en œuvre de l'ACP comme prétraitement des données s'accompagne, dans notre cas, d'une dégradation des performances du système de localisation.

4.3. Performance de localisation et variabilité temporelle

À cause de la variabilité temporelle des mesures de puissance observée sur la Figure 2.7, nous avons cherché à estimer expérimentalement l'influence de cette variabilité sur les performances de localisation du système proposé. Nous avons montré que les classifieurs *un contre un* sont inutilisables sur des données de test acquises six mois après les données qui ont servi à l'apprentissage. Dans ce cas, une mise à jour des classifieurs est nécessaire. En plus de la variabilité temporelle, due à l'influence de l'environnement de propagation, la modification cellulaire effectuée par l'opérateur peut également être la cause de ce phénomène.

Afin de vérifier la validité du système sur la période d'un mois, un ensemble d'apprentissage/validation a été formé avec les mesures effectuées sur les trois premières semaines. Les mesures de la dernière semaine ont été utilisées comme ensemble de test. Cette expérience a été menée sur la base *Lab*, qui a un plus grand nombre d'exemples de test que la base *Home*. Les performances étant faibles pour les *fingerprints* sélectionnés sur critère de puissance, seuls les *fingerprints* composés de toutes les porteuses GSM ou des porteuses pertinentes ont été considérées. La procédure de construction des classifieurs est la même que dans les expériences précédentes. Les performances obtenues sont résumées dans le Tableau 4.5.

Tableau 4.5. Performances sur l'ensemble de test mesuré sur une semaine.

	Sélection par pertinence			Pas de sélection		
	K-PPV	LDA	SVM	*K*-PPV	LDA	SVM
Base *Lab*	93,10	95%	94%	92,1%	-	97%

Les bonnes performances de localisation sur cette base de test confirment la possibilité d'utiliser les classifieurs sur une échelle de temps d'un mois sans aucune modification. Pour l'application finale, une phase d'adaptation des modèles devra être envisagée dans un délai compris entre un et six mois. Pour des raisons de temps, nous n'avons pas pu effectuer d'expériences sur des durées intermédiaires entre un et six mois.

4.4. Performances de localisation et variabilité entre dispositifs de mesure

L'étude menée sur les signaux GSM mesurés a mis en évidence des différences de niveaux des puissances reçues par deux dispositifs de mesure différents. Ce constat a également été confirmé par des tests dans une chambre anéchoïde (voir section 2.5). Dans cette section, nous cherchons à estimer l'influence de la variabilité des dispositifs sur les performances de classification. Les résultats que nous présentons ont été obtenus par application d'un classifieur SVM linéaire sur les données de la base *Minipegs*. Pour cette

expérience, compte tenu des résultats obtenus dans les sections précédentes, nous utilisons uniquement les *fingerprints* incluant toute les porteuses GSM.

L'objectif de cette étude est de déterminer si un classifieur construit avec une base de données de *fingerprints* mesurés par un de nos appareils peut être utilisé pour localiser un appareil mobile différent. Pour cela, un classifieur, séparant les deux pièces considérées dans la base *Minipegs*, a été construit à partir des mesures de l'appareil numéro 1 (voir Figure 3.4). Les fingerprints collectés dans les mêmes pièces par les autres dispositifs (le même jour pour les dispositifs 2, 3, 4, un jour plus tard pour les autres) ont été présentés à ce classifieur. Les performances de localisation obtenues sur les données de chacun des appareils sont résumées dans le Tableau 4.6.

Tableau 4.6. Performances du classifieur construit à partir des données du dispositif numéro 1.

Dispositif	2	3	4	5	6	7	8
Performance	80.8%	82.9%	75.6%	64.3%	51.1%	61.2%	54.8%

Les performances présentées dans ce tableau sont largement inférieures à celles qui sont obtenues quand le même dispositif est utilisé pour les mesures d'apprentissage et de test. Cependant, le classifieur testé reste plus performant sur des données prises, par des dispositifs différents, au même moment que les données d'apprentissage (dispositifs 2, 3 et 4).

Afin de contourner ce problème de variabilité du matériel de mesure, nous avons utilisé pour l'apprentissage une base de données construite avec plusieurs dispositifs. Ceci permet au classifieur d'apprendre la variabilité entre les différents dispositifs. Le Tableau 4.7 montre les résultats obtenus en présentant les données collectées par les dispositifs 2 à 7 à un classifieur conçu à partir des données collectées par les dispositifs 1 et 8.

Tableau 4.7. Performance du classifieur construit à partir des données des dispositifs 1 et 8 sur le reste des *fingerprints*.

Dispositif	2	3	4	5	6	7
Performance	99.6%	99.6%	99.6%	98.2%	99.6%	99.2%

Les taux de *fingerprints* bien localisés atteignent les niveaux obtenus quand un seul appareil de mesure est considéré. Cette approche peut être adoptée pour des applications où plusieurs utilisateurs sont à localiser. Chacun disposant d'un équipement différent, un apprentissage sur l'un ou l'autre n'est alors pas envisageable. Dans ce cas, le système de localisation doit être construit à partir de mesures issues de différents dispositifs afin de pouvoir fournir une précision acceptable.

4.5. Localisation par des classifieurs conçus par apprentissage semi-supervisé

Comme indiqué dans la section 3.4.5, étiqueter les *fingerprints* au fur et à mesure des balayages est une tâche coûteuse en temps. Afin de tirer profit de toutes les mesures, même si celles-ci ne sont pas étiquetées, il existe des méthodes d'apprentissage semi-supervisé. Des performances de localisation d'environ 80% ont été atteintes avec les puissances reçues de points d'accès WiFi [57] : il s'agissait de localiser un mobile sur un trajet à partir de *fingerprints* constitués de mesures de puissances reçues des points d'accès WiFi. Les données d'apprentissage dans ce cas étaient partiellement étiquetées. La méthode d'apprentissage semi supervisé dite de « propagation des étiquettes » a été adoptée pour ce problème. Cette dernière tire profit de la connaissance de l'étiquette de position à l'instant t-1 pour attribuer au *fingerprint* une étiquette à l'instant t. Comme nous l'avons indiqué dans la section 1.1, nous ne disposons pas de données exogènes de ce type. Nous avons donc mis en œuvre une autre méthode d'apprentissage semi-supervisé : les Machines à Vecteurs Supports Transductives (TSVM), fondées sur les SVM dont nous avons vu, dans les sections précédentes, qu'elles nous permettent d'obtenir d'excellents résultats de localisation.

À partir des bases de données *Home* et *Lab*, des situations de bases partiellement étiquetées ont été simulées. Pour cela, les ensembles de test construits dans la section 4.2 ont été utilisés. Sur les ensembles d'apprentissage, 70% des étiquettes ont été retirées. Dans ce cas, deux éléments différents par rapport à l'apprentissage supervisé :

- en l'absence des étiquettes pour une partie des exemples, un système de classifieurs *un contre un* ne peut être envisagé.

- toujours en raison de l'absence des étiquettes pour une partie des exemples, la validation croisée décrite dans la section 3.4 n'est pas possible.

Nous avons donc, d'une part, réalisé des classifieurs *un contre tous*, et, d'autre part, modifié la procédure de validation croisée. Afin de sélectionner les hyperparamètres des classifieurs possédant les meilleures capacités de généralisation, la partition illustrée par la Figure 4.5 a été utilisée. À chaque itération de la validation, deux sous-ensembles sont extraits aléatoirement des données étiquetées de l'ensemble d'apprentissage/validation. Le premier est consacré au test. Le second est mélangé aux exemples non étiquetés afin de former l'ensemble d'apprentissage.

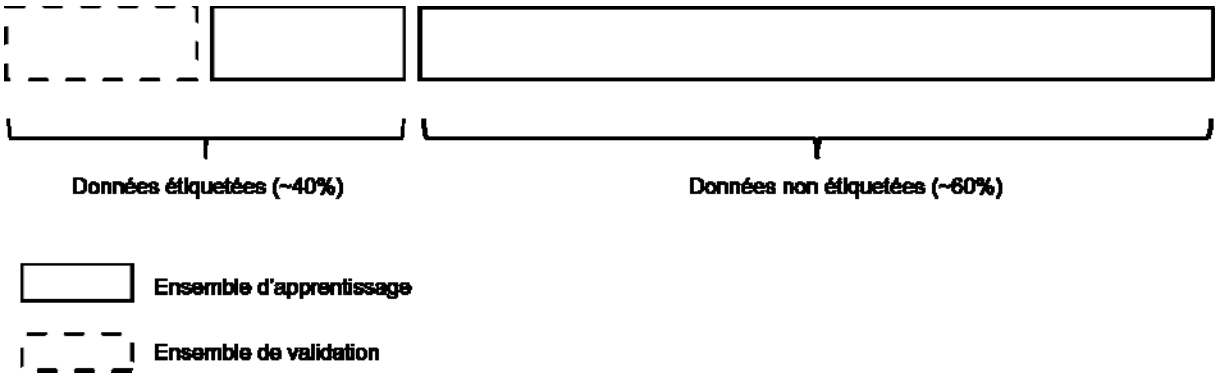

Figure 4.5. Partition de données pour chaque itération de la procédure de validation des TSVM.

Afin d'évaluer les performances de cette méthode, seuls les *fingerprints* contenant toutes les porteuses GSM ont été testés. L'utilisation de classifieurs non linéaires n'apportant pas d'amélioration à la performance de localisation, seuls les résultats obtenus avec des classifieurs linéaires sont présentés dans ce mémoire. Les résultats obtenus sur les bases *Home* et *Lab* sont résumés dans le Tableau 4.8.

Tableau 4.8. Performance de test des classifieurs TSVM linéaires sur des *fingerprints* contenant la totalité des porteuses GSM.

	TSVM
Base *Home*	98.4%
Base *Lab*	92.1%

On peut constater que le système qui met en œuvre les classifieurs linéaires TSVM utilisant toutes les porteuses GSM est aussi performant que son homologue supervisé. On remarque une baisse de la performance obtenue sur la base *Lab*, mais elle reste acceptable si l'on considère que les classifieurs ont été construits sur une base d'apprentissage dont 70% des exemples sont non étiquetés.

Cette expérience a permis de mettre en évidence la robustesse du système de localisation en présence d'une base de données partiellement étiquetées. Ainsi des ensembles d'apprentissage plus importants pourront être collectés sans qu'il soit nécessaire de les étiqueter entièrement.

4.6. Conclusion

Dans ce chapitre, les résultats de localisation obtenus avec le système proposé ont été présentés et discutés. Les différentes méthodes de sélection des *fingerprints* et de classification proposées pour chaque étape du processus de localisation ont été testées et leurs performances comparées. De plus, des solutions permettant de faire face à la variabilité temporelle des signaux et à la variabilité entre dispositifs de mesure ont été proposées. Ainsi, un système de localisation robuste aux phénomènes observés sur les signaux utilisés (voir chapitre 2) peut être conçu. Enfin, le problème d'étiquetage des balayages lors de la construction des bases de données a également été abordé. Une méthode d'apprentissage semi-supervisé a permis de tirer profit de tous les balayages effectués, même de ceux qui n'ont pas été étiquetés.

Conclusion générale et perspectives

L'objectif de ce travail était la conception d'un système de localisation en absence d'un signal GPS, permettant l'intégration des services nécessitant une localisation à l'intérieur de bâtiments ou dans des situations où la localisation par GPS est inopérante. Les méthodes existantes fournissent des performances acceptables, mais nécessitent, pour la plupart, l'installation d'infrastructures spéciales ou des modifications importantes des infrastructures existantes. Le coût de ces modifications a empêché le déploiement de ce type de solutions. Il convenait donc d'étudier la faisabilité d'un système de localisation sans GPS, ne nécessitant pas de modification d'infrastructure et pas ou peu de modification des récepteurs. L'ubiquité du système de téléphonie mobile GSM faisait de celui-ci un excellent candidat pour une telle application.

Nous avons démontré, dans ce travail, la faisabilité d'un système, reposant sur des techniques d'apprentissage statistique, qui modélise la relation entre la position d'un récepteur mobile et les puissances des porteuses GSM reçues par celui-ci.

Cette approche nécessite une première phase d'acquisition de données. Celle-ci a été réalisée par deux dispositifs de mesures, dans deux bâtiments différents et à des périodes suffisamment séparées dans le temps pour que l'on puisse tester les performances du système dans des conditions significativement différentes. Le comportement des signaux enregistrés a été étudié, ce qui nous a permis de mettre en évidence la variabilité temporelle des puissances reçues à la même position.

La seconde phase de conception du système de localisation consiste en l'extraction de l'information pertinente des mesures effectuées. Dans ce but, différentes techniques de sélection de variables (porteuses) ont été mises en œuvre. Dans un premier temps, des méthodes de classement des porteuses ont été utilisées. Deux critères ont été considérés pour ce classement : le critère de pertinence, c'est-à-dire la capacité des porteuses à discriminer différentes positions, et le critère de puissance reçues dans lequel on considère que les porteuses reçues avec de fortes puissances sont les plus informatives pour la localisation. Une procédure de validation croisée est ensuite nécessaire pour définir le nombre de porteuses à retenir pour la construction des modèles de localisation.

La troisième phase du processus consiste en la construction du modèle de localisation. Cette phase sert à définir les paramètres des classifieurs qui discriminent les pièces dans lesquelles des mesures du spectre GSM ont été effectuées. Une stratégie *un contre un* a été adoptée pour résoudre le problème de localisation à plusieurs pièces, abordé dans cette étude comme un problème de classification multi-classe. Dans cette phase, plusieurs méthodes d'apprentissage statistique ont été implémentées. Les hyperparamètres des différents classifieurs, ainsi que le nombre de porteuses considérées par chacun d'eux, ont été déterminés par validation croisée.

Une fois les paramètres des classifieurs estimés, le système peut être utilisé pour identifier les positions dans lesquelles des mesures de test ont été effectuées. Pour cela, les *fingerprints* de test sont présentés aux classifieurs construit. L'identification de la pièce où a été mesuré chaque *fingerprint* est effectuée par un vote à partir des réponses des différents classifieurs.

Les résultats obtenus montrent l'apport de méthodes d'apprentissage statistique au problème de la localisation en intérieur par *fingerprint* GSM. Les performances obtenues sont extrêmement encourageantes et ouvrent des perspectives nouvelles pour la localisation des personnes et des biens.

Pour passer de notre étude de faisabilité à un appareil et des services opérationnels, plusieurs étapes restent à franchir. Tout d'abord, il convient de valider notre concept de manière extensive, dans différents types de bâtiments et de locaux. Ceci nécessitera un investissement matériel, pour la conception et la réalisation de dispositifs de mesure plus évolués que ceux que nous avons utilisés ici, et dont l'ergonomie se rapproche de celle d'un téléphone. Cela nécessitera aussi une évolution de nos algorithmes ; la modification dans le temps de l'environnement électromagnétique, dont nous avons vu qu'elle exige une mise à jour des classifieurs avec une périodicité de six mois au maximum, nécessitera probablement la mise au point d'algorithmes adaptatifs [57].

D'autre part, nous avons supposé, dans toute notre étude, qu'aucune information exogène (cartographie des locaux, vitesse et direction de déplacement du mobile, …) n'était disponible. Dans une application réelle, il est très probable que de telles informations pourront être mises à profit. Il est également possible que d'autres capteurs (magnétomètre, RFID, …) fournissent des données qu'il faudra fusionner avec le spectre de puissance GSM.

Enfin, pour des applications dans le domaine de la santé, notamment le suivi des personnes atteintes de troubles cognitifs (maladie d'Alzheimer), le problème de l'ergonomie et de l'acceptabilité du dispositif par les patients concernés devra être abordé, dans le cadre de collaborations avec des professionnels du domaine de la santé.

Publications

I. Ahriz, B. Denby, G. Dreyfus, R. Dubois, P. Roussel, *The ARPEGEO Project : A New Look at Cellular RSSI Fingerprints for Localization*, Workshop on Advances in Positioning and Location-Enabled Communications, 26-29 Septembre 2010, Istanbul, Turquie.

I. Ahriz, Y. Oussar, B. Denby, G. Dreyfus, *Full-Band GSM Fingerprints for Indoor Localization using a Machine Learning Approach*, International Journal of Navigation and Observation, Hindawi, Mai 2010.

I. Ahriz, Y. Oussar, B. Denby, G. Dreyfus, *Carrier Relevance Study for Indoor Localization Using GSM*, 7th Workshop on Positioning, Navigation and Communication 2010, 11-12 Mars 2010, Dresden, Allemagne.

B. Denby, Y. Oussar, I. Ahriz, G. Dreyfus, *High-Performance Indoor Localization with Full-Band GSM Fingerprints*, actes de IEEE International Conference on Communications, 14-18 Juin 2009, Dresden, Allemagne.

B. Denby, G. Dreyfus, Y. Oussar, I. Ahriz, *Système de localisation de personnes à base de réseaux cellulaires en absence d'un signal GPS adéquat*, demande de dépôt de brevet n° 09 02 863 déposée le 12 Juin 2009.

B. Denby, Y. Oussar, I. Ahriz, *Geolocalisation in Cellular Telephone Networks*, actes de NATO 2007 Advanced Study Institute on Mining Massive Data Sets for Security, IOS Press, Amsterdam, The Netherlands, F. Fogelman-Soulié, D. Perrotta, J. Piskorski & R. Steinberger, Eds., IOS Press, Amsterdam, Netherlands.

Références

[1] C. Drane, R and C. Rizos, *Positioning Systems for Intelligent Transportation Systems*. Boston: Artech House, 1997.

[2] [Online], "http://heberge.lcpc.fr/lavia/," 2006.

[3] A. Küpper, *Location-based Services: Fundamentals and Operation*. Germany: John Wiley & Sons, 2005.

[4] G. Glanzer and U. Walder, "Self-Contained Indoor Pedestrian Navigation by Means of Human Motion Analysis and Magnetic Field Mapping," in *Proceedings of the 7th Workshop on Positioning, Navigation and Communication*, Dresden, Germany, 2010, pp. 177-181.

[5] J. Haverinen and A. Kemppainen, "Global indoor self-localization based on the ambient magnetic field," *Robotics and Autonomous Systems,* vol. 57, pp. 1028-1035, Oct 31 2009.

[6] P. Robertson, M. Angermann, and B. Krach, "Simultaneous Localization and Mapping for Pedestrians using only Foot-Mounted Inertial Sensors," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, Orlando, Florida, USA, 2009, pp. 93-96.

[7] R. Want, A. Hopper, V. Falcao, and J. Gibbons, "The Active Badge Location System," *Acm Transactions on Information Systems,* vol. 10, pp. 91-102, Jan 1992.

[8] J. Ansari, J. Riihijarvi, and P. Mahonen, "Combining particle filtering with cricket system for indoor localization and tracking services," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Athens, Greece, 2007, pp. 2270-2274.

[9] N. Ravi, P. Shankar, A. Frankel, A. Elgammal, and L. Iftode, "Indoor localization using camera phones," in *Proceedings of the 7th IEEE Workshop on Mobile Computing Systems & Applications, Proceedings*, 2006, pp. 19-19.

[10] M. Heidari and K. Pahlavan, "A Markov model for dynamic behavior of ToA-based ranging in indoor localization," *Eurasip Journal on Advances in Signal Processing,* pp. -, 2008.

[11] P. Kemppi and S. Nousiainen, "Database Correlation Method for Multi-System Positioning," in *Proceedings of the IEEE 63rd Vehicular Technology Conference*, Melbourne, Australia, 2006, pp. 866-870.

[12] V. Ashkenazi, C. H. J. Chao, W. Chen, C. J. Hill, and T. Moore, "A new high precision wide area DGPS system," *Journal of Navigation,* vol. 50, pp. 109-119, Jan 1997.

[13] J. Z. Li and M. Q. Wu, "A Positioning Algorithm of AGPS," in *Proceedings of the International Conference on Signal Processing Systems*, Singapore, Singapore 2009, pp. 385-388.

[14] F. van Diggelen, "Indoor GPS theory & implementation," *IEEE Position Location and Navigation Symposium,* pp. 240-247, 2002.

[15] S. S. Chawathe, "Beacon Placement for Indoor Localization using Bluetooth," in *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China, 2008, pp. 980-985.

[16] G. Goncalo and S. Helena, "Indoor Location System using ZigBee Technology," in *Proceedings of the 3rd International Conference on Sensor Technologies and Applications* Athens, Greece, 2009, pp. 152-157.

[17] M. Sugano, T. Kawazoe, Y. Ohta, and M. Murata, "Indoor localization system using RSSI measurement of wireless sensor network based on ZigBee standard," in *Proceedings of the 6th International Multi-Conference on Wireless and Optical Communications*, Alberta, Canada, 2006, pp. 503-508.

[18] E. Trevisani and A. Vitaletti, "Cell-ID location technique, limits and benefits: an experimental study," in *Proceedings of the 6th IEEE Workshop on Mobile Computing Systems and Applications*, Lake District National Park, UK, 2004, pp. 51-60.

[19] M. Kuhn, C. C. Zhang, B. Merkl, D. P. Yang, Y. Z. Wang, M. Mahfouz, and A. Fathy, "High Accuracy UWB Localization in Dense Indoor Environments," in *Proceedings of the IEEE International Conference on Ultra-Wideband*, Hannover, Germany 2008, pp. 129-132.

[20] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, "Field Strength and Its Variability in Vhf and Uhf Land-Mobile Radio Service," *Review of the Electrical Communications Laboratories,* vol. 16, pp. 825-&, 1968.

[21] E. Damosso and G. deBrito, "COST 231 achievements as a support to the development of UMTS: A look into the future," *IEEE Communications Magazine,* vol. 34, pp. 90-96, Feb 1996.

[22] C. Yongguang and H. Kobayashi, "Signal Strength Based Indoor Geolocation," in *Proceedings of the IEEE International Conference on Communication*, 2002, pp. 436 - 439.

[23] X. Lagrange, P. Godlewski, and S. Tabbane, *Réseaux GSM*. Paris: Hermes Science, 2000.

[24] A. A. Ali and A. S. Omar, "Time of Arrival Estimation for WLAN Indoor Positioning Systems using Matrix Pencil Super Resolution Algorithm," in *Proceedings of the 2nd Workshop on Positioning, Navigation and Communication*, Hannover, Germany, 2005, pp. 11 - 20.

[25] M. Pettersen, R. Eckhoff, P. H. Lehne, T. A. Worren, and E. Melby, "An experimental evaluation of network-based methods for mobile station positioning," in *Proceedings of the 13th IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications*, Lisboa, Portugal, 2002, pp. 2287-2291.

[26] H. Laitinen, J. Lahteenmaki, and T. Nordstrom, "Database Correlation Method for GSM location," in *Proceeding of the 53rd IEEE Vehicular Technology Conference*, Rhodes, Greece, 2001, pp. 2504-2508.

[27] P. Bahl and V. N. Padmanabhan, "RADAR: An In-Building RF-based User Location and Tracking System," in *Proceedings of the 9th Annual Joint Conference of the IEEE Computer and Communications Societies*, Tel-Aviv, Israel, 2000, pp. 775-784.

[28] Y. Y. Jin, W. S. Soh, and W. C. Wong, "Indoor Localization with Channel Impulse Response Based Fingerprint and Nonparametric Regression," *IEEE Transactions on Wireless Communications,* vol. 9, pp. 1120-1127, Mar 2010.

[29] A. Varshavsky, A. LaMarca, J. Hightower, and E. de Lara, "The SkyLoc floor localization system," in *proceedings of the 5th Annual IEEE International Conference on Pervasive Computing and Communications*, New York, USA, 2007, pp. 125-134.

[30] B. D. S. Lakmali and D. Dias, "Database Correlation for GSM Location in Outdoor & Indoor Environments," in *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, Colombo, Sri Lanka 2008, pp. 423-428.

[31] V. Otsason, A. Varshavsky, A. LaMarca, and E. de Lara, "Accurate GSM indoor localization," in *Proceedings of the Ubiquitous Computing Conference*, Tokyo, Japan, 2005, pp. 141-158.

[32] D. Zimmerman, J. Baumann, M. Layh, F. Landstorfer, R. Hoppe, and G. Wölfle, "Database correlation for positioning of mobile terminals in cellular networks using wave propagation models," in *Proceedings of the IEEE 60th Vehicular Technology Conference* Los Angeles, CA, USA, 2004, pp. 4682-4686.

[33] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless LANs," *The International Journal of Computer and Telecommunications Networking,* vol. 47, pp. 825-845, Apr 22 2006.

[34] F. M. Landstorfer, "Wave Propagation Models for the Planning of Mobile Communication Networks," in *Proceedings of the 29th European Microwave Conference* Munich, Germany, 1999, pp. 1-6.

[35] T. Roos, P. Myllymäki, H. Tirri, P. Misikangas, and J. Sievänen, "A Probabilistic Approach to WLAN User Location Estimation," *International Journal of Wireless Information Networks,* vol. 9, pp. 155 -164
July 2002.

[36] M. Youssef and A. Agrawala, "The Horus WLAN location determination system," in *Proceedings of the Third International Conference on Mobile Systems, Applications, and Services* Seattle, Washington, USA, 2005, pp. 205-218

[37] [Online], "http://www.ascom.com/en/index/group/company/divisions/network-testing-home.htm," 2010.

[38] [Online], "http://www.ascom.com/en/index/products-solutions/our-solutions/product/tems-pocket-3," 2010.

[39] [Online], "http://www.telit.com/en/products/gsmgprs.php," 2010.

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

[41] S. Tabbane, *Ingénierie des réseaux cellulaires*. Paris: Hermes Science, 2002.

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

[43] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society Series B-Statistical Methodology,* vol. 36, pp. 111-147, 1974.

[44] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *Ieee Transactions on Information Theory,* vol. It13, pp. 21-+, 1967.

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

[46] [Online], "http://www.kyb.tuebingen.mpg.de/bs/people/spider/," 2010.

[47] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Massachusetts: MIT Press, 2006.

[48] T. Joachim, "Transductive Inference for Text Classification using Support vecter Machines," in *Proceedings of the International conference on Machine Learning* Bled, Slovenia, 1999, pp. 200-209. .

[49] J. Jia and L. Cai, "A TSVM-Based Minutiae Matching Approach for Fingerprint Verification," in *Proceedings of the International Workshop on Biometric Recognition Systems* Beijing, China, 2005, pp. 85-94.

[50] [Online], "http://svmlight.joachims.org/," 1999.

[51] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," *Advances in Neural Information Processing Systems,* vol. 12, pp. 547-553, 2000.

[52] Y. C. Ho and R. L. Kashyap, "An Algorithm for Linear Inequalities and Its Applications," *Ieee Transactions on Electronic Computers,* vol. Ec14, pp. 683-&, 1965.

[53] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[54] B. Denby, Y. Oussar, I. Ahriz, and G. Dreyfus, "High-Performance Indoor Localization with Full-Band GSM Fingerprints," in *Proceedings of the IEEE International Conference on Communication Workshops*, Dresden, Germany 2009, pp. 654-658.

[55] I. Ahriz, Y. Oussar, B. Denby, and G. Dreyfus, "Full-Band GSM Fingerprints for Indoor Localization Using a Machine Learning Approach," *International Journal of Navigation and Observation,* vol. 2010, p. 7 pages, 2010.

[56] I. Ahriz, Y. Oussar, B. Denby, and G. Dreyfus, "Carrier relevance study for indoor localization using GSM," in *Proceedings of the 7th Workshop on Positioning, Navigation and Communication*, Dresden, Germany, 2010.

[57] Q. Yang, S. J. Pan, and V. W. Zheng, "Estimating location using Wi-Fi," *IEEE Intelligent Systems,* vol. 23, pp. 8-13, Jan-Feb 2008.

Annexe 1 : The ARPEGEO Project: A New Look at Cellular RSSI Fingerprints for Localization

I. Ahriz, Bruce Denby, Gérard Dreyfus, Rémi Dubois, Pierre Roussel, *The ARPEGEO Project : A New Look at Cellular RSSI Fingerprints for Localization*, Workshop on Advances in Positioning and Location-Enabled Communications, 26-29 Septembre 2010, Istanbul, Turquie.

Annexe 2 : Carrier Relevance Study for Indoor Localization Using GSM

I. Ahriz, Y. Oussar, B. Denby, G. Dreyfus, *Carrier Relevance Study for Indoor Localization Using GSM*, 7th Workshop on Positioning, Navigation and Communication 2010, 11-12 Mars 2010, Dresden, Allemagne.

Annexe 3 : Full-Band GSM Fingerprints for Indoor Localization using a Machine Learning Approach

I. Ahriz, Y. Oussar, B. Denby, G. Dreyfus, *Full-Band GSM Fingerprints for Indoor Localization using a Machine Learning Approach*, International Journal of Navigation and Observation, Hindawi, Mai 2010.

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 that 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. In section III, our two datasets, recorded in different locations using different platforms, are presented, along with the definitions of the fingerprints used in the article. A detailed description of the classification and variable selection algorithms employed appears in section IV. The final indoor localization results, presented in section V, clearly show the importance of including large numbers of carriers in the fingerprint scans. 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 VI 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 VII.

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 chipsets, 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 AND FINGERPRINTS

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 RSSI fingerprints were defined: 1) "Standard", containing about 40 carriers, which includes all carriers which appeared in the set of the 7 strongest ones at least once in the training set; 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.C.

IV. ALGORITHMS

A. Support Vector Machine classifiers (SVM)

Our method aims at using the fingerprints to indicate 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 ($i, j = 1,\ldots,5,\ i \neq j$) 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

$$\mathbf{x}.\boldsymbol{\theta} = 0 \qquad (1)$$

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 will be presented in Table I in the next section.

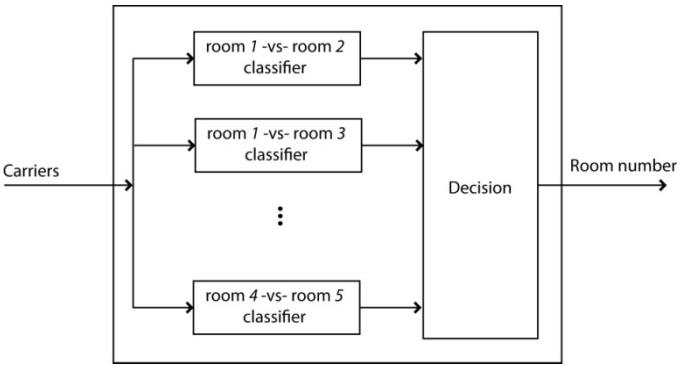

Figure 1. Overview of the system: the input "carriers" of the classifier can be all available carriers, "All"; or a subset thereof, "Standard" or "Relevant".

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 simultaneously with the value of the regularization constant, as explained above.

V. 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. The three fingerprints were compared on our two databases in such a static mode, on an indoor, room-level classification problem. All results presented here are based on the system presented in figure 1. For each database, 80% of the data were used for SVM training and validation; 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 VI.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 VI.

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

VI. 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. While, as shown in Table II, including all GSM carriers gives somewhat better results, it may be necessary to include complementary information in order to obtain further performance improvements. This could take the form of RSSI measurements from other frequency bands, for instance, or of 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.

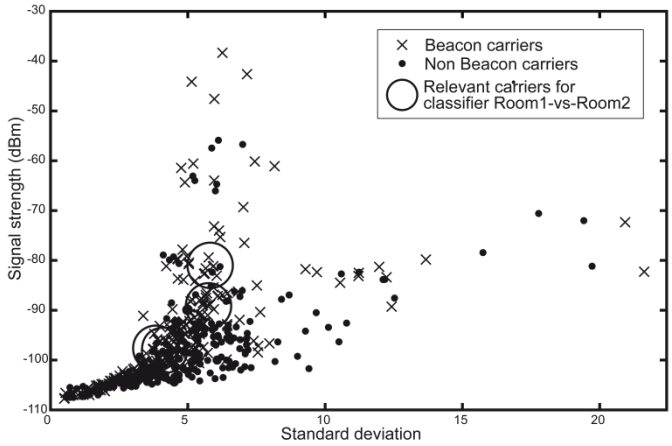

Figure 2. Signal strength and standard deviation of 500 available carriers scanned in room 1.

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.

VII. 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.

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.

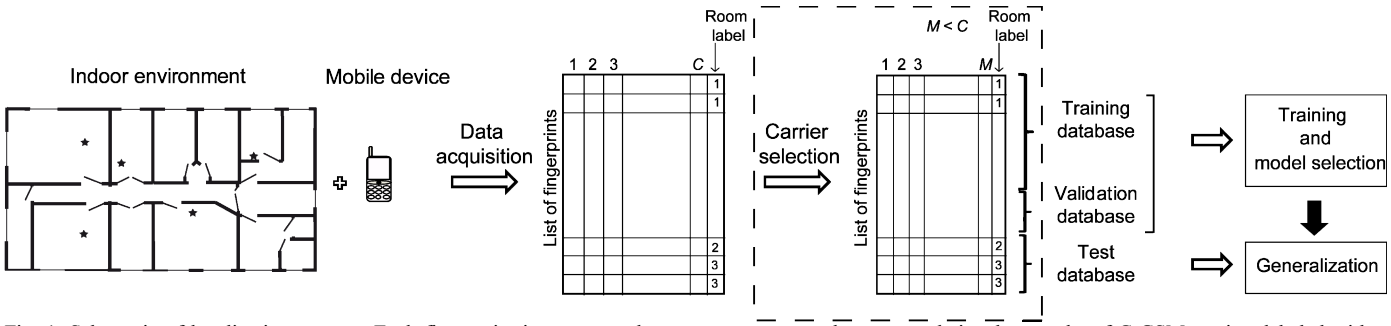

Fig. 1. Schematic of localization process. Each fingerprint is a vector whose components are the measured signal strengths of C GSM carriers labeled with a room number.

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 α_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.

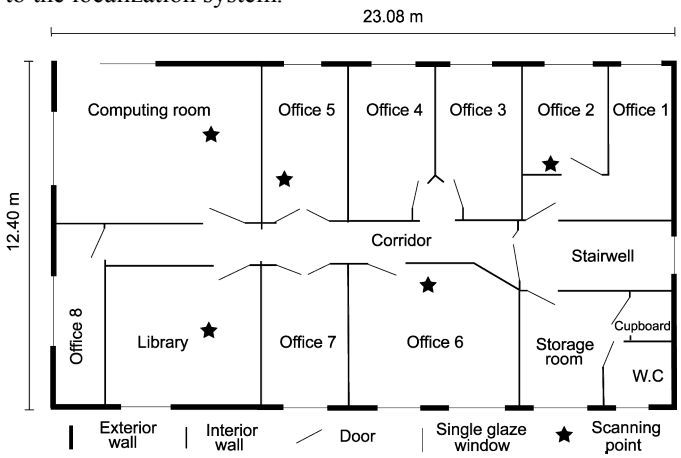

Fig. 2. Laboratory floor plan.

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.

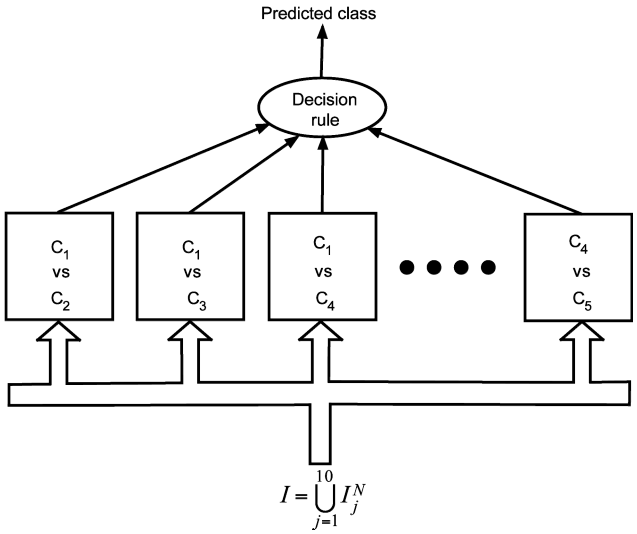

Fig. 3. Localization system composed of 10 "one versus one" classifiers.

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].

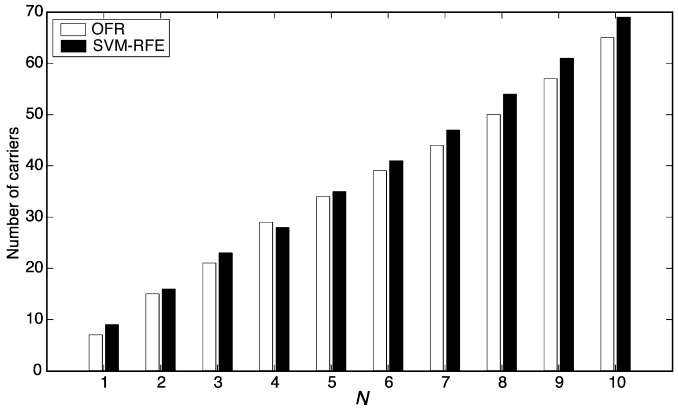

Fig. 4. Number of carriers $M(N) = |I|$ taken into account by the classifier as a function of N.

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).

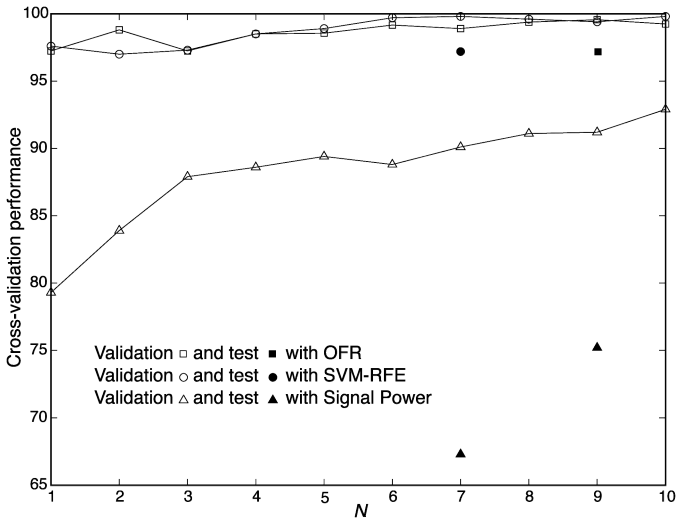

Fig. 5. Validation (symbols and lines) and test set (symbols) performance as a function of N.

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).

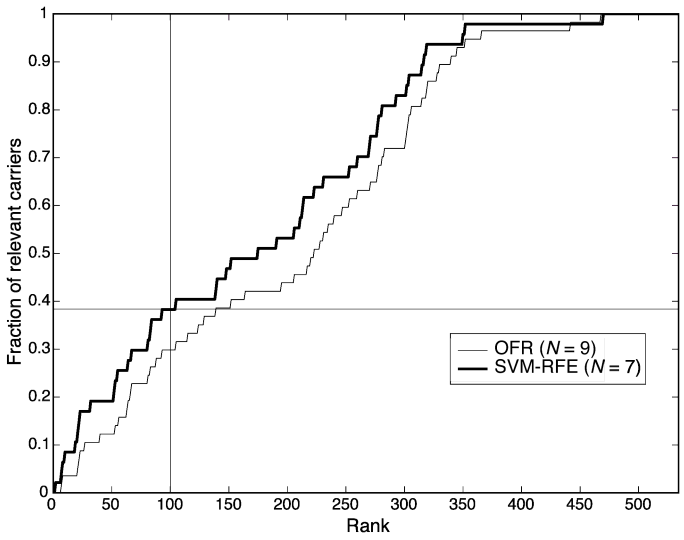

Fig. 6. Cumulative percentage of relevant carriers as a function of their rank in the power ranking of all available carriers, for OFR and SVM-RFE relevance ranking methods. For instance, the graph shows that approximately 40% of the relevant carriers selected by SVM-RFE are among the 100 strongest carriers.

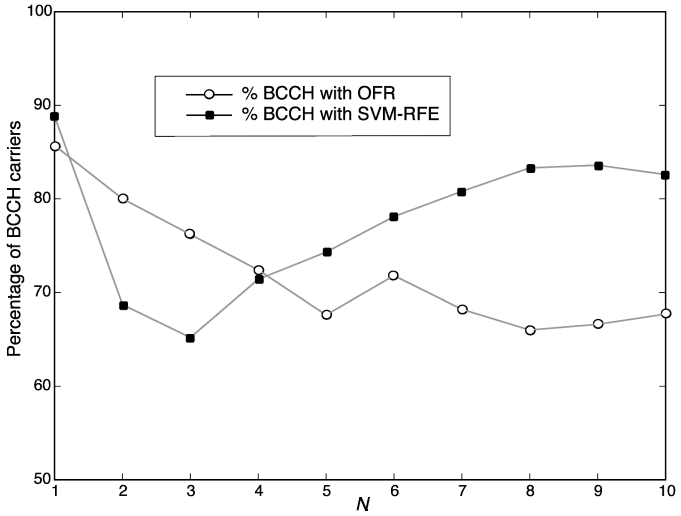

Fig. 7. Percentage of beacon channels in the set of relevant carriers, for several values of N and for the two ranking methods.

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/

Hindawi Publishing Corporation
International Journal of Navigation and Observation
Volume 2010, Article ID 497829, 7 pages
doi:10.1155/2010/497829

Research Article

Full-Band GSM Fingerprints for Indoor Localization Using a Machine Learning Approach

Iness Ahriz,[1] Yacine Oussar,[1] Bruce Denby,[2,1] and Gérard Dreyfus[1]

[1] *Signal Processing and Machine Learning (SIGMA) Laboratory, ESPCI—ParisTech, 10 rue Vauquelin, 75005 Paris, France*
[2] *Université Pierre et Marie Curie—Paris VI, 4 place Jussieu, 75005 Paris, France*

Correspondence should be addressed to Bruce Denby, denby@ieee.org

Received 1 October 2009; Accepted 25 March 2010

Academic Editor: Simon Plass

Indoor handset localization in an urban apartment setting is studied using GSM trace mobile measurements. Nearest-neighbor, Support Vector Machine, Multilayer Perceptron, and Gaussian Process classifiers are compared. The linear Support Vector Machine provides mean room classification accuracy of almost 98% when all GSM carriers are used. To our knowledge, ours is the first study to use fingerprints containing all GSM carriers, as well as the first to suggest that GSM can be useful for localization of very high performance.

1. Introduction

Location-based services for cellular telephone networks are today very much in the public eye [1]. Global Positioning System, or GPS, receivers integrated into cellular handsets can provide very accurate positioning information; however, few mobiles are so equipped at present, and GPS furthermore performs poorly in the indoor and urban canyon environments which are prevalent in wireless networks. For these reasons, the study of localization techniques based upon the radio networks themselves is also a very active area. Most commercially installed systems still rely on cell-ID, in which the mobile station's position is reported as that of the serving base station. Although improvement is possible using triangulation, time of arrival, and the like, the accuracy of such methods is in practice compromised by the path loss and multipath characteristics inherent in the radio channel [2].

The database correlation method [3] allows to overcome channel effects to a certain extent. In this method, a mobile is localized by comparing one of the regularly emitted Received Signal Strength (RSS) measurements to a position-labelled database of such measurements, which are often called fingerprints. Existing localization services implemented in some GSM networks rely on Network Measurement Reports (NMR), which are a part of the GSM norm and contain the RSS and Base Station Identity Code (BSIC) of the serving cell and six strongest neighboring cells. The resulting 7-component vector allows a localization precision of some tens of meters in outdoor environments (see, e.g., [4, 5]).

As for indoor radio-based localization, most studies which have appeared in the literature have involved WiFi networks, describing "corridor waveguide" scenarios in the workplace, and obtaining performance which, though interesting, can still be improved [6–8]. Another approach, using the household power lines as an antenna, appears in [9]. The notion of using GSM or CDMA networks for localization in indoor environments, particularly in domestic settings, is still somewhat new (see, e.g., [10, 11]). The basic idea is that inside a building, the RSS of the external base stations will be strongly correlated with a mobile's exact position, due to for example the varying absorption of electromagnetic energy by different building materials, and the exact placement of doors and windows. There has also been evidence that including more than the standard 7-carriers of the NMR fingerprint is advantageous in indoor GSM localization [10, 12].

In this article, we present tests of indoor GSM localization using scans containing large numbers of carriers—up to the full GSM band. In order to keep working with such large numbers of carriers tractable, we propose to create a mathematical model mapping fingerprints to position using machine-learning techniques, in this case Support Vector

Machines (SVM), and Multilayer Perceptrons (MLP), often also referred to as neural networks. We demonstrate the superiority of the machine learning approach, for problems with such high input dimensionality, over more traditional classifiers based on Euclidean (K-Nearest Neighbor) and Mahalanobis (Gaussian process) distances. Our results show that in an urban apartment setting, the room in which a handset is located can be identified with nearly 98% accuracy when the full set of GSM carriers is included. To our knowledge, this study, which is an extension of that described in [12], is the first to use fingerprints of all carriers in the GSM band, and the first to demonstrate very good performance on indoor localization using GSM.

The structure of the article is as follows. The data sets used in our study are presented in Section 2, while a discussion of preprocessing and the classifiers tested are given in Section 3. Our results are discussed in Section 4, while our conclusion, as well as some perspectives for the future is outlined in the final section.

2. Data Sets

The TEMS [14] trace mobile system was used to take twice-daily scans of the entire set of 498 GSM carriers in 5 rooms of a 5th floor apartment (top floor) in Paris, France. Both the RSS and the BSIC, where readable, were requested for each carrier in the scans. The layout of the apartment is shown in Figure 1. Acquisitions could be made anywhere within a room; however, in practice, the scans were recorded in those areas where the necessary laptop and cellphone could be conveniently set down and accessed. An exhaustive coverage of all rooms was thus not assured.

3. Data Analysis

3.1. Preprocessing. Ten of the carriers were found to contain no energy and were removed from the study. As the BSICs of the remaining 488 proved unreadable in many instances, a decision was made to exclude the BSICs entirely from the subsequent analysis, despite the possibility this engenders of confusing carriers at the same frequency in separate cellular motifs. The data set contained a total of 241 scans—approximately 48 scans per class, where a class is defined here simply as the index of the room within the apartment, indicated in Figure 1. To obtain a measure of the statistical significance of our classification results, cross-validation was performed with ten independent randomly selected splits of our data, each one containing 169 training examples and 72 validation examples. In a given split, the training and validation examples were uniformly distributed over time during the one-month acquisition period.

3.2. Dimensionality Reduction and Fingerprint Types. The relatively small size of our dataset is a reflection of the difficult, time-consuming nature of obtaining labeled scan data—a point to which we will return later. Its high dimensionality (488 carriers) also limits the complexity of the classifiers which may be applied. To deal with these issues, signal strength-based carrier selection was initially carried out so as to define the four fingerprint types defined below. Further dimensionality reduction of any fingerprint can be obtained by a subsequent application of Principal Component Analysis (PCA).

Three vectors are used in defining the fingerprints:

$$\mathbf{g}_j^7 = \left\{ i = 1 \cdots 488, \sum_k \mathbf{1}_{\mathrm{RSS}(i,j) < RSS(k,j)} \leq 6 \right\},$$

$$\mathbf{G}^7 = \bigcup_j \mathbf{g}_j^7, \tag{1}$$

$$\mathbf{G}^{35} = \left\{ i = 1 \cdots 488, \sum_k \mathbf{1}_{\langle \mathrm{RSS}(i,j) \rangle_j < \langle \mathrm{RSS}(k,j) \rangle_j} \leq 34 \right\},$$

where $\mathbf{1}$ is the so-called indicator function, and $\langle\rangle_j$ represents the mean over the index j. The first, $\mathbf{g}_j^7$, contains the indices of the 7 strongest carriers, i, in example j. The vector $\mathbf{G^7}$, composed of the indices of the carriers which were among the strongest 7 in at least one scan of the training set, contains between 36 and 40 of such "good" carriers, depending upon the random split used. The third vector, $\mathbf{G^{35}}$, consists of the indices of the 35 carriers which were the strongest on average, over the whole training set. The fingerprints may then be defined as follows.

(1) *Current Top 7.* These seven carrier fingerprints, $\mathrm{RSS}(\mathbf{g}_j^7)$, are meant to mimic standard "top 7" NMRs, which were not present in our scans. Indeed, NMRs are only logged during a communication, while our scans were obtained in idle mode. Validation set fingerprints can in fact contain less than 7 elements if certain carriers were not represented in the training set. For classifiers requiring fixed labeling of input vectors, such as KNN, SVM, and MLP, the seven $\mathrm{RSS}(\mathbf{g}_j^7)$, values are entered at the corresponding positions in a vector of length $\|\mathbf{G}^7\|$, and the rest of the elements are set to zero.

(2) *Top 7 with Memory.* This fingerprint, defined as $\mathrm{RSS}(\mathbf{G}^7)$, includes the values of all of the 36–40 "good" carriers; they are thus "wider" than the *Current Top 7* fingerprint defined above.

(3) *35 Best Overall.* The *35 Best Overall* fingerprint, of length 35, is defined as $\mathrm{RSS}(\mathbf{G}^{35})$. It thus gives another way of assessing the "goodness" of a carrier, by the size of its average RSS value over the whole training set.

(4) *All 488.* All of the active carriers' RSS values are included in the fingerprint, that is, no selection is in fact made.

3.3. Classifiers. Four types of classifier were tested:

(1) *Support Vector Machines (SVM).* A 2-class SVM classifier [15] finds the separating surface which maximizes the distance (or "margin") between that surface and the data points on either side of it. The SVM can be linear and operate

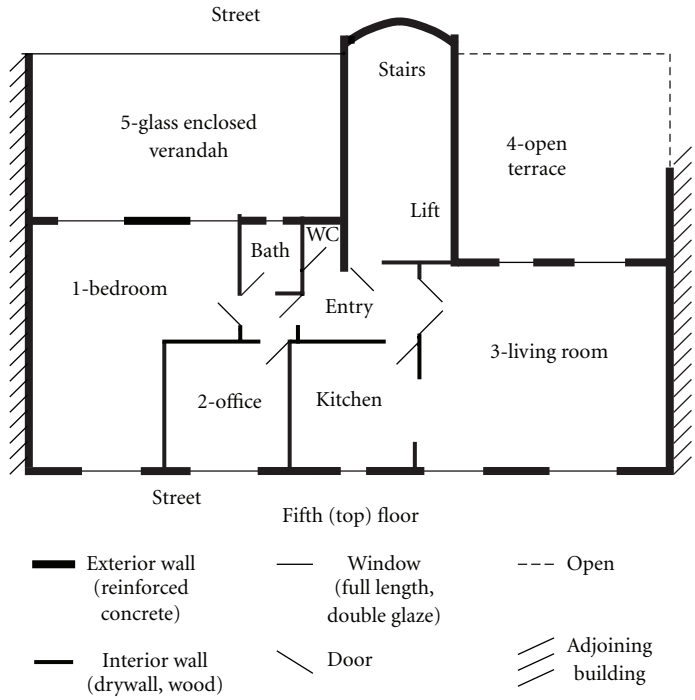

Figure 1: Schematic of apartment layout.

directly upon the data, or first map the data onto a higher-dimensional space using a non-linear transformation, before finding the maximum margin surface. The SVM decision rule takes the sign of

$$f(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b \tag{2}$$

with $\mathbf{x}$ the RSS vector to be localized, N_s the number of support vectors $\mathbf{s}_i$ (training vectors which are on the boundary of the optimal margin), and $y_i = \pm 1$ the class label of the vector $\mathbf{s}_i$. $K(\cdot)$ here is the selected kernel, and b as well as the α_i are parameters determined in the search for the optimal separating surface. For large, well-behaved data sets, the SVM rule approximates the Bayes decision rule [15].

In the case of a linear SVM, the kernel function is just the scalar product $K(\mathbf{s}_i, \mathbf{x}) = \mathbf{s}_i \cdot \mathbf{x}$. The standard Gaussian kernel was adopted in our tests of non-linear SVMs,

$$K(\mathbf{s}_i, \mathbf{x}) = e^{-|\mathbf{s}_i - \mathbf{x}|^2/\sigma^2}, \tag{3}$$

where the variance, σ^2, is optimized in the cross-validation stage. Since a "soft margin" approach was used (i.e., some training examples were allowed to lie within the margin), a regularization parameter controlling the complexity of the separating surface [15] was also estimated by cross-validation. For m classes, it is traditional, using the "conventional recipe" [16], to construct m binary, one-versus-rest classifiers, and take as the output class that of the classifier having the largest output value before thresholding. This procedure is illustrated for the case of $m = 5$ in Figure 2. The Spider SVM package [17] was used in all of our analyses.

(2) *Multilayer Perceptron (MLP).* A multilayer Perceptron is a multivariate, nonlinear, scalar or vector function, which is a combination of parameterized elementary nonlinear functions called neurons [18]. A neuron is usually a function of the form $f = \tanh(\boldsymbol{\theta} \cdot \mathbf{x})$ where $\boldsymbol{\theta}$ is the vector of parameters of the neuron and $\mathbf{x}$ is the vector of variables. A single-output "multilayer Perceptron" $g(\mathbf{x})$ is a combination of N_h "hidden" neurons $f_i (i = 1 \text{ to } N_h)$ and of a constant equal to 1. Denoting by Θ_1 the vector of parameters of the linear combination (of size $N_h + 1$), by Θ_2 the $(N + 1, N_h)$ matrix whose elements are the parameters of the "hidden" neurons, and by $\mathbf{f}$ the vector (of size N_h+1) of functions computed by the N_h hidden neurons with an additional component equal to 1, the multilayer Perceptron function is of the form

$$g(\mathbf{x}) = h(\Theta_1 \cdot \mathbf{f}(\Theta_2 \mathbf{x})). \tag{4}$$

Multilayer Perceptrons are frequently described pictorially as shown in Figure 3.

The parameters of the multilayer Perceptron are estimated from the available training data by minimizing the least squared cost function

$$J(\Theta_1, \Theta_2) = \sum_{k=1}^{n} (y_k - g(\mathbf{x}_k))^2 \tag{5}$$

with respect to all parameters, where $\mathbf{x}_k$ is the vector of variables pertaining to example k and y_k is the measured value of the quantity of interest for example k. In the present study, the gradient of the cost function was computed by a computationally efficient algorithm known as "backpropagation", and the optimization of the cost function was

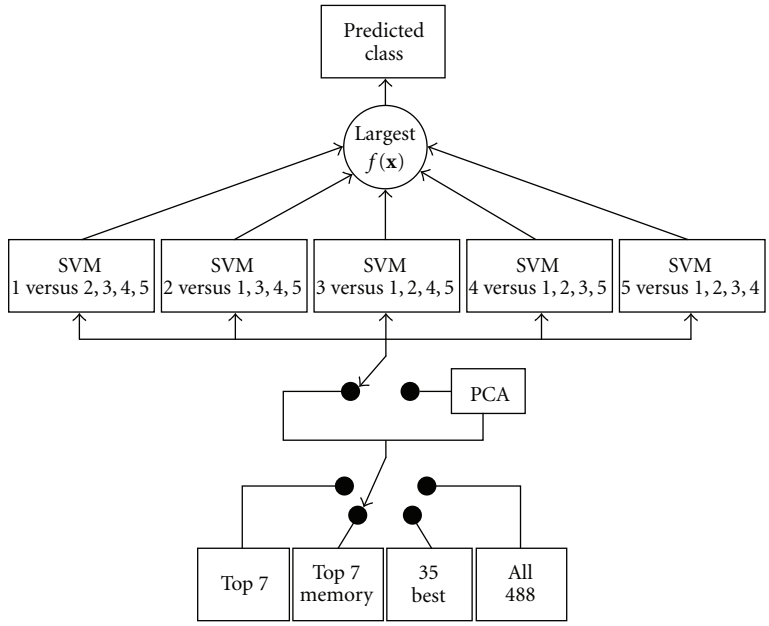

Figure 2: Architecture combining five one-versus-rest SVM classifiers to predict the class of an RSS vector from one of the carrier sets.

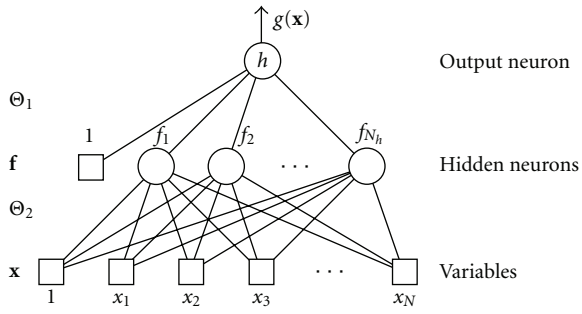

Figure 3: A multilayer Perceptron with a single output.

performed by the conjugate gradient algorithm with Powell-Beale restarts [19].

In a two-class (A, B) classification problem, $y_k = +1$ for all examples of class A and $y_k = -1$ for all examples of the other class. After training, an unknown example described by vector $\mathbf{x}$ is assigned to class A if $\text{sgn}(g(\mathbf{x})) = +1$, and to class B otherwise. In the present study, function h was taken identical to function f. For a c-class problem, example k, belonging to class i ($1 \le i \le c$), is assigned a vector $\mathbf{y}_k$, of dimension c, that encodes the class in a 1-out-of-c code: all components are equal to -1, except component i, which is equal to $+1$. The number of output neurons is equal to the number of classes, so that the output of the multilayer Perceptron is a vector $\mathbf{g}(\mathbf{x})$ of dimension c. The cost function that is optimized during training is

$$J(\Theta_1, \Theta_2) = \sum_{k=1}^{n} \left\| \mathbf{y_k} - \mathbf{g}(\mathbf{x}_k) \right\|^2. \tag{6}$$

In the present study, two strategies were compared for multiclass classification with multilayer Perceptrons.

(i) All functions h were taken identical to f (sigmoid functions), so that the output vector of the multilayer Perceptron was

$$\mathbf{g}(\mathbf{x}) = \mathbf{h}(\Theta_1 \mathbf{f}(\Theta_2 \mathbf{x})), \tag{7}$$

where Θ_1 is the (c, N_h+1) matrix of the parameters of the output neurons.

(ii) Output i ($1 \le i \le c$) of the multilayer Perceptron was computed as

$$g_i(\mathbf{x}) = \frac{\exp[(\Theta_1 \mathbf{f}(\Theta_2 \mathbf{x}))_i]}{\sum_{j=1}^{c} \exp\left[(\Theta_1 \mathbf{f}(\Theta_2 \mathbf{x}))_j\right]} \quad \text{(softmax function)}. \tag{8}$$

In either case, and example described by $\mathbf{x}$ was assigned to the class j such that

$$j = \text{argmax}_{1 \le i \le c}(g_i(\mathbf{x})). \tag{9}$$

In the second case, the components of vector $\mathbf{g}$ belong to $[0, 1]$ and sum to 1, so that they can be interpreted safely as estimates of the posterior probability of class c given the observed vector $\mathbf{x}$.

(3) *K-Nearest Neighbor (K-NN).* As a first step, *K-NN* ranks the training vectors according to their RSS-space Euclidean distances from a test vector to be localized. The predicted class of this test vector is then the class most often represented in the K "nearest" vectors according to the defined metric. The K parameter is chosen empirically, to optimize performance. When a single best neighbor is used, $K = 1$, and the classifier is called *1-NN*.

Table 1: Percentage of correct radio fingerprint classifications on the 4 carrier sets described in the text. Figures quoted are averages and standard deviations over 10 randomly selected validation sets. All classifiers achieve their best performance when all 488 carriers are included. The most effective classifier for this case is the linear SVM.

Classifier		Fingerprint Type: Current Top 7 (≤7 carriers)[1]	Top 7/Memory (36–40 carriers)	35 Best Overall (35 carriers)	All 488 (488 carriers)
Linear SVM		71.3 ± 7.2	84.6 ± 3.6	90.4 ± 3.5	**97.8 ± 1.5**
Gauss. SVM	w/o PCA	72.2 ± 3.6	89.2 ± 2.9	93.2 ± 3.4	—[2]
	w/PCA[3]	71.8 ± 3.2	85.6 ± 5.3	92.0 ± 3.0	96.4 ± 1.5
Linear Perceptron		66.9 ± 4.1	73.2 ± 5.1	79.7 ± 5.1	94.4 ± 2.6
MLP (one versus all)	w/o PCA	66.9 ± 7.1	87.2 ± 3.3	91.8 ± 3.4	—[2]
	w/PCA[3]	68.1 ± 3.4	87.5 ± 4.5	89.6 ± 2.5	95.7 ± 2.1
MLP (multiclass) sigmoids	w/o PCA	56.8 ± 7.1	80.4 ± 12.9	92.6 ± 3.2	—[2]
	w/PCA[3]	66.4 ± 5.7	85.1 ± 9.5	89.4 ± 3.6	96.1 ± 1.1
MLP (multiclass) softmax	w/o PCA	64.3 ± 7.5	85.7 ± 15.8	91.2 ± 4.2	—[2]
	w/PCA[3]	67.7 ± 5.7	88.2 ± 3.9	90.4 ± 3.1	96.6 ± 2.4
K_{best} K-NN		5 59.3±3.5	26 85.1±3.0	20 93.3±2.1	20 94.9±1.9
1-NN		58.1 ± 5.2	74.7 ± 3.7	86.0 ± 2.9	87.2 ± 2.8
GP ($\sigma = 5$ dB)		78.8 ± 3.7		—[4]	

[1]SVM and K-NN can have < 7 carriers if some did not show up in the training set.
[2]Small training set size precludes training a nonlinear classifier due to Cover's theorem [13].
[3]Best result obtained using the first 4 principal components.
[4]Gaussian process is equivalent to 1-NN for fixed input vector length.

(4) *Gaussian Process (GP)*. As in the case of K-NN, GP starts by comparing the test RSS vector to be localized to every vector in the training set. The probability, P_1, that the compared vectors correspond to measurements at the same geographical position is assumed to be Gaussian in the Euclidean RSS distance between the two vectors, using a fixed variance σ^2 which is determined empirically. If a carrier appears in one of the compared vectors, but not in the other, GP presumes that the missing value was below the reception threshold in the vector lacking it. A penalty term probability, P_p, is introduced, in which the missing RSS value is replaced by an estimate of the reception threshold, taken to be the smallest RSS in the vector which is missing the carrier. The overall GP probability, P, is the product of P_1 and P_p.

To be more precise, let A and B be sets of indices of carriers contained in a training set vector, and a test set vector, respectively. We define the set of common carriers as $C = A \cap B$, and the noncommon carrier sets as $D = A - C$ and $E = B - C$, for the train and test sets, respectively. We then have

$$
\begin{aligned}
P_1 &= \sqrt[|C|]{\prod_{i\in C} e^{-|\mathrm{RSS}_i^A - \mathrm{RSS}_i^B|^2/\sigma^2}}, \\
P_p &= \sqrt[|D|]{\prod_{j\in D} e^{-|\mathrm{RSS}_j^D - \min_B(\mathrm{RSS}^B)|^2/\sigma^2}} \\
&\quad \times \sqrt[|E|]{\prod_{k\in E} e^{-|\mathrm{RSS}_k^E - \min_A(\mathrm{RSS}^A)|^2/\sigma^2}}, \\
P &= P_1 \cdot P_p,
\end{aligned}
\tag{10}
$$

where RSS_i^A is the signal strength of the ith carrier of set A, and the order of each root normalizes the probability to the number of carriers in the corresponding term. GP is actually the only classifier tested which is able to handle missing carriers in a natural way. When input vectors are of fixed length—a requirement for SVM, MLP, and KNN—and all variables must be represented, GP is equivalent to a 1-NN classifier. As a caveat, however, as we do not use the BSIC information, in some cases, carriers with the same index can belong to different cellular motifs, which would penalize the GP method.

4. Results

We define the localization performance of a given classifier as the average of the validation scores obtained over our ten random splits, expressed as a percentage of correctly identified locations. The standard deviation over the ten splits is also calculated. The results are shown in Table 1.

A few preliminary remarks about the table are in order. First, when the *All 488* fingerprint is used, it is not meaningful to apply a non-linear classifier to the data. This is because of Cover's theorem [13], which states that the examples of a training set are always linearly separable when the number of input variables exceeds the number of examples. The corresponding table entries are thus left blank (footnote 2 in the table). Secondly, on the other hand, dimensionality reduction by principal component analysis is known to often make examples nonlinearly separable, giving poor performance (nonlinear separability of the training examples was verified using the Ho-Kashyap algorithm [20]).

Table 2: Confusion Matrices for 35 Best Overall and All 488 carrier sets, using a Linear SVM classifier. Figures quoted are in percent. Using the full number of carriers tightens up the diagonal to give individual room classification efficiencies near 100%.

(a)

Confusion Matrix	True class 35 Best overall				
Pred. Class	1	2	3	4	5
1	95	5.3			3.3
2	1.4	93.3	3.6		
3	0.7	1.3	77.9	11.4	
4			16.4	87.9	0.7
5	2.9		2.1	0.7	96

(b)

Confusion Matrix	True class All 488				
Pred. Class	1	2	3	4	5
1	100		0.7		
2		100			
3			91.4	1.4	1.3
4			5.7	98.6	
5			2.2		98.7

For this reason, linear classifiers are not applied in those cases where PCA is used. A few further details are explained in the remaining footnotes of Table 1.

The table shows that the performance of all classifiers tested improves as more carriers are added to the fingerprint, but that very good performance—for example, our best result of 97.8% in the case of the linear SVM—is only obtained on the *All 488* carrier fingerprint. The implication is that indoor position can indeed be deduced from the RSS of GSM cell towers, but that commonly used 7-carrier NMRs and even "wide" fingerprints are insufficient: high performance requires fingerprints of very high dimensionality. It is reassuring to see that this conclusion is supported by all the classifiers tested, including a simple *K-NN*, even if the best results are obtained with SVM and MLP machine learning techniques. MLP performance appears slightly worse than that of linear SVMs, within the statistics of our sample, with the best MLP performance, 96.6%, obtained on a multiclass MLP with the softmax output function applied to *All 488* carriers, after an input dimensionality reduction.

A more detailed look at our conclusion is given in Table 2, where the confusion matrices for the linear SVM classifier on the *35 Best Overall* and *All 488* fingerprints appear. The table shows once again that the ability to sharply discriminate between rooms comes only with the inclusion of the full GSM carrier set. The deviation of our global result from 100% is in fact dominated by the confusion between class 3 and class 4, which appears to be the most difficult case.

5. Conclusions and Perspectives

We believe this study, which is an extension of that presented in [12], to be the first to include the full set of GSM carriers in RSS fingerprints for localization. Although confirmation with more extensive databases will be required, our results strongly suggest that high-performance room-level localization is possible through the use of such fingerprints. The fact that good performance is obtained irrespective of the machine learning technique used (MLPs or SVMs,) is a further confirmation that the useful information for localization is obtained by taking into account many GSM carriers, including those which may be rather weak. Finally, it is interesting to note that our result is robust against time-dependent effects—network modifications, propagation channel changes, meteorological effects, and so forth, as our dataset was acquired over a period of one month.

Acquiring datasets and labeling scans is a tedious and time-consuming activity. To address this issue, two independent solutions are currently being investigated. First, experiments with semisupervised classification techniques using kernel methods (see, e.g., [21]) are being carried out, which will permit to take advantage of the unlabeled scans during the training procedure. The second approach entails the design and construction, in our laboratory, of a set of ten autonomous scanning devices which will allow the acquisition of large datasets simultaneously in different rooms, labeled with very little human intervention. These devices will also enable to test the efficiency of our approach when implemented using mixed datasets of scans acquired both indoors and in nearby outdoor areas. For larger outdoor areas, preliminary results indicate that a regression approach using x-y coordinates seems more suitable than the room-by-room classification used here for indoor localization.

References

[1] A. Küpper, *Location-Based Services: Fundamentals and Operation*, John Wiley & Sons, New York, NY, USA, 2005.

[2] H. Laitinen, et al., "CELLO: cellular network optimization based on mobile location," Internal Report IST-2000-25382-CELLO, EU IST Project, October 2001, http://www.telecom.ntua.gr/cello/documents/CELLO-WP2-VTT-D03-007-Int.pdf.

[3] D. Zimmermann, J. Baumann, M. Layh, F. Landstorfer, R. Hoppe, and G. Wölfle, "Database correlation for positioning of mobile terminals in cellular networks using wave propagation models," *Proceedings of the 60th IEEE Vehicular Technology Conference*, vol. 7, pp. 4682–4686, September 2004.

[4] M. Chen, T. Sohn, D. Chmelev, et al., "Practical metropolitan-scale positioning for GSM phones," in *Proceedings of the 8th International Conference on Ubiquitous Computing*, P. Dourish and A. Friday, Eds., vol. 4206 of *Lecture Notes in Computer Science*, pp. 225–242, Springer, Orange County, Calif, USA, September 2006.

[5] B. Denby, Y. Oussar, and I. Ahriz, "Geolocalisation in cellular telephone networks," in *Proceedings of the NATO Advanced Study Institute on Mining Massive Data Sets for Security*, F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger, Eds., IOS Press, Amsterdam, The Netherlands, 2007.

[6] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless LANs," *Computer Networks and ISDN Systems*, vol. 47, no. 6, pp. 825–845, 2005.
[7] A. M. Ladd, K. E. Bekris, A. P. Rudys, D. S. Wallach, and L. E. Kavraki, "On the feasibility of using wireless ethernet for indoor localization," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 3, pp. 555–559, 2004.
[8] Q. Yang, S. J. Pan, and V. W. Zheng, "Estimating location using Wi-Fi," *IEEE Intelligent Systems*, vol. 23, no. 1, pp. 8–9, 2008.
[9] S. N. Patel, K. N. Truong, and G. D. Abowd, "Powerline positioning: a practical sub-room-level indoor location system for domestic use," in *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp '06)*, Orange County, Calif, USA, September 2006.
[10] V. Otsason, A. Varshavsky, A. LaMarca, and E. de Lara, "Accurate GSM indoor localization," in *Proceedings of the International Conference on Ubiquitous Computing (UbiComp '05)*, M. Beigl, et al., Ed., pp. 141–158, Springer, Berlin, Germany, 2005.
[11] W. ur Rehman, E. De Lara, and S. Saroiu, "CILoS: a CDMA indoor localization system," in *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*, pp. 104–113, Seoul, Korea, 2008.
[12] B. Denby, Y. Oussar, I. Ahriz, and G. Dreyfus, "High-performance indoor localization with full-band GSM fingerprints," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '09)*, Dresden, Germany, June 2009.
[13] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. 14, pp. 326–334, 1965.
[14] Test Mobile System, http://www1.ericsson.com/solutions/tems.
[15] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.
[16] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.
[17] The Spider, http://www.kyb.tuebingen.mpg.de/bs/people/spider.
[18] G. Dreyfus, *Neural Networks: Methodology and Applications*, Springer, New York, NY, USA, 2005.
[19] M. J. D. Powell, "Restart procedures for the conjugate gradient method," *Mathematical Programming*, vol. 12, no. 1, pp. 241–254, 1977.
[20] Y.-C. Ho and R. L. Kashyap, "An algorithm for linear inequalities and its applications," *IEEE Transactions on Electronic Computers*, vol. 14, no. 5, pp. 683–688, 1965.
[21] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, Mass, USA, 2006.